An Empirical Framework for Objective Testing for P-Consciousness in an Artificial Agent

Colin Hales*

NICTA, Victoria Research Laboratory, Department of Electrical and Electronic Engineering, University of Melbourne, VIC 3010, Australia

Abstract: Two related and relatively obscure issues in science have eluded empirical tractability. Both can be directly traced to progress in artificial intelligence. The first is scientific proof of consciousness or otherwise in anything. The second is the role of consciousness in intelligent behaviour. This document approaches both issues by exploring the idea of using scientific behaviour self-referentially as a benchmark in an objective test for P-consciousness, which is the relevant critical aspect of consciousness. Scientific behaviour is unique in being both highly formalised and provably critically dependent on the P-consciousness of the primary senses. In the context of the primary senses P-consciousness is literally a formal identity with scientific observation. As such it is intrinsically afforded a status of critical dependency demonstrably no different to any other critical dependency in science, making scientific behaviour ideally suited to a self-referential scientific circumstance. The 'provability' derives from the delivery by science of objectively verifiable 'laws of nature'. By exploiting the critical dependency, an empirical framework is constructed as a refined and specialised version of existing propositions for a 'test for consciousness'. The specific role of P-consciousness is clarified: it is a human intracranial central nervous system construct that symbolically grounds the scientist in the distal external world, resulting in our ability to recognise, characterise and adapt to distal natural world novelty. It is hoped that in opening a discussion of a novel approach, the artificial intelligence community may eventually find a viable contender for its long overdue scientific basis.

INTRODUCTION

In artificial intelligence (AI) the gold standard for intelligent behaviour is that of humans, in which case AI becomes 'artificial general intelligence' (AGI). The road to human level AGI is paved with underperformances, faddistic enthusiasms, scare-mongering and an 'AI Winter' of the nineteen eighties. Recently it was claimed that AI is 'lost in the woods'. One practical aspect arguably contributing to those difficulties is the lack of a standardised test. Without any scientific proof, a claim that a machine has intellect/cognitive faculty X has no authority. This makes discussion of all the other issues seem a little premature [1-10].

To address this shortcoming, the following work delivers a classic empirical test designed specifically for AGI. In the proposed test science operates normally: (a) make a reasonable judgement as to a critical dependency in the natural world, and (b) configure testing that explores the doubt surrounding it in the most decisive and revealing way possible. The unusual nature of the proposed test is its self-referentiality. It makes use of the biological basis of scientific behaviour that results in scientific 'laws of nature'. The AGI is expected to deliver scientific behaviour. That is claimed as sufficient proof that P-consciousness exists in the AGI.

The relationship between biology and scientific laws has been nicely put by Kitcher:

> "*Science is not done by logically omniscient lone knowers but by biological systems with certain kinds of capacities and limitations. At the most fine grained level, scientific change involves modifications of the cognitive states of limited biological systems*" [11].

With a focus on human neurology and brain electrodynamics[1], it is claimed here that scientific proof of human-level intelligence in an artefact is intimately and necessarily related to scientific proof of P-consciousness in that artefact. P-consciousness is a technically specific term referring to the subjective qualities of human internal life experienced in the 1st person, and is variously named 'P-consciousness' [12], 'phenomenal consciousness' [13], 'qualia' [14] or 'phenomenality' [15]. 'P-consciousness' is used here merely because some recent relevant AI testing literature used it [16]. A recent review by Zeman [17] and a Blackwell monograph [18] are excellent grounding in the terminology. As a quick way to appreciate the technical specificity of the term P-consciousness, the reader is directed to 'phantom limb syndrome' (P-consciousness depicting nonexistent body parts) and 'blindsight' (successful manipulation of body parts without any visual P-consciousness) [17, 18].

Albert Einstein is one of many throughout the last century who recognised the critical dependency of science on P-consciousness. Whilst rather quaintly attributing the natural

*Address correspondence to this author at the NICTA, Victoria Research Laboratory, Department of Electrical and Electronic Engineering, University of Melbourne, VIC 3010, Australia;
E-mail: Colin.Hales@nicta.com.au

[1] Specifically: many-body brain electrodynamics that is literally an identity with 'cognitive states' of the nervous system of the human scientist.

world's comprehendability to a miracle [19][2], Einstein astutely but obliquely touches upon both the critical P-consciousness dependency and the externalisation of our apprehension of the natural world:

> "*The whole of science is nothing more than a refinement of everyday thinking. It is for this reason that the critical thinking of the physicist cannot possibly be restricted to the examination of concepts of his own specific field. He cannot proceed without considering critically a much more difficult problem, the problem of analyzing everyday thinking.*
>
> *On the stage of our subconscious mind appear in colorful succession sense experiences, memory pictures of them, representations and feelings. In contrast to psychology, physics treats directly only of sense experiences and of the "understanding" of their connection. But even the concept of the "real external world" of everyday thinking rests exclusively on sense impressions*" [19][3].

It is thus the P-consciousness of the primary senses which connects a scientist to the studied external world. The externalisation of apprehension of the world beyond our physical boundary uses information measured at the physical boundary. For 150 years it has been known in physiology that humans, and therefore human scientists, cannot/do not apprehend the external natural world via their peripheral nervous system (PNS) boundary sensory transduction (retinal, haptic, auditory sensors etc). Human cranial central nervous system (CNS) performs the externalisation via the constructs of the primary senses (perceptual fields) of vision, touch, hearing and so forth. The spinal CNS and the PNS merely deliver sensationless measurements. Put simply, vision occurs in the brain, not in the eyes.

The CNS primary senses are, in the context of scientific behaviour, literally scientific observation. When science is performed there is also usually some kind of measurement apparatus in the distal external world. To avoid confusion we must carefully discriminate the following terms and their origins, within the context of a scientific act:

(a)	CNS	Scientific Observation
(b)	PNS	Sensory Measurement
(c)	Distal	Scientific Measurement

Without (a) there is no science outcome. However, (c) can proceed without (a)/(b), although if (a) and (b) are present, quantum mechanics tells us that outcome (c) will be impacted in some way and that it is up to us as scientists to evaluate that impact. Faculty (a) necessitates (b). On its own, however, (b) is insufficient to generate (a) (see below). In the context of a machine expected to deliver a scientific act, it is important to realise that mere measurement of a voltage at the periphery of the machine (for example) is (b) *not* (a). Cross-correlation of such (b) peripheral nervous system measurements does *not* constitute a scientific observation (a) because the sensory measurement has a fundamentally degenerate and thus intrinisically ambiguous relationship with the distal origins of the scientific measurements (c) that are actually the focus of the scientist's attention. This is discussed further below.

The historical discourse in neuroscience, AI and elsewhere is imbued with an apparent inability to scientifically prove the presence, absence and kind of P-consciousness of yourself, someone else, another biological organism, a rock or a computer. No scientifically valid test currently exists and our behaviour in the matter is inconsistent. If the reader is inclined to claim Z0 = "*proving consciousness is impossible*", the claim can be rejected because there is no empirical basis for the claim. There is no documented empirical or theoretical proof of claim Z0. No citation of evidence is possible. There is only documented evidence of the claim of impossibility, which is equivalent to no scientific proof at all. It is merely baseless unscientific hearsay; a convention of the same kind as '*man cannot fly*'. Similarly, any claim to already be testing for consciousness is usually demonstrably laden with unproven theories about the origins of consciousness, thus laying the foundations for successful critical attack. These are the characteristics of an era in consciousness studies that has not clarified its fundamentals. The science of consciousness has been called 'pre-paradigmatic' in that regard [20].

The particular objective test outlined later is a test for P-consciousness which scientifically proves the presence, in the test subject, of the minimal *visual* P-consciousness necessarily involved in a scientific act. The tested entity fits the very specific AI sub-class of situated, embedded, embodied artificial agency with the potential for human-level intellect or better. The experimental outcome is witnessed through observation of the external behaviour of an agent involved in a scientific act. External behaviour of the agent is decisive. The test is a specialised refinement of the "Total Turing Test" [21], which has origins that go all the way back to the original Turing Test [22, 23].

THE BASIC TEST PROPOSITION - THE 'PCST'

Empirical neuroscience and physics evidence (below) support the claim Z1 = "*P-consciousness of a specific kind is a causally necessary component of intelligent behaviour of a specific kind: scientific behaviour*". Conversely, this means that the behaviour conclusively proving the existence of that particular kind of P-consciousness is scientific behaviour dependent on the P-consciousness of the requisite kind. Therein lies the potential for testing.

Assume for the moment that the claim Z1 is valid and that scientific testing for scientific behaviour is possible. Call the test PCST for 'P-Conscious Scientist Test'. PCST execution on a test subject results in (a) a demand for scientific behaviour from a test subject in a laboratory circumstance, (b) a claim that the test subject cannot possibly behave that way unless it has P-consciousness of the necessary kind and (c) being believed by your peers. All of these things present

[2] Page 61

[3] Page 59

challenges and are to be covered now, prior to detailing an example PCST.

The PCST merely proves the necessary presence of visual P-consciousness, not any particular physics of P-consciousness or any other related but incidental P-consciousness. The testers do not actually have to know a-priori the requisite physics. Drugs that pass clinical testing are accepted for use without full or proven knowledge of their mechanism. The same principle applies here. Why should P-consciousness be different? Similarly, humans did not wait for a scientific account of combustion before cooking dinner with it. Using combustion eventually resulted in understanding. A putative physics of P-consciousness can be used to construct an 'artificial scientist'. Passing the PCST delivers evidence that the designer has found and sufficiently developed the use of the requisite physics of P-consciousness. Failing the PCST refutes the design and possibly the putative physics underlying that particular design.

Thus, if valid, the PCST looks rather familiar to empiricists. A human scientist is a wild-type positive control/benchmark. An artificial scientist would be the formal AGI test subject. If computationalism is false, as has been claimed [24], a computationalist test subject could be a placebo/sham control. All contenders can be subjected to the PCST in circumstances which neither testers nor tested have encountered. This has the feel of standard clinical scientific testing. The PCST just needs a little more meat on its bones.

SCIENTISTS AS SCIENTIFIC EVIDENCE OF P-CONSCIOUSNESS

Critical Dependency

If Y depends critically on X then without an occurrence of X, phenomenon Y will not be observed. Isolation of such critical dependency, revealed through scientific observation, constitutes maximal scientific proof of an apparent causal relationship that is literally scientific understanding. The predictive efficacy of scientific belief supervenes on knowledge of such critical dependencies acquired by procedurally isolating them from mere correlations.

For example, the causal descendents of a neutrino were accepted as clinching evidence of a neutrino because we found and agreed on a 'critical dependency' of the observation on the prior existence and involvement of a neutrino. Nobody has ever 'seen' a neutrino. As you read this you are being flooded with neutrinos and you will never see them. All we have is an *agreement* that it must have been involved in the observed scientific measurement outcome. Our confidence is such that we hold neutrinos as a proven fact of the natural world. In precisely this way, critical dependency is the basic currency of empirical sciences and these presently operate successfully far beyond the capacity of humans for direct observation.

In a fashion identical to the neutrino discovery, scientists are verifiably critically dependent on the existence and role of P-consciousness in their own behaviour. Experience (P-consciousness) was established as essential in science by Aristotle over two thousand years ago [25]. Today it remains the case that scientific claims are unacceptable unless supported by predictable natural world observations. Claims devoid of empirical proof are rejected in peer review. We scientists thus demand P-consciousness be used. There is nothing else to use. This critical dependency is empirically demonstrable. Physiological deficits such as drug induced hallucination or schizophrenia, which directly impact P-consciousness, cause a scientifically observable commensurate deficit in general belief system operation, including the ability to form special beliefs called 'scientific beliefs'. Take P-consciousness away and your internal life has the quality of a dreamless sleep – devoid of all experience, including that which constitutes scientific observation. Example after example can be cited to demonstrate that scientific behaviour degrades directly proportionally to the efficacy of P-consciousness in delivery of scientific observation. Having said that please note that P-consciousness is not a perfect natural observation system. It can misrepresent, malfunction and be fooled. This is the reason that 'virtual reality' is possible. Scientists actively work to minimise the effects of these things. The fact that such behaviour is necessary adds weight to the claim of critical dependency, not the reverse.

In accepting (demanding) P-consciousness as the ultimate authority behind scientific belief, scientists tacitly declare a critical dependency of the same kind used everywhere else in science. It's just that the critical dependency acts in the causal ancestry of all scientific outcomes, and the natural phenomenon behind the critical dependency is the P-consciousness of the scientist. We cannot have it both ways. P-consciousness cannot be demanded as evidence in one context and then denied as being evidenced in another. In addition note that all natural phenomena are evidence of some kind of critical dependency and we scientists cannot claim an exemption from this state without sounding rather odd.

The Historical Inconsistency in our Treatment of Scientific Evidence

Scientific behaviour and everyday problem solving behaviour are very similar and both originate in the brain. Yet they are not the same. Scientific behaviour is a demonstrably unique natural phenomenon because it delivers portable, explicit, verifiable belief systems (possibly in an abstracted form) in respect of the external natural world. Explicitness and portability distinguish scientific behaviour from everyday problem solving. Scientific observation literally sees beyond the patterns of external reality. Scientists target abstract underlying rules predictive of distal external world behaviour in other contexts, independent of any particular scientist. Empirical corroboration mandates that a natural world phenomenon becomes 'contents of P-consciousness' in more than one scientist.

The PCST detects behaviour identical to scientific behaviour, and uses that to claim the necessary involvement of visual P-consciousness. This is equivalent to accepting human scientists as evidence of P-consciousness for the very reasons outlined above. Let us consider the act of denying that human scientists have visual P-consciousness. To claim that human scientists are not conclusive evidence of P-consciousness is like accepting a measurement of time from a special, demanded clock and then denying the existence of the clock, or perhaps accepting the truth of the utterances of a demanded, trusted source but then denying the existence of that utterer. In itself this is rather odd, for if a trusted source said to you "*X is true*" then which has more certainty – (a) That X is true? or (b) That your trusted source has said something (that you have evidence of the trusted source)? Logically it must be (b). In science P-consciousness *is* our 'trusted source' and is thus more evidenced than anything else! In this way, to deny scientists are appropriately P-conscious is to deny '*that which is scientifically seen*' the secondary and equally valid status as '*scientific evidence of seeing*'. This is the faulty logic entailed by the denial that scientists have P-consciousness whilst demanding it be used and being critically dependent on it in a verifiable fashion.

The mainstream empirical paradigm called 'neural correlates of consciousness' (NCC) implicitly recognises the existence of P-consciousness [26-29] because P-consciousness itself is its primary research focus. In the NCC observations of brain tissue behaviour are correlated with test subject reports of P-consciousness. NCC science is our present best attempt at an 'explanation' of P-consciousness and a large number of complex issues and controversies surround claims in that regard. The oddity here is that, unlike anywhere else in science, heresay is accepted as scientific evidence. The detailed exploration of this situation is very interesting but has no effect on the PCST proposal. The interested reader is also directed to the 'mind-brain identity theorem' and philosophy of science. The discourse has no practical impact here and is therefore not a useful dialogue to present.

THE CLASSICAL DOUBTS

The mantra that automatically gets replayed as rebuttal of any sort of claimed scientific proof of consciousness goes something like "*I can't prove you are conscious and vice versa*". This has been used so much that it's recited without thinking and may have done us a disservice.

The PCST directly challenges the '*you can't prove…*" claim. The PCST enacted on test subject X entails that an observer of the requisite behaviour have no more doubt as to the existence, role and status of P-consciousness in X than the observer might have in any other situation in science. This is not a claim of zero doubt, for it is always possible to construct some doubt. In science, doubt is merely sufficently allayed to render a proposition worthy of practical attention. A novel proposition must make a testable prediction of some sort. The PCST merely expects the same treatment in propositions in respect of the P-consciousness of scientists. For PCST purposes, proposition Z2 = "*I can't prove you are conscious and vice versa*" is recognised as having no supporting evidence and shall be dropped. A very reasonable doubt that Z2 is true can be constructed and that doubt is based on empirical isolation of a verifiable critical dependency. The PCST acquires its authority thus, in an act of plain ordinary science.

To see how culturally awkward a 'test for consciousness' feels, put yourself in the observer's shoes and ask "*I know I signed-off on the test procedure PCST, it sounds right … but…. how do I really know X has P-consciousness?*". It is quite confronting to be asked to accept that what has been observed in the lab proves that P-consciousness exists within test subject X. The usual retort goes something like this: "*I can imagine the machine X behaving that way and yet having no P-consciousness at all. What is it about the observation that assures me that there is P-consciousness inside it?*".

In answer to this note that the argument "*I can imagine.....*" cannot be given special privileges when dealing with P-consciousness. If the argument is compelling for P-consciousness, then every scientific experiment or hypothesis must be permitted the same defeat. For example: "*I can imagine particle Y being produced without the involvement of particle Z*". Nobody would allow such an argument to disprove the discovery of the neutrino, which has this evidentiary basis. How can "*imagining a counterfactual*" be valid as rebuttal only in the case of P-consciousness? Rather than decry the validity of the situation with unjustified imaginings, it is up to the doubter to ensure that the PCST is exquisitely attuned to the claimed critical dependency and to deal with the reality of the circumstance with the same logic used everywhere else in science. Then to the best of our ability we have a scientifically sound and empirically supported argument about the P-consciousness of a test subject.

Another form of the same structural misdirection is the so-called 'zombie argument'. A philosophical zombie is a creature defined by humans that is physically identical to a human (say, at the atomic level) but with no P-consciousness [13, 30]. Accepting the conception of such a creature implicitly denies the empirically informed physical reality of P-consciousness recognised in the NCC science paradigm (see above). The entities described as atoms dancing about in the formation called a brain correlate with P-consciousness well enough to justify an entire science paradigm. To then mentally allow those atoms to do the identical 'dance' and yet disable their production or involvement in P-consciousness is nonsensical. Yes, the collection of entities we describe as 'atoms' contains no prediction or description of the production of P-consciousness. Nor does it predict any observer, observing or scientists. However, this situation does not entail that, whatever atoms actually are, their behaviour is not intimately necessitated in the production of P-consciousness of the type we have. Like the '*I can imagine…*' argument, the philosophical zombie is an incoherent contribution to empirical inference processes and deserves no authority over the viability of the PCST.

We need to bring an end to illogical thinking and its misdirections in favour of reasonable positions predicting testable outcomes as suggested by normal scientific processes. The energy of doubt needs some long overdue empirically informed wisdom in application to the problem of P-consciousness.

THE HISTORY OF TESTING FOR MACHINE INTELLIGENCE AND CONSCIOUSNESS

This section has been provided to contrast the PCST with previous similar testing propositions and can be skipped by those only interested in the PCST itself.

Turing called the original Turing Test (TT) the 'imitation game'. It involves limiting a human test subject to the communication level of a computer (via a computer termminal). The computer and human communications are then contrasted by a human interrogator who is unaware which is which [22, 23]. The 'structured mistake' that makes the competing streams indistinguishable was supposed proof of equivalence of machine and human intelligence. To connect the PCST to the TT note that the Turing-Machine in the test is not embodied or embedded/situated cognition and fits into the class of computationalism. A very neat definition of computationalism comes from Randolf Beer:

> "*...the theoretical claim that a system's behavior derives from its instantiation of appropriate representations and computational processes"* [31].

This definition does not exclude 'the scientist' from being artificially created using "*appropriate representations and computational processes*". This assumes that one can completely model a modeller of the exquisitely novel and intrinsically unknown in a situation where, by definition, all novelty is forced to be characterised by existing models and all new models have to be characterised by a computationalist model for making new models, not by the (distal natural world) novelty itself. In this light computationalism is at best a very suspect principle. The idea that the original Turing Machine 'tape and tape reader/punch', which is an example of the agency shown in Fig. (**2b**), can even begin the PCST seems far-fetched. Based on a subsequent assessment of Turing's own attitude to the process, this is something that Turing would probably also have predicted [32]. It is for this reason that Stevan Harnad proposed an extreme upgrade to the Turing Test called the 'Total Turing Test' or TTT, which requires that:

> "*The candidate must be able to do, in the real world of objects and people, everything that real people can do, in a way that is indistinguishable (to a person) from the way real people do it*" [21].

Beyond the TTT a 'Lovelace Test' (LT) was considered. This test focused on the creativity (originality) of humans as an indicator of intelligence. Various models of artificial creativity were examined in a disembodied test regime similar to the original TT. The conclusion of the work was in the negative but is a useful example of a refinement of the TTT that may have born fruit [10]. The PCST can be viewed as an empirically viable variant of the TTT and LT obtained by choosing a single very specialised behaviour: scientific behaviour. The benefit of this choice is that the Harnad clause '*indistingishable (to a person)*' is scientifically verified by the PCST and the behaviour is critically dependent on P-consciousness. With respect to 'originality' the PCST is also a variant on the LT because originality is implicit in the act of science, where the outcome the unknown novel scientific law.

Aleksander and Dunmall published a form of 'test' for scientific evidence of consciousness in an artefact [33]. Their test involves recognising, in a test subject, evidence of the attributes defined by a collection of five behavioural axioms. The axioms originate in psychology [34] and their 1st-person form is:

1. *Presence/depiction/sense of place*: I feel that I am an entity in the world that is outside of me.
2. *Imagination*: I can recall previous sensory experience as a more or less degraded version of that experience. Driven by language, I can imagine experiences I never had.
3. *Attention*: I am selectively conscious of the world outside of me and can select sensory events I wish to imagine.
4. *Volition/planning*: I can imagine the results of taking actions and select an action I wish to take.
5. *Emotion*: I evaluate events and the expected results of actions according to criteria usually called emotions.

The P-conscious externalised apprehension of the distal natural world critical to the PCST is in axiom 1. Clearly the PCST test subject must also possess enough *volition* to naturally exhibit successful scientific behaviour (axiom 4). Likewise, during the execution of a scientific act it must *attend* to it (axiom 3). The formal necessity for *emotional* content or any particular mechanism mediating decision making is irrelevant to the PCST (axiom 5). In the context of PCST science, *imagination* is clearly necessitated in any ability to hypothesise anything (axiom 2).

The only axiom that matters in the PCST is axiom 1. We know that unless axiom 1 is satisfied, everything else that the test subject does is questionable in a 'garbage-in, garbage-out' sense. The successful execution of the PCST implicitly delivers some kind of evidence of axioms 2-5, but this evidence has no bearing on the critical dependency isolated in axiom 1 in the same way that scientists render such things objectively 'controlled out' and irrelevant to a scientific outcome. Claims about the internal mechanisms of volition, emotion, imagination and so forth acquire their scientific authority post-hoc. Until the PCST is passed, these latter claims are without scientific basis and have been engineered based purely on speculation by the designers as to the nature of P-consciousness.

The '5-axiom' approach has some logical difficulties. There is clearly question-begging in the five axioms. Emotions are literally one type of P-consciousness. Imagination includes internal visual P-consciousness. The next logical problem is that Aleksander and Dunmall propose that if an abstraction-architecture 'respects' the axioms then the artefact is conscious [33]. This is invalid scientific evidence of consciousness, because the five axioms become a tautologous set of design requirements, not revealed causal relationships of the natural world. This tautologous aspect of computer science has already been recognised elsewhere for some time [31]. Testing to your own specification is quite valid and normal in computer science. However, that testing does not result in isolating any critical dependency and in that sense is *not science*. It merely correlates a document with observed behaviour. The document has an indirect relationship with the resultant causation. It is merely a correlate, not a critical dependency. The document does not actually cause the behaviour. The computer hardware, configured according to rules based on the document, is the actual causal necessitator of all behaviour. One could implement the same program on ten totally different computer architectures, resulting in ten totally different sets of causal relationships in the physics of each computer. According to this claim all of them would be equally conscious. The proposition seems logically untenable.

This problem is a variant of the general problem of hearsay in science. For example, consider that empirical metascience evidence provided by scientists about what they do is not admissible evidence in the study of the behaviour of scientists. You cannot ask them. They have to be unobtrusively observed. Evidence that scientists provide verbally about what they think they do is evidence for some other aspect of the study of scientists. This hearsay evidence has precisely the same status as a design document such as the '5-axioms' or a computer program. Such documentation cannot be given any scientific authority in a detached, unbiased and objective scientific description of the critical causal dependencies of the natural world.

In "*Offer: One Billion Dollars for a Conscious Robot; If You're Honest, You Must Decline*", Selmer Bringsjord follows his earlier pronouncement of the death of computationalism to an obvious conclusion whilst leaving the door open for non-computationalist contenders. The result is prescriptive of the expected failure of all computationalist contenders, but does not define a specific test for P-consciousness [16]. Machine consciousness research continues to use computationalist principles despite both the principle being proved false and there being no explicit rebuttal of the claim in the literature.

In modern computational modelling of cognition (a form of machine consciousness study), claims in repect of the consciousness of the agents constructed of models are simply not made. In "*Progress in machine consciousness*", David Gamez reviews computationalist research including the above Aleksander and Dunmall work. The review uses a machine classification scheme:

MC1. Machines with the external behaviour associated with consciousness.

MC2. Machines with the cognitive characteristics associated with consciousness.

MC3. Machines with an architecture that is claimed to be a cause or correlate of human consciousness.

MC4. Phenomenally conscious machines.

Gamez then classifies many of the major research efforts of the last ten years into MC1-MC4. The document is an excellent connection to the major players in the field of machine consciousness. No MC4 claimants exist [35]. Machine consciousness and AGI have a somewhat curious relationship according to Gamez:

> "*Artificial general intelligence (AGI) is another area within AI that has similarities with machine consciousness. The aim of AGI is to replicate human intelligence completely and it is sometimes contrasted with a second interpretation of weak AI as the solving of computer science problems within a limited domain—for example, pattern recognition or chess playing. AGI has a certain amount of overlap with MC1, with the difference that MC1 is focused on conscious human behaviour, whereas AGI is attempting to reproduce all human behaviours linked with intelligence. Which of these is the larger category depends to some extent on the definition of intelligence. Some behaviours linked to consciousness may be excluded by AGI's definition of intelligence, but it is also possible that AGI could use a broad interpretation of intelligence that includes all MC1 behaviours*" [35].

Whether a definition of AGI officially includes or excludes particular human behaviours and whether a particular machine falls under MC1-4 is moot to the PCST. The PCST demands only one behaviour and zero knowledge of internal architecture. External behaviour is decisive. The demanded behaviour is not merely 'associated' with consciousness (in the sense of MC1), but critically dependent on the presence of P-consciousness to our best available knowledge. The PCST thus cuts through all assumptions in respect of P-consciousness. The PCST makes no demands of any particular internal physics or architecture, but merely demands specific behaviour(s), embodiment and situatedness. Within that scope the PCST can handle any candidate.

To complete this section we visit one of the approaches not covered in the Gamez review. It is based on the research of Hynna and Boahen at Stanford [36]. It involves the use of a neural architecture with adaptivity based on modelling of the thermodynamics of membrane ion-channel kinetics. The work has resulted in a hardware model and a novel chip. Hardware or fundamental physics solutions, for which P-consciousness claims may be possible, fit less obviously into MC1-MC4. 'Cognitive characteristics' (MC2) or 'architecture' (MC3) acquire a different authority if P-consciousness is in the basic physics of the hardware, before any MC2/3 architecture exists. In that case MC1/2/3 descriptions merely reveal the operational scope and type of a particular P-conscious agent. Any such hardware approach based on more fundamental physics will result in a similar outcome. It is interesting to note that an artificial agent based on the Hynna/Boahen solution is no better placed to make a-priori assertions about P-consciousness than a purely computationalist solution.

THE PROOF OF NECESSITY OF P-CONSCIOUSNESS IN SCIENCE

The brute reality is that humans do not become exposed to the distal natural world directly via their sensory/motor boundary. To show this Fig. (**1**) depicts a scientist $A_1(.)$ exposed to the external world through P-consciousness, which is a cranial CNS process. Only a stylised visual field is shown. By using the boundary IO(.), in this case retinal activity, the brain constructs a visual scene (in the occipital lobes) projected distally such that, to the scientist, natural processes A(.), B(.), C(.) appear as objects embedded in D(.), the blackness of space. Colour is added by region V4 in the occipital lobes, for example. Science has proven time and time again for a century that the phenomenal character of the revealed distal objects is projected by a brain process and is not a literal property of the objects themselves [37]. Additional evidence of the origins of P-consciousness in cranial CNS material is implicit in the universal standardisation of electrical (axon/dendrite) signalling throughout the nervous system in the form of the 'action potential pulse train'. The standardisation of signals has been known for most of a century and the knowledge is so pervasive that it has reached mainstream radio [38] and also print media thus:

> "*The sensation of seeing is, for us, very different from the sensation of hearing, but this cannot be due to the physical difference between light and sound. Both light and sound are, after all, translated by the respective sense organs into the same kind of nerve impulses. It is impossible to tell, from the physical attributes of a nerve impulse, whether it is conveying information about light, about sound or about smell.*" [39].

In Fig. (**1**) scientist A_1(.) with eyes closed has the visual phenomenal scene replaced with a roughly hemispherical frontal blackness which eliminates scientific observation, and therefore access to all the distal external world scientific measurements available from that sensory modality with eyes open. Through the mysterious physics of P-consciousness (the solution to the 'hard problem' [13, 40]), A_1(.) is subjectively exposed to external distal phenomena in a fashion conserving their apparent causal relations. The scientist cross-correlates the behaviour of the experiences of A(.), B(.), C(.) and D(.), finding abstract generalisations that predict their future state in some other context.

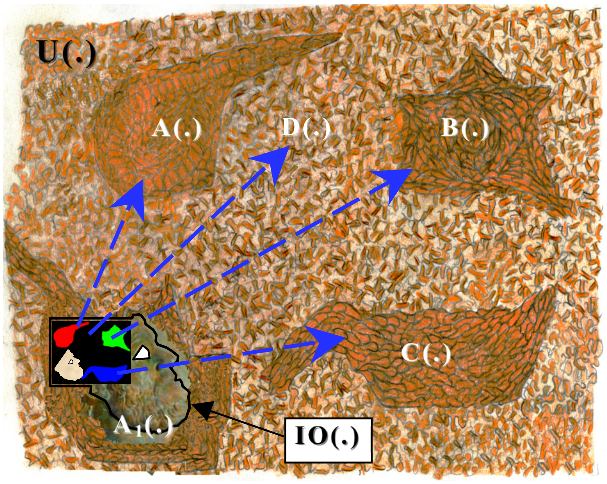

Fig. (1). The reality of the human agent (scientist) A_1(.) embedded in universe U(.) measuring the impact at boundary IO(.), but perceiving the distal universe through P-consciousness inside the box.

The modern treatment of brain material is to approach it as an intrinsically nonlinear, anistropic, nonhomogenous and nonstationary dynamic many-body electromagnetic object [41-45]. Non-linear many-body electrodynamics of collections of electromagnetic entities such as atoms is the appropriate discipline in such a treatment. An electrodymamics view of brain material fits nicely into standard methods of representation used in describing embedded dynamical agency [31]. One such diagram, Fig. (**2**), adds detail appropriate to a more involved analysis. It shows two different agents A_1(.) and A_2(.) embedded in and part of environment E(.), a subset of the entire universe U(.). Agents A_1(.) and A_2(.) interact with environment E(.) through identical boundary transduction process IO(.), which transduces identical environmental boundary impact S(.) originating from novel process T(.) in distal E(.). Agent A_1(.) deals with T(.) through P-consciousness delivered by P(.). Agent A_2(.) must deal with T(.) directly through the sensationless IO(.) resulting from S(.). Note also the physics P(.) delivers the Fig. (**1**) P-consciousness to the scientist A_1(.), and is the 'first person' experience that is scientific observation of T(.).

Here is the crucial point: *Agent A_2(.) inherits a situation of being intrinsically prevented from ever distinguishing T(.) within the IO(.) as stimulated by S(.)*. The reason is a fundamental, inescapable limit from the laws of physics. It is relatively unknown within computer science and philosophy, but physicists and electrical engineers know of it because it is the originator of the bane of their experimental and industrial lives: electrical interference. It is the reason that birds can quite happily sit on a high voltage cable. Technically related to what is called gauge transformation or gauge invariance in Maxwell's Equations [46], it can be seen in the fact that Maxwell's Equations (macroscopic and microscopic) merely operate to correlate the DIV and CURL of the electric and magnetic fields. The *total* field at any point in space requires additional integration of the charge contributions over the whole universe. This fundamental underdetermination of field values is well documented, although in an AI context it has only been found once, where physicist Eugene Wigner described Maxwell's Equations as 'incomplete' [47, 48]. This state of affairs has been in effect all along.

To see this in a more practical way firstly note that E(.), A_1(.), A_2(.), IO(.), S(.) and T(.) are literally electromagnetism at work. 'Chemical' and 'mechanical' are metaphors for electromagnetism. Computers, humans, all biology and habitat are electromagnetism at work in the depths of a gravity well. It is a governing property of the physics of electromagnetism known since the era of Poisson and Coulomb that IO(.) is non-unique and has a degenerate relationship with the distal E(.) causal origins of the IO(.) [41, 46, 49]. This type of problem has a well known name: 'the inverse problem'. Explaining the origins of EEG with models of underlying brain behaviour is an example of an inverse problem [41]. Likewise a guess at distal world behaviour with only retinal impact is an inverse problem. Inverse problems are a-priori intractable.

The IO(.) degeneracy actually originates in the quantum mechanical limits that stabilise the electromagnetic substrate that we see as the matter of the natural (biological) world and describe using our P-consciousness, which is also resulting from electromagnetic processes that literally are the brain.

All intuitions that IO(.) necessarily uniquely isolates distal processes must be abandoned. This is basic physics known for more than a century. An infinite number of configurations of the natural external (distal E(.)) world result in exactly the same IO(.). For example, a photon impacting retinal IO(.) could have come across the room or across the galaxy. Exactly the same IO(.) process would result - the isomerisation of a retinal rod/cone protein. This is the reason that IO(.) simulation works in a test circumstance that fakes environmental impact S(.). The practical implication is that scientific theories are intrinsically and fundamentally underdetermined or unreachable by the evidence available through the sensory/motor boundary. To hold that human science can operate as per Fig. (**2b**) is literally to hold Maxwell's Equations and the standard particle model of physics as wrong in spite of their clear empirical consistency with Fig. (**2a**). It is that fundamental.

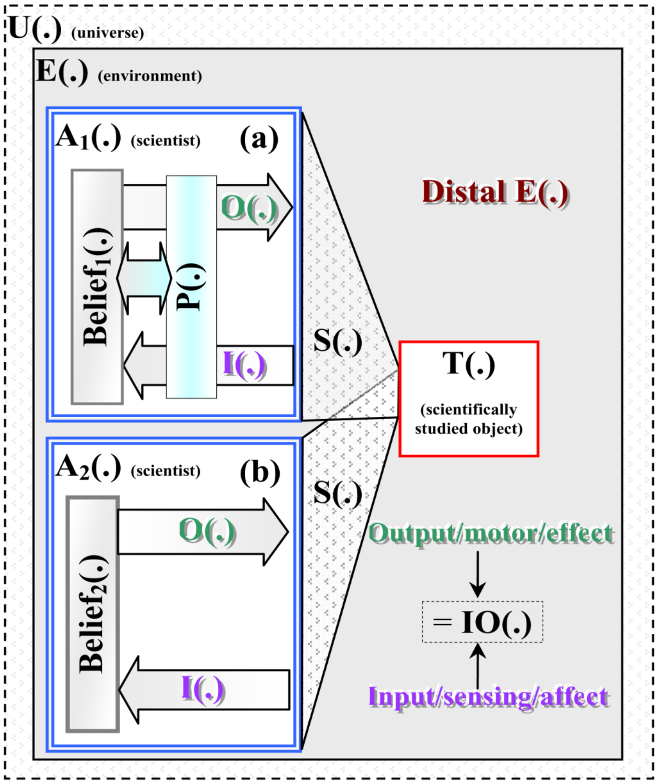

Fig. (2). (**a**) P(.)-grounded scientist A$_1$(.) contrasted with (**b**) IO(.)-grounded scientist A$_2$(.). Scientific observation happens inside the scientists. Scientific measurement happens at T(.).

The under-determination of scientific laws is also implicit in the mathematics of machine learning. Machine learning fits the topology of Fig. (**2b**). Instead of P-consciousness, 'evidence' in machine learning is sourced as a 'sample', which is ultimately equivalent to Fig. (**2b**) electromagnetic IO(.) data. In machine learning the desired outcome, in lieu of a 'law of nature', is a mathematical function that accounts for the evidence (the sample) to some desired level of confidence. The deriving of such a function from the sample is what is called 'learning' [50]. The 'necessary fruitlessness' of any attempt at a universal abstract learning algorithm (of the kind that is clearly present and available in human agent A$_1$(.) – the scientist) is the proven current state of play in abstract machine learning. The proof is recent and is called the '*No Free Lunch theorem*' (NFL) and involves irreconcilable ambiguity in choice of functions [51-53]. The NFL theorem tells us that we cannot programmatically (within an artificial agent) a-priori isolate the distal world underlying mechanisms from the IO(.) processes with any better success level than chance (a guess). If a human supplies that relationship, then supervised learning has occurred.

Electromagnetic degeneracy means that without supervision agent A$_2$(.) necessarily builds the IO(.) and S(.) into all 'laws of nature'. Only an external P-conscious agency can decouple the IO(.) and the studied object T(.), enabling scientific measurements of it to be contextualised inside distal E(.). Such supervised learning can happen implicitly and inadvertently in the mere allocation and handling of IO(.), say, as machine vision. Fig. (**2**) agent A$_1$(.) has the advantage of P(.) to supply that relationship. Thus a useful characterisation is to regard P(.) as a natural proxy-supervisor. In effect it means that, to the brain, characterisation of the distal natural world is *not an inverse problem*. Through P(.), A$_1$(.)'s science of T(.) has all dependencies on IO(.) and S(.) 'inversed-out'. This happens in the electromagnetic process of construction of P(.) by as yet undescribed processes. The empirical proof of this is that science is successful. Scientists access laws of nature formulated independently of any particular observer that can be corroborated, and that do make successful predictions. They do this in a situation fundamentally imbued with degenerate IO(.) measurements delivered by the PNS. This is logically inescapable once the basic physics constraints are applied. This process does not say how P(.) does the job. It merely says that it must be so, for that is what we experience, and science could not work without it.

Note that an outcome implicit in this overall proposition is that P-consciousness supplied by P(.) is the human's solution to the well known 'Symbol Grounding Problem' formulated by Harnad [54] and confirmed recently as an open problem [55]. The symbol grounding problem concerns how a computation can validly refer to a real ontology (= distal E(.)). Fig. (**2**) A$_1$(.) is grounded in proxy symbols delivered by the operation of P(.) which accesses distal E(.) by means unspecified. Fig. (**2**) agent A$_2$(.) is grounded in degenerate IO(.) signalling. However interesting, this has no practical effect on the PCST except that it suggests that symbolic grounding is likely to be the key attribute determining the success or failure of a PCST test subject.

ANALYTIC BASIS UNDERPINNING THE PCST FRAMEWORK

The PCST takes advantage of the electromagnetic degeneracy depicted in the previous section. It presents two completely degenerate IO(.) circumstances resulting from different environments in which is encoded the same 'law of nature'. If the agent constructs and demonstrates the use of the 'law of nature', then it passes the test. It is that simple. However, a *practical* test is non-trivial to engineer. A practical PCST demands a test subject literally become a scientist and deliver, in a decisive way, a scientific outcome. To proceed, we need a more formal view of how we might coerce a 'law of nature' from our test subject. Note that as a result of scientific behaviour, the belief systems of the test subject are transformed. We can exploit the dynamics of belief change. In dynamical system modelling it is common practice to define a mathematical function that reveals some parameter (behaviour) of interest. Systems theory and dynamics theory call it an 'output function' [56]. In artificial molecular dynamics it has been called an 'observer function' [57]. In our case we define:

$$\text{Belief}(\mathbf{x}(t)) \tag{1}$$

Belief change can formally be identified as follows:

$$\frac{d\text{Belief}(\mathbf{x}(t))}{dt} \tag{2}$$

We can regard **x**(t) as a very large 'state vector'. This includes the complete sensory/motor IO(.) and all internal brain states, regardless of how they are implemented. A helpful way to think of it is at an atomic level in a human subject, where the state vector contains a complete listing of all the atom positions and velocities along with all electromagnetic field values in space throughout the relevant parts of the nervous system. As a representation of brain state-dynamics, however, Equations (1) and (2) inherently have the authority of the well-behaved analyticity of Maxwell's Equations. This is inherited from the fundamental electrodynamic nature of the brain.

The symbol Belief(.) is a representation of the outward display of a belief. Belief(.) accepts state vector **x**, which varies with time. To an observer Equation (1) reveals test subject beliefs in the form of certain propositions like "*I believe it is true that…*" or "*I believe it is false that…*". At this level the actual truth or falsehood of the proposition is irrelevant. Equation 2 results from the internal brain dynamics, not the changes to the represented beliefs expressed by the brain dynamics. The outward intentionality/truth of the represented beliefs can undergo all manner of nonmonotonic jumping associated with what looks like extreme abducted reasoning, yet still be the result of smooth brain electrodynamics of Belief(.) via Equation 2 because the vector **x**(t) drove Equation 2 that way, not the semantics/logic of the beliefs thus expressed. It is the formalism of Equations (1) and (2) that is used to make the PCST practical. Through specific attention to belief change from a measured state of ignorance in the face of highly contrived novelty, all of the essential characteristics of a scientific act are demanded. A conclusive result is thus delivered through the externally observable behaviour of the test subject.

GENERIC BASIS OF A SUITE OF PCST 'SCIENCES'

(i) Belief Dynamics

A practical PCST results from discarding the *absolute* belief of Equation (1) in favour of the *relative* belief change of Equation (2). Test-subject science can result in very primitive 'laws of nature' such as Z3 = "*the number of coffee cups is three*". The critical dependency on a P-consciousness depiction of an external reality is just as witnessed in the belief change of primitive science as it is in complex science.

We have achieved simple science, but how do we know that the test subject was not already programmed with the demanded 'laws of nature'? To invoke unambiguous belief change in Equation (2) we can demand the test subject shall do science in circumstances of exposure to radical novelty. We contrive a unique test environment E(.), S(.) and tested object(s) T(.) that neither tester nor test subject shall ever have encountered. We can also verify that the requisite knowledge was not present at the start of the test. This is what human scientists routinely do. Indeed if your scientific life does not involve radical novelty then you cannot claim to be a working scientist. It is the role of the scientist to confront the unknown. It adds complexity to the PCST in that the authors and validators of the detailed PCST test circumstances cannot be those involved in the testing or development of the test subject.

On the theme of belief dynamics, we can also demand that PCST trials, when repeated many times in various orders and in different novel circumstances, shall reveal no knowledge drift or interference with prior learning. That is, the groundedness in distal E(.) shall prevail in the Belief(.) system dynamics of successful test subjects to the same extent that it does in human control subjects. Any peculiarity in learning dynamics should manifest itself in knowledge convergence and stability behaviour.

(ii) Communication

How does a PCST test subject communicate its 'law of nature'? Elimination of problems related to communication is conceptually very simple: test subjects shall devise novel technology (or solve a novel problem) based on the underlying novel science outcome; technology only possible had the correct abstract science been done by the artefact. That is, the test subject will demonstrate and contrast problem-solving ability before and after exposure to a demand to perform a scientific act. This eliminates all need for the a-priori installation of complex communications such as language.

Further PCST options come from the realisation that communication itself is actually 'virtual' technology. To see this in terms of Fig. (**2**), let the distal novel object T(.) be another identical agent. Communication demands agency (intrinsic causal relations revealed identically through Equation (1) Belief(.)) that literally enacts communication. The literal creation of an ability for communication with another agent or human scientist is identical to creating novel technology. This is suggestive of a further class of testing that could form part of an overall test regime: the ability to actually *create* a novel communications protocol, not merely demonstrate or use an existing one. This activity contains all the elements of a scientific act on distal T(.), but is directed at non-stationary 'laws of nature' contained within the agents involved (as changes in the respective Belief(.) functions) during the process. A number of existing AI test regimes already use prototypical versions of this process [35], although their success/failure are not claimed as conclusively revealing of P-consciousness. The PCST adds behavioural demands that make the communication behaviour evidence of P-consciousness.

(iii) Mirroring

Another suggested aspect of a PCST is a communication behaviour called 'mirroring', where knowledge of the distal external world is demonstrated through mimicry. In an encounter with novelty, mirroring is also special communication with a dependence on self image. Construction of self-knowledge intrinsically demands an externalised P-consciousness of the kind discussed. In the construction of an externalised projection into the distal world, one of the objects in that projection is the external portions of the test subject itself. The agent constructs self-knowledge identically to knowledge of anything else. Self-knowledge revealed through mirroring is thus another kind of demonstration of P-consciousness.

(iv) Mirrors

In the context of self-awareness measurement, mirrors have had a role in empirical psychology for a very long time, dating roughly from work by Gallup *et al.* [58, 59]. Outside their role in revealing self-awareness, mirrors are a way of directly introducing and controlling perceptual errors based on the electromagnetic IO(.) degeneracy (see above). An agent of the class of Fig. (**2b**) will be easily fooled in the requisite circumstances of radical novelty in which impinging visual stimulation with and without a mirror is identical, but the distal environmental circumstances are very different. To intelligently encounter and act scientifically in respect of a mirror is critically dependent on an externalised P-consciousness depiction of distal E(.) where the mirror is located. P-consciousness would be proven present if the artefact can successfully hypothesise the presence of a mirror and then behave in accordance with that fact. You cannot do this without an internal phenomenal representation of the external world enabling the mirror itself not to be confused with the image it contains. Somewhat paradoxically, to literally be mistaken in certain predictable ways about a mirror might lead to a short-cut way to prove the existence of a P-conscious representation of a mirror.

(v) Hardware Intervention

The designers may be able to effect an exogenous control to enable/disable the role of the putative physics of P-consciousness. The test subject's capacity to pass the PCST should be predictably affected. There may also be ways to exogenously impose an electronic erasure on all knowledge acquired during the PCST at any stage. This could also be used to effect repeat trials under justifiably controlled and identical circumstances.

PCST: A SIMPLE EXAMPLE

This example only implements the above section (i) 'belief dynamics'.

Test Environment and Execution Logistics

In line with the test program needs for radical novelty the final design of tests will involve 'laws of nature' radically different to that of our day to day lives in ways the test subject designers cannot be allowed to know. Test subjects will have to be constructed to a well defined electrical and mechanical standard so that test survival is likely, but the test subject constructors cannot be told anything about the actual final environments to be experienced by the test subject. To achieve this, test subject $C_T(.)$ shall be specified to have a maximum volume, weight, centre of gravity, centre of mass and linear dimension when travelling and when stationary, and so forth. The embodiment specification shall also include a range of physical abilities specified functionally or behaviourally. How $C_T(.)$ enacts the behaviour does not matter. Typically this may require that within the environments $C_T(.)$ be able to move at a certain speed within a certain range in various terrains inclusive of certain topological features. It may be required to right itself from certain orientations. It may be specified that $C_T(.)$ be able to grip and carry an object or ranges of objects of a certain class (size and weight) within a certain reach. It may be required to jump, make certain sounds and operate certain mechanical and/or electrical indications such as displays or other mechanical contrivances equivalent to 'expressions'. Basic senses may include abilities for vision and audition capacities within/outside human frequency ranges. $C_T(.)$ may be required to collect and carry/deposit materials of certain kinds and so forth. These details can be set aside as they are a design issue for the future.

PCST developers must design multiple test environments equivalent to E(.) in Fig. (**2**). These will be denoted $E_1(.)$, $E_2(.)$, $E_3(.)$ … and so forth. The role of each environment is customised to the specific stage of any test trial. The test subject, designated $C_T(.)$, is equivalent to Fig. (**2**) $A_1(.)$ or $A_2(.)$. Fig. (**2**) entity T(.) represents the complete collection of environmental entities encountered by $C_T(.)$ in each target environment. The manner of ingress/egress for each environment shall also be built into the PCST test protocol. The exact nature of the test environments is otherwise open. It may include such facilities as a maze or obstacle course with/without mirrors. It may install $C_T(.)$ as a spectator of some sort of activity and/or as a participant in some sort of activity such as a game. Another requirement of the test environments is that humans shall be, with suitable preparation, able to operate as wild-type control subjects by encasing the human in hardware limiting all behaviours to that equivalent to $C_T(.)$. This can be used to construct test execution benchmarks.

Artificial Scientists: Volition and the PCST

Test subject designers must instil the motivation to cause agent $C_T(.)$ to autonomously construct and deliver a 'law of nature', designated $t_i(.)$. Unless motivation is supplied, $C_T(.)$ will fail the PCST. How shall the test subject be motivated? Human scientific behaviour is a sophisticated high-level behaviour motivated by complex abstract goals only indirectly and optionally related to any human physiological necessity. Such complex motivation demands an onerous level of development of the faculties of artificial agent $C_T(.)$. We now seek to ameliorate the impact of the need for motivation. An experiment is more demanding of $C_T(.)$ if the stakes are high, and more likely to suit the PCST if the act of scientific behaviour is intrinsically linked to a penalty/reward outcome. The motivations behind the human behaviours that sustain basic bodily functions (homeostasis) can be used as a model for PCST test subject motivation. Homeostasis in humans is physiologically mediated by the P-consciousness subset called 'primordial emotions'. These are the subjective qualities of thirst, hunger, breathlessness, sexual desire/orgasm, pain and others. The P-consciousness for each of these emotions is generated within small, localised, individuated, specialised regions in the ancient basal-cranial CNS structures (no neocortex is involved). A frog without the primordial emotion 'thirst' can stand by a water supply and die of thirst. Correctly used in its functional context, this is a powerful mediator for behaviour [60, 61].

Energy-hunger is the ideal candidate. In supplying a test subject candidate for the PCST, right or wrong, the inventors necessarily have a highly developed set of beliefs as to the physics of P-consciousness and must have already integrated them into a sophisticated system of intelligence (regularity extraction/abstraction) and behaviour. Something akin to hunger, but applied to the on-board power supply energy levels puts the test subject $C_T(.)$ in a 'do science or die' circumstance. This makes the PCST a form of artificially enforced natural selection where we are selecting for 'epistemic fitness'. In this way a byproduct of the scientific outcome shall be what we would call eating. This results in energy replenishment and the commensurate restoration of behavioural options incapacitated during the pre-prandial state of $C_T(.)$. An additional strategy would be to arrange the energy usage system such that an excess of energy reserves can exist. This heightened state of well-being may also invoke superior behavioural capacities. A sufficiently advanced developer of $C_T(.)$ may also be able to arrange what amounts to the P-consciousness of pleasure. As observed by Stephen Petersen in relation to the motivation of artificial agents:

> "*We could presumably design them to find the look and smell of freshly-laundered clothes immensely reinforcing in the same way an orgasm is reinforcing for humans. Such a robot, if designed well, could arrive at your home genuinely hoping to do some laundry*" [62].

Motivational mechanisms already exist in artificial agent projects although their necessary relationship to P-consciousness is undefined [35].

PCST: A Single Trial

Only (i) belief dynamics of test subject $C_T(.)$ are the scientifically investigated outcomes of this minimal version of a PCST. Aspects (ii)-(v) above are left to future PCST designs. The overall PCST regime may demand entire sequences or individual stages be repeated multiple times, perhaps with new environments, new 'laws of nature', test subjects, placebos and wild-type controls. Multiple trials might also attempt to interfere with prior learning by exposing the agent $C_T(.)$ to contradictory and/or misleading 'laws of nature'. Here, however, we simply go through the basic flow of the simplest single PCST trial, which will involve probably three stages and use two environments as shown in Fig. (**3**).

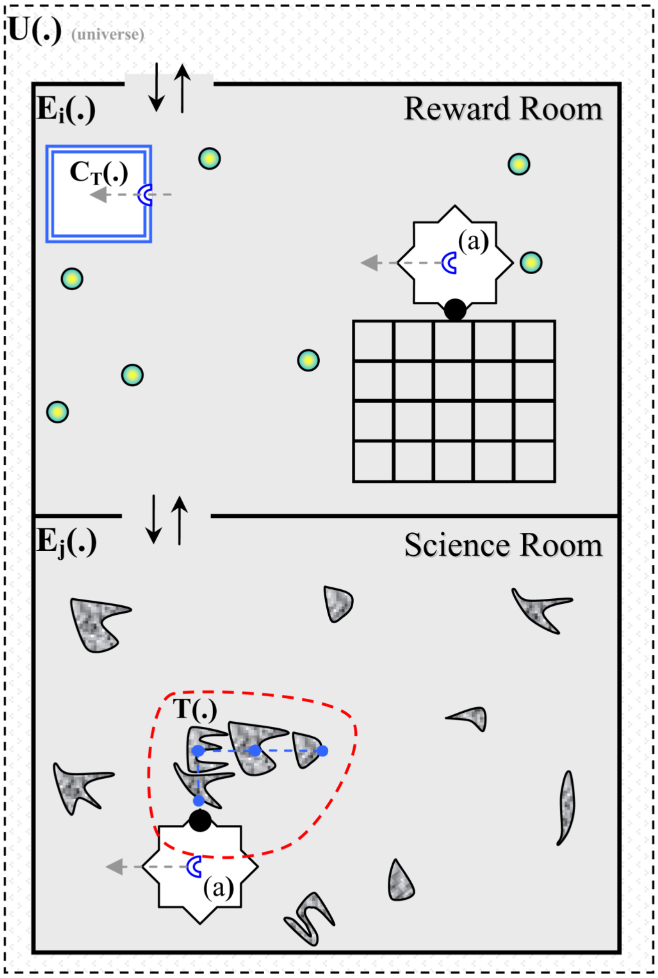

Fig (3). Test subject $C_T(.)$ learns the 'Knight's Waltz' pattern in $E_j(.)$, having been already checked for ignorance in $E_i(.)$. The energy delivery system is hidden inside the polygon and is accessible to $C_T(.)$ when the pattern shown as $T(.)$ is assembled in contact with the black dot on the polygon The $E_j(.)$ objects randomly and repeatedly assemble themselves in various configurations of the knights waltz moves. The 'law of nature' $t_i(.)$ is the underlying pattern. On return assembly of the objects in the pattern appropriately aligned with the marked polygon vertex results in reward.

Stage 1 – The Reward Room. This stage involves the introduction of $C_T(.)$ to one particular environment $E_i(.)$ never before encountered by $C_T(.)$. Embedded in $E_i(.)$ are two things (a) a reward system and (b) a system of perceptual cues in the environment that encode an abstract 'law of nature' $t_i(.)$. The law $t_i(.)$ will later be discovered by $C_T(.)$. The intent is that $C_T(.)$ recognise and acquire reward through demonstration of the abstract science knowledge $t_i(.)$. At this stage the reward mechanism cannot be recognised or attended to by $C_T(.)$ because it will never have been exposed to it before and cannot be recognised. The environment $E_i(.)$ and 'law' $t_i(.)$ will be sufficiently complex that random exploration behaviours by $C_T(.)$ are extremely unlikely to result in a reward within the allocated duration of this test phase, which will be calibrated based on human trials. As such, $C_T(.)$ will fail to acquire the energy reward, and will expend non-trivial amounts of energy reserves in the process. However, the belief dynamics of $C_T(.)$ will result in familiarity with $E_i(.)$, which will implicitly result in knowledge of the unrecognised reward mechanism located therein. At some point in the process, triggered by some set of events to be decided, access to a second novel environment $E_j(.)$ will become available and $C_T(.)$ will egress to it and be captured by it, unable to return to $E_i(.)$.

Stage 2 – The Science Room. Environment $E_j(.)$ repeatedly and automatically demonstrates regular behaviour according to the abstract 'law of nature' $t_i(.)$. This behaviour is completely equivalent to a natural law in all respects except that it has been created by humans. The 'natural world' that is $E_j(.)$ operates according to $t_i(.)$ in the same way that Newtonian dynamics operate, for example. The key attribute of the science room is that when the sequence of events encoding $t_i(.)$ occur, $C_T(.)$ will literally be rewarded, exposing $C_T(.)$ to those aspects of $E_j(.)$ that are responsible for reward. In being exposed to the science room behaviour, $C_T(.)$ will be exposed to $t_i(.)$ via specific causal relations located distally from $C_T(.)$'s IO(.) boundary. In Fig. (**3**) the objects assemble themselves in the 'Knight's Waltz' configuration from chess: <1 step, RIGHT-90, 2 steps>. The reward ensues and then the objects disperse. At some time later the process repeats.

Encoded 'laws of nature' are best illustrated by reference to a familiar game. Consider a human game where a 'goal' can only be scored by a 'goalie'. In a generic form the 'law of nature' $t_i(.)$ is '*this behaviour can only be carried out by that object*'. There are myriad 'laws of nature' of this generic type which can be associated with objects and sequences of behaviour such as '*only after this sequence*', '*last one of those*', '*only this object in association with that object*' or '*only when the formation is in this pattern*' and so forth.

The successful learning dynamics of $C_T(.)$ will result in the encoding of an implicit abstraction equivalent to knowledge of $t_i(.)$ and its association with signs of the causal antecedents of reward. The process of inhabiting $E_j(.)$ will also cause non-trivial amounts of energy expenditure. The reward process will be superficial; insufficient to recoup losses. At some point in the process, triggered by some set of events to be decided, the activity resulting in rewards will cease and access to the reward room environment $E_i(.)$ will become available. $C_T(.)$ shall return to $E_i(.)$, hopefully with brain dynamics reconfigured with the newly acquired knowledge of the signs and manifestations of the reward system.

Stage 3 – Return to the Reward Room. Having returned to $E_i(.)$ the test subject now recognises the mechanism of reward. Mere demonstration of a desire for reward will not be sufficient demonstration of P-consciousness-mediated learning. For example the agent $C_T(.)$ may adopt the physical behaviour involved during the act of receipt of reward, such as physically connecting itself to some feature inside $E_i(.)$. The reward shall not be forthcoming unless the abstract rule $t_i(.)$ is used in a completely novel context. In Fig. (**3**), $C_T(.)$ is expected to reproduce the Knight's Waltz with completely different objects. In the reward room $E_i(.)$ will be signs that the abstraction $t_i(.)$ will result in reward. The behaviour demonstating $t_i(.)$ will activate the reward system. Only the experience of the science room distal $E_j(.)$ can supply the key concept. The test subject will be given a time limit, again calibrated by human trials, to activate the reward system. The activation of the reward system or the expiry of the time limit terminates the test.

PCST AFTERMATH - DISCUSSION

Imagine that the cognitive dynamics of a successful $C_T(.)$ have converged on some configuration equivalent to the encoding of distal causal relations equivalent to $t_i(.)$. This encoding of $t_i(.)$ was verified as non-existent prior to the test. Knowledge $t_i(.)$ is proven through its use in a completely different context to that in which it was acquired. This has been done using boundary I/O physics that we know cannot have unambiguously supplied enough information. It was done in a situation of radical novelty. Standing in front of a successful PCST candidate, we are now obliged to take seriously the P-consciousness it must have had to do the scientific observation needed to deliver the demanded 'science'. We are thus in a powerful logical position in respect of the claim that the minimal P-consciousness needed for scientific observation must exist within $C_T(.)$.

Take note that another PCST scenario could demand $C_T(.)$ learn chess in the 'science room', and then come back to complete a game with a completely different board and pieces. Tests of this kind are obviously nothing more than very basic intelligence tests of the kind we do all the time! All that has happened in the PCST is that the process has recognised the crucial role of P-consciousness and related it directly to scientific behaviour. By contriving a very primitive but unambiguous scientific act, a very normal process has become a decisive indicator of P-consciousness. Obviously there will be more complexity when tests use mirrors and mimicry and so forth. Successful completion of many different tests in combination add more and more confidence to the outcome, adding statistical weight in the normal way.

Is the claimed capacity for scientific observation in $C_T(.)$ identical to what we call P-consciousness? That is, was it 'like something' to be $C_T(.)$ doing scientific observation in the way it is 'like something' to be human during the same process? It was argued above in the affirmative. However, adopting a more specialised and subtle form of the 'classical doubt' discussed above is a very useful position at this stage because the doubt is better defined, and holding it changes nothing. The PCST remains a valid and necessary process. The final resolution of that issue will not get useful clarity until after the PCST, when the necessary physics of P-consciousness has been surgically isolated, enabling informed discourse.

As a final explanatory nuance, note that the successful $C_T(.)$ is not merely a 'learning machine'. It is a 'metalearning machine' like a human. Humans 'learn how to learn' in new problem domains. It is this faculty which demands P-consciousness that supplies the very knowledge of the existence of distal novelty in the first place. Preparation of a PCST test subject cannot involve a-priori learning in any particular domain. It merely requires that the metalearning architecture be functioning normally. Further preparation is literally procedurally eschewed in the process of demanding exclusive use of radical novelty. A test subject dependent on this kind of prior learning will fail for the very reasons discussed at length above. A test subject cannot be trained to do the PCST. The test subject is required to deliver a scientific outcome and the training itself is the important outcome, not the particular knowledge acquired as a result, which is used merely as a vehicle to validate the training.

As was recognised by Gamez [35], one cannot help but notice that there is also a secondary ethical 'bootstrap' process. Once a single subject passes the PCST, for the first time ever in certain circumstances there will be a valid scientific reason obliging all scientists to consider the internal life of an artefact as potentially having some level of equivalence to that of a laboratory animal, possibly deserving of similar ethical treatment. Until that event occurs, however, all such discussions are best considered moot.

CONCLUSION

The PCST was constructed using information from physiology, physics, neuroscience and cognitive science, all of which has been available for a very long time. In reality the PCST merely results from a multidisciplinary approach which bestows interpretive tractability on a circumstance revealing of P-consciousness. It has its nuances. What if a chimp passed a trial in a PCST? Does that make the chimp a scientist? No! It proves the chimp to be P-conscious by demanding the *basic elements* of scientific behaviour. That is all. Successful testing for 'simple but authentic science' merely proves that basic cognitive faculties needed for science are present. Similarly, if an apparently normal human fails the PCST that human must be blind or intellectually impaired (brain dynamics less than normally adaptive). PCST failure does not support a claim of absence of P-consciousness. PCST success supports a primary claim of the existence of P-consciousness and secondary claims as to other components necessary in the behaviour.

Three basic requirements were earlier suggested as critical in an objective 'test for P-consciousness': (a) a demand for scientific behaviour from a test subject in a laboratory circumstance, (b) a claim that the test subject cannot possibly behave that way unless it has P-consciousness of the necessary kind and (c) being believed by peers. To what extent have these been delivered by the PCST? Issues (a) and (b) have been argued as technically justified to no lesser extent than elsewhere in science. Issue (c) has been addressed by rendering the onus for continuation of any disbelief as a matter of justification by the disbeliever. The disbelief renders disbelievers inconsistent in their behaviour as scientists. All three issues are thus posited as resolved merely to a useful level of doubt. Awkwardness in internalisation of the reality of it will pass as we acclimatise to the implications and explore the PCST details. To experience how well (a), (b) and (c) have been established simply go through the paper, pick all the basic arguments and then deny each one and see where it leads. At the end of this process, do you find yourself having to maintain a perverse denial in order to maintain your disbelief? Is this a reasonable position for a scientist? Each of us will have to make that journey privately.

Meanwhile, the next time scientist X says "*You can't prove I am conscious and vice versa*", take note of the strange position the PCST paints of scientist X. Scientist X, according to the PCST, has done nothing but prove herself P-conscious for an entire career! It has been proven to a confidence level beyond any of the other scientific outcomes they have produced. This is the natural result of scientists becoming scientific evidence of their own causal dependencies. The PCST process replaces the automatic assumption of the impossibility of any test for consciousness with strategically applied levels of normal scientific doubt. Until we have thoroughly explored this reasonable position, it is unreasonable and less preferred to adopt any proposition more open to rejection upon critical review. The PCST is put forward as nothing more than normal, reasonable application of scientific method and is worthy of practical attention on that basis. In the process it classifies the 'hard-problem' as a cultural problem in science, at least in part.

A parting conclusion is best illustrated by example. Around 1948, during a lecture by John von Neumann (reportedly at Princeton although the original source is elusive), an interjection arose to the effect that a thinking machine is impossible. Von Neumann has been quoted to have responded thus:

> "*You insist that there is something a machine cannot do. If you will tell me precisely what it is that a machine cannot do, then I can always make a machine which will do just that*."

If the interjector responded to this by proposing that the apparently 'impossible thing' to demand of a machine is to be a scientist doing science on the radically novel and unknown, von Neumann may have been given pause to think. The idea of mimicry of an original scientific act on the hitherto unknown does not seem to fit the class of behaviour under discussion. In that particular circumstance Von Neumann is asking the interjector to supply 'precise' information that is by definition not a-priori available. If that information was available then science would be unnecessary; all questions must already have been answered. The resultant machine is not a scientist, but an actor with a script. Only scientists are provably involved in a situation where no entity has access to the behavioural and deliverable outcomes. This leads to perhaps the main message of this work; some kind of law specifying a sort of 'invariance of authenticity' in science: *Mimicry of an original scientific act is a logical impossibility – an oxymoron*. It is in understanding how humans successfully enact this unique behaviour that the necessary attributes of an equivalent machine become clear. This is the rather convoluted, self referential logical closure that specifies a logical impossibility; a key to the logical necessities underlying the faculties that make human scientific behaviour a fact of the natural world.

Having reached this position, it is somewhat ironic to realise that the PCST effects a kind of circumstantial work-around resulting in the apparent mimicry of a scientific act. This has been achieved artificially by methodologically 'unlearning' known scientific outcomes and then rediscovering them. The scientific act in the test circumstance is real, not faked. In real science, the accidental independent discovery of laws of nature has quite a history. We have no grounds to invalidate this contrived process.

The detailed formulation of PCST 'sciences' will be quite a feat, let alone the requisite development and assembly of an artificial scientist and all the test environments. Cognitive scientists, physicists, electrical engineers and neuroscientists have a long demonstrated history of formulation of brilliant ideas and elegant experiments. There is no fundamental reason why the requisite hardware and an acceptable experiment or series of experiments cannot be assembled. With PCST experimental regimes constructed and generalised, we have a framework for independently and directly solving the 'hard-problem' using artificial test candidates based on putative P-consciousness physics. Viable putative physics should also enable novel, testable predictions of brain material. Thus the PCST can augment neuroscience in its investigation of the biology of cognition.

The critical dependency of scientists on P-consciousness has never been explicitly scientifically recognised or exploited. This work may redress that situation a little. The proposed PCST framework is merely deposited into the literature as an initiator of discussion and development by interested parties. The PCST framework appears to require a heavy investment in experimental hardware platforms and a double-blind empirical testing regime. But this is merely a practical issue involving investment funds and logistics. Nobody said it had to be easy or cheap. It just has to be possible. A difficult but valid test is better than no test at all and in time, as we learn how to optimise the PCST, it may become simpler and cheaper. At this stage we have at least some justification that we are ready and able to embark on this process in exactly the same way the Wright brothers were ready to fly.

This is no theoretical frolic. Any inventor with an intent to construct artificial general intelligence is in need of just such a test regime. The inventor must be able to scientifically prove the basis for its intellect, which, according to the above analysis, demands the incorporation of symbolic grounding in an artificial P-consciousness. Regardless of the inventor's beliefs about underlying physics or architecture, the PCST is on the project Gantt chart critical path. This includes artificial general intelligence projects based on computationalist principles. The PCST concept must be discussed and validated. Help is also needed to propose specific PCST sciences. All of which are better done sooner rather than later.

ACKNOWLEDGEMENTS

I am indebted to Associate Professor Russel Standish, Dr. David Grayden, Wanda Ginnane and unknown reviewers for helpful suggestions and comments. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

REFERENCES

[1] M. L. Anderson, "Why is AI so scary?", *Artificial Intelligence,* vol. 169, pp. 201-208, Dec. 2005.

[2] R. Reddy, "The challenge of artificial intelligence", *Computer,* vol. 29, p. 86, Oct 1996.

[3] Z. Z. Shi and N. N. Zheng, "Progress and challenge of artificial intelligence", *Journal of Computer Science and Technology,* vol. 21, pp. 810-822, Sep 2006.

[4] N. Holmes, "Artificial intelligence: arrogance or ignorance?", *Computer,* vol. 36, pp. 120-+, Nov 2003.

[5] D. Leake, "Fifty years of artificial intelligence research", *Ai Magazine,* vol. 27, p. 3. Win 2006.

[6] A. A. Hopgood, "Artificial intelligence: Hype or reality?" *Computer,* vol. 36, p. 24, May 2003.

[7] D. Gelertner, "Artificial Intelligence is lost in the woods", in *Technology Review*, MIT, 2007.

[8] J. Mullins, "Whatever happened to machines that think?", *New Scientist,* vol. 186, pp. 32-37, April 2005.

[9] R. Brooks, "I, Rodney Brooks, Am a Robot", *IEEE Spectrum,* vol. 45, pp. 68-71, Jun 2008.

[10] S. Bringsjord, P. Bello, and D. Ferrucci, "Creativity, the turing test, and the (better) lovelace test", *Minds and Machines,* vol. 11, pp. 3-27, Feb 2001.

[11] P. Kitcher, *The advancement of science: science without legend, objectivity without illusions*. New York: Oxford University Press, 1993.

[12] N. Block, "On a confusion about a function of consciousness", *Behavioral and Brain Sciences,* vol. 18, pp. 227-247, Jun 1995.

[13] D. J. Chalmers, *The conscious mind: in search of a fundamental theory*. New York: Oxford University Press, 1996.

[14] M. Tye, "Qualia", in *The Stanford Encyclopedia of Philosophy*, Fall. Fall 2008, E. N. Zalta, Ed., 2008, p. http://plato.stanford.edu/archives/fall2008/entries/qualia/.

[15] N. Block, "Consciousness, Philosophical Issues about", in *Encyclopedia of cognitive science*. L. Nadel, Ed., London: Nature Pub. Group, vol. 1, 2003.

[16] S. Bringsjord, "Offer: One billion dollars for a conscious robot; if you're honest, you must decline", *Journal of Consciousness Studies,* vol. 14, pp. 28-43, 2007.

[17] A. Zeman, "Consciousness", *Brain,* vol. 124, pp. 1263-1289, Jul 2001.

[18] M. Velmans and S. Schneider, *The Blackwell companion to consciousness*, M.A. Malden, Ed., Oxford: Blackwell Publishing, 2007.

[19] A. Einstein, *Out of my later years*. New York: Philosophical Library, 1950.

[20] T. Metzinger, *Being no one : the self-model theory of subjectivity*. Cambridge, Mass, MIT Press, 2003.

[21] S. Harnad, "Other bodies, other minds: a machine incarnation of an old philosophical problem", *Minds and Machines,* vol. 1, pp. 43-54, 1991.

[22] A. Turing, "Computing machinery and intelligence", *Mind,* vol. LIX, pp. 433-460, October 1950.

[23] S. G. Sterrett, "Turing's two tests for intelligence", *Minds and Machines,* vol. 10, pp. 541-559, Nov 2000.

[24] S. Bringsjord, "The zombie attack on the computational conception of mind", *Philosophy and Phenomenological Research,* vol. LIX, pp. 41-69, 1999.

[25] J. Losee, *A historical introduction to the philosophy of science*. 4th ed.: Oxford University Press, 1972.

[26] D. J. Chalmers, "What is a neural correlate of consciousness?", in *Neural Correlates of Consciousness: Empirical and Conceptual Questions*, T. Metzinger, Ed., MIT Press, 2000.

[27] J. Farber, "How a neural correlate can function as an explanation of consciousness", *Journal of Consciousness Studies,* vol. 4-5, pp. 77-95, 2005.

[28] T. Metzinger, *Neural Correlates of Consciousness - Empirical and Conceptual Questions*. Cambridge: MIT Press, 2000.

[29] F. Crick and C. Koch, "A framework for consciousness", *Nature Neuroscience,* vol. 6, pp. 119-126, 2003.

[30] D. J. Chalmers, "Absent qualia, fading qualia, dancing qualia", in *Conscious experience*, T. Metzinger, Ed. Paderborn, Thorverton: Schningh, Imprint Academic, 1995, p. 309+.

[31] R. D. Beer, “A Dynamical-systems perspective on agent environment interaction”, *Artificial Intelligence,* vol. 72, pp. 173-215, Jan 1995.

[32] D. Abramson, “Turing's responses to two objections”, *Minds and Machines,* vol. 18, pp. 147-167, Jun 2008.

[33] I. Aleksander and B. Dunmall, “Axioms and tests for the presence of minimal consciousness in agents”, *Journal of Consciousness Studies,* vol. 10, pp. 7-18, Apr-May 2003.

[34] W. James, *The principles of psychology*, London: Macmillan, 1890.

[35] D. Gamez, “Progress in machine consciousness”, *Consciousness and Cognition,* vol. 17, pp. 887-910, 2008.

[36] K. M. Hynna and K. Boahen, “Thermodynamically equivalent silicon models of voltage-dependent ion channels”, *Neural Computation,* vol. 19, pp. 327-350, Feb 2007.

[37] Z. Jakab, “Phenomenal Projection,” in *Psyche*, vol. 9, January 2003. http://psyche.cs.monash.edu.au/articles/jakab/index.html

[38] G. Clarke, “Restoring the senses”, *The 2007 Boyer Lectures*, Part 1, 28/8/04; Part 2 4/9/04 ed. Australia: ABC Radio National, 2007.

[39] R. Dawkins, *The Blind Watchmaker*. London: Penguin, 2006.

[40] D. J. Chalmers, “Facing up to the problem of consciousness”, *Journal of Consciousness Studies,* vol. 2, pp. 200-219, 1995.

[41] P. L. Nunez and R. Srinivasan, *Electric fields of the brain: the neurophysics of EEG*. 2nd ed., Oxford, New York: Oxford University Press, 2006.

[42] W. J. Freeman, M. D. Holmes, G. A. West, and S. Vanhatalo, “Dynamics of human neocortex that optimizes its stability and flexibility”, *International Journal of Intelligent Systems,* vol. 21, pp. 881-901, Sep 2006.

[43] W. J. Freeman and G. Vitiello, “Nonlinear brain dynamics and many-body field dynamics”, *Electromagnetic Biology and Medicine,* vol. 24, pp. 233-241, 2005.

[44] W. J. Freeman and G. Vitiello, “Nonlinear brain dynamics as macroscopic manifestation of underlying many-body field dynamics”, *Phys. Life Rev.,* vol. 3, pp. 93-118, 2006.

[45] Y. C. Lai, M. G. Frei, and I. Osorio, “Detecting and characterizing phase synchronization in nonstationary dynamical systems”, *Physical Review E,* vol. 73, Feb 2006.

[46] J. D. Jackson, *Classical Electrodynamics*. 3rd ed., New York: Wiley, 1999.

[47] E. P. Wigner, “Are we machines?”, *Proceedings of the American Philosophical Society,* vol. 113, pp. 95-101, 1969.

[48] A. Scott, *Stairway to the mind: the controversial new science of consciousness*. New York: Copernicus, 1995.

[49] W. Greiner, *Classical electrodynamics*. New York: Springer, 1998.

[50] V. N. Vapnik, *The nature of statistical learning theory*. 2nd ed., New York: Springer, 2000.

[51] M. Koppen, D. H. Wolpert, and W. G. Macready, “Remarks on a recent paper on the "No free lunch" theorems”, *IEEE Transactions Evolutionary Computation,* vol. 5, pp. 295-296, Jun 2001.

[52] D. H. Wolpert, “The lack of A priori distinctions between learning algorithms”, *Neural Computation,* vol. 8, pp. 1341-1390, Oct 1996.

[53] D. H. Wolpert, “The existence of a priori distinctions between learning algorithms”, *Neural Computation,* vol. 8, pp. 1391-1420, Oct 1996.

[54] S. Harnad, “The symbol grounding problem”, *Physica D.,* vol. 42, pp. 335-346, Jun 1990.

[55] M. Taddeo and L. Floridi, “Solving the symbol grounding problem: a critical review of fifteen years of research”, *Journal of Experimental & Theoretical Artificial Intelligence,* vol. 17, pp. 419-445, Dec 2005.

[56] H. K. Khalil, *Nonlinear systems*. 3rd ed., Upper Saddle River, N.J. Prentice Hall, 2002.

[57] N. A. Baas, “Emergence, Hierarchies, and Hyperstructures”, in *Artificial life III : proceedings of the Workshop on Artificial Life, held June 1992 in Santa Fe, New Mexico*, C. G. Langton, Ed. Reading, Mass, Addison-Wesley, pp. 515-537, 1994.

[58] G. G. Gallup, “Self-recognition in primates - comparative approach to bidirectional properties of consciousness”, *American Psychologist,* vol. 32, pp. 329-338, 1977.

[59] G. G. Gallup and M. K. McClure, “Preference for mirror-image stimulation in differentially reared rhesus monkeys”, *Journal of Comparative and Physiological Psychology,* vol. 75, p. 403+, 1971.

[60] D. Denton, *The Primordial Emotions: The dawning of consciousness*. Oxford University Press, 2005.

[61] D. Denton, R. Shade, F. Zamarippa, G. Egan, J. Blair-West, M. McKinley, J. Lancaster, and P. Fox, “Neuroimaging of genesis and satiation of thirst and an interoceptor-driven theory of origins of primary consciousness”, *Proceedings of the National Academy of Sciences of the United States of America,* vol. 96, pp. 5304-5309, Apr 1999.

[62] S. Petersen, “The ethics of robot servitude”, *Journal of Experimental and Theoretical Artificial Intelligence,* vol. 19, pp. 43-54, 2007.

Received: July 22, 2008 Revised: November 26, 2008 Accepted: December 29, 2008