On the Solvability of the Mind-Body Problem

Jan Scheffel
Department of Fusion Plasma Physics, KTH Royal Institute of Technology,
SE-100 44 Stockholm, Sweden

Abstract

The mind-body problem is analyzed in a physicalist perspective. By combining the concepts of emergence and algorithmic information theory in a thought experiment employing a basic nonlinear process, it is shown that epistemically strongly emergent properties may develop in a physical system. Turning to the significantly more complex neural network of the brain it is subsequently argued that consciousness is epistemically emergent. Thus reductionist understanding of consciousness appears not possible; the mind-body problem does not have a reductionist solution. The ontologically emergent character of consciousness is then identified from a combinatorial analysis relating to universal limits set by quantum mechanics, implying that consciousness is fundamentally irreducible to low-level phenomena.

1 Introduction

Understanding consciousness is a central problem in philosophy. The literature produced through the centuries, relating to the 'mind-body' - problem, is also vast. A subset of some 2500 articles on theories of consciousness can be found in PhilPapers (2018). An apparent difficulty lies in the fact that while we normally seek scientific understanding from a reductionist perspective, in which the whole is understood from its constituents, consciousness has for millions of years naturally evolved into an extremely complex system with advanced high-level properties.

The theoretical difficulties we have faced strongly suggest that fundamentally new ideas are needed for the mind-body problem to reach its resolution. In this work it is argued that emergence, combined with results from algorithmic information theory and quantum mechanics, is such an idea. The meaning of these concepts will shortly be discussed; we may here briefly state that emergence relates to complex systems with characteristics that are difficult or impossible to reduce to the parts of the systems and algorithmic information theory concerns relationships between information and computing capacity. We reach the conclusion that the mind is epistemically emergent, which by definition implies that the mind-body problem cannot be solved reductionistically. Reductionistic understanding of the subjective aspects of consciousness, like introspection and qualia, therefore does not appear possible. The concept of the 'explanatory gap' (Levine, 1983) is thus justified.

McGinn, in his influential work "Can we solve the mind-body problem?" (McGinn, 1989), also concludes that the mind cannot be understood, but on other grounds. He focuses on the ability to understand phenomenal consciousness (Block, 1995), and finds that we humans, because of 'cognitive closure' are not able to solve this 'hard problem of consciousness' (Chalmers, 1995). With the reservation "the type of mind that can solve it is going to be very different from our" McGinn does not fully exclude that consciousness can be given some kind of explanation, an optimism not supported in this work.

We will present a thought experiment, featuring a process shown to produce an emergent property in the epistemic sense. From a subsequent comparison with the neurological

functions of consciousness it is argued that consciousness is epistemically emergent. The ontologically emergent character of consciousness is then discussed in light of its complexity considered as a global system. Chalmers (2006) finds, on intuitive rather than on formal grounds, that the mind is 'strongly emergent'; a term used here in the same meaning as 'ontologically emergent'.

Definitions are important in this work. There are at least two reasons for this. The first is that several aspects of the concepts of consciousness, in particular emergence, are often used in different ways by different philosophers, neuroscientists and others. This may be understandable on the basis of that consciousness, not least semantically, is an elusive concept. The problem is rooted in its unique character, causing attempts for a definition to contain circular elements of some kind. The influential early characterisation of Locke (1690) "consciousness is the perception of what passes in a man's own mind" suffers from reference to the subjective term 'perception'. Nagel's (1974) characterisation "there is something that it is like to be that organism - something it is like for the organism" has gained popularity, although "is like" refers back to the subject itself, that is to consciousness. A more exhaustive and recent discussion of possible criteria for and meanings of conscious states can be found in Van Gulick (2014). However, either of the above formulations sufficiently catches the subjective components of consciousness that are referred to in this work, thus we here consider phenomenal consciousness. When we discuss other key concepts, attempts will be made to render the treatment more precise, in some cases using formalisations from physics and mathematics.

A second reason for the need for clear definitions is simply that binding arguments requires precision (Carnap, 1950). The consequence of such specifications may of course be that the definitions of some philosophers are excluded; the results should be seen in this perspective.

We begin by discussing what requirements must be placed on a solution of the mind-body problem. It is then argued that such a solution cannot be found. The core of the argument is that consciousness is epistemically emergent, precluding understanding of consciousness in a reductionist sense. Interestingly, this conclusion has bearing on the problem of free will since if it really were the case that a true theory of consciousness could be designed, then free will seems excluded. Free will implies that the mind is epistemically emergent, a circumstance that may deserve more attention in the literature. The reason is that if an individual's behaviour would be computable or could be simulated, this behaviour would be predictable and thus not free.

The paper ends with conclusions.

2 What is required of a solution to the mind-body problem?

The goal of the mind-body problem research is to find a theory that explains the relationship between mental and physical states and processes. The sub-problem which by far has attracted the most interest concerns the question how consciousness can be understood. We may initially ask the question: what is required from an adequate theory?

Chaitin (1987) has clarified the meaning of the necessary requirement that a theory must be inherently less complex than what it describes; in his terminology it must to some extent be 'algorithmically compressible' in relation to what it should explain. Let us illustrate this by an example. A relationship $y = f(x)$ has been established to explain a phenomenon, but the precise dependence is not known. A series of experiments that generate N data points (x_i, y_i); i:1…N has thus been performed. Clearly, a polynomial $Y(x)$ of degree N-1 (N coefficients) can always be fitted to be drawn through all the data points in an xy-diagram. Is it a theory? The answer is no, for the simple reason that $Y(x)$ does not

explain anything; it is always possible to draw a polynomial of degree N-1 exactly through N data points. Had we instead adapted a polynomial of lower degree through all the points, say a second order polynomial through 10 data points, then we would have a theory worthy of the name; it predicts more than it must. That it is algorithmically compressible means that it can be formulated using fewer bits of binary information than those required for $Y(x)$. Simply put: *a proper theory must be simpler than the phenomenon it describes, otherwise it does not explain anything*.

In this work we will make use of the discrete logistic equation for later comparisons with the neurological processes that form the basis of consciousness. This equation can be formulated as the discrete recursive relation $x_{n+1} = \lambda x_n(1\text{-}x_n)$, n:0…n_{max}, where the positive integer n_{max} can be chosen freely. The discrete logistic equation then iteratively generates new numbers x_{n+1} for increasing values of numbers n. The parameter λ and the start value x_0 must first be selected. We can now ask: is there an explicit theory for the value x_{n+1}, that is is there a function $u(k)$ which satisfies the relation $x_k = u(k)$, being algorithmically compressible as compared to repeated use of the iterative relationship $x_{n+1} = \lambda x_n(1\text{-}x_n)$? Of course we can form $x_1 = \lambda x_0(1\text{-}x_0)$, $x_2 = \lambda x_1(1\text{-}x_1) = \lambda\lambda x_0(1\text{-}x_0)(1\text{-}\lambda x_0(1\text{-}x_0))$ and so on. This latter route is not feasible; for large n we will find that the symbolic expression for x_{n+1} becomes extremely complex; this way of searching $u(k)$ does not result in a valid theory. Alternatively formulated: the binary bits needed to represent the characters of these symbolic terms is at least of the same order as the bits representing the numbers x_1, x_2, x_3... themselves. Unfortunately, it can be shown that the question we posed must be answered in the negative; no matter how we try, it is not possible (except for a very few values of λ) to derive a theory, that is a compact, explicit expression for $u(k)$.

The cause of the problem is that the discrete logistic equation is a *nonlinear recursive equation*. Let us, for a moment, instead consider the simpler *linear recursive equation* $x_{n+1} = A+\lambda x_n$, n:0…n_{max}, where A is a constant, for which the general term x_k can be derived in *explicit* form simply as $x_k = x_0\lambda^k + A(1\text{-}\lambda^k)/(1\text{-}\lambda)$ for $\lambda \neq 1$ and as $x_k = x_0 + Ak$ when $\lambda = 1$. The formal solution is expressed using only a few mathematical symbols; it is thus algorithmically compressible (may be represented by fewer digital bits of information) as compared to the solution x_k obtained iteratively by forming $x_1 = A+\lambda x_0$, $x_2 = A+\lambda x_1 = A+\lambda(A+\lambda x_0)$, $x_3 = A+\lambda x_2 = A+\lambda(A+\lambda(A+\lambda x_0))$ and so on. This explicit solution was analytically available because of the low complexity involved in the solution of linear equations as compared to nonlinear. Furthermore, the solution for x_k is derived mathematically by using well-known axioms and theorems; consequently we can *theoretically explain* the values x_k for the linear recursive equation.

We have here employed examples from mathematics, but the reasoning applies generally when we seek any kind of formal explanation or theory for a phenomenon. As a result, a theory cannot explain consciousness if it relates to systems of the same level of complexity (like other minds). Understanding is only reached from theories that are less complex than consciousness itself and relate to already established knowledge; in other words they should be algorithmically compressible in relation to consciousness.

3 Emergence stands in the way

The emergent character of consciousness is persistently debated in the philosophical literature (Kim 1999, Kim 2006, Chalmers 2006). We will here argue that consciousness is both epistemically strongly emergent and ontologically emergent; our definitions of these concepts are provided below. Consciousness thus has features that are not reducible to the properties of its components. The standard expression 'not reducible to' expresses that the characteristics of the low-level components, taken separately, of the phenomenon

are insufficient to establish high-level properties. By 'low-level' and 'high-level' we refer to the parts of and integrated wholes of a system or phenomenon, respectively.

This conclusion is central to the mind-body problem since it settles the issue of the 'explanatory gap' (Levine 1983); an unbridgeable gap exists between the theories we can formulate on the basis of the basic physiology of the brain and the subjective, cognitive function of consciousness. A consequence is that behaviour of consciousness is in principle unpredictable, a relationship being of importance when addressing the problem of freedom of will. It should however be noted that even if consciousness is an emergent, unexplainable property of the brain there is in principle nothing that precludes design of artificial consciousness. The possibility of imitating evolution is always open. We now turn to investigate the emergent character of consciousness.

3.1 Definitions of epistemical and ontological emergence

Emergence as a concept emerged in the literature in the late 19th and early 20th centuries, mainly through the philosopher John Stuart Mill, the psychologists George Henry Lewes and Conwy Lloyd Morgan and the philosopher Charlie Dunbar Broad, although already Aristotle had touched upon the subject in his Metaphysics. See Corning (2012) for a concise review. The literature in the field has since expanded significantly and criticism against careless use of emergence has been put forth (Goldstein 2013). Thus we find it essential to define the varieties of emergence discussed in the present work.

Epistemically strong emergence is defined in the following way: *A high-level property is epistemically strongly emergent with respect to properties on low-level if the latter form the basis for the high-level property and if the theories that describe the low-level properties cannot predict properties at high-level.*

Ontological emergence, in turn, can be defined by replacing "*if the theories that describe the low-level properties cannot predict properties at high-level*" with "*if it is not reducible to properties at low-level*". It may be argued that any property or behaviour in a physicalist world would be reducible to low-level properties since we assume that the physical is all there is. But this is not what is meant by reduction; supervenience does not imply reducibility (see also Francescotti (2007) for a clarification of the relation between supervenience and emergence). An ontologically irreducible property, if it exists, could not be determined by its low-level-properties or behaviour; it could not be characterised by a statistical or law-like behaviour in relation to its low-level components. Loosely formulated it can be said that its behaviour comes as a surprise to nature. This distinction is crucial and we will indeed see that even if causality holds, there are systems where extreme complexity can, in an ontological sense, 'shield' the dynamics of a high-level phenomenon from that of its associated low-level phenomena. An important consequence is that these systems are uncontrollable in principle. Furthermore we will, for simplicity and clarity, follow Schröder (1998) and preferentially refer to emergent *properties* rather than to emergent things, behaviour, processes or laws.

The requirements for ontological emergence are indeed harder to satisfy than those for epistemic emergence; the former relates to intrinsic properties of the system rather than to knowledge about and theories for the system. As defined above, ontological emergence implies epistemical emergence. Note also that epistemic emergence is here defined in an *a priori* sense (high-level properties should be predictable from those of low-level) rather than in the weaker *a posteriori* sense (high-level properties should be explainable from those of low-level). This differentiation is, however, not important for the analysis presented here. What we are looking for is the possible existence of relations between particular high-level properties and low-level properties or states. In the epistemical case

this amounts to the existence of theories that connect the two levels. In the ontological case we are concerned with establishing whether a direct relation between the two levels is possible in principle. Assuming, as we do here, supervenience and causality we know that complex properties of the mind, like consciousness, have evolved. But evolution is complex and does not, as we will se, guarantee that ontological reduction is possible.

The term *epistemically weakly emergent* is frequently used for systems that can be simulated on a computer but otherwise would be characterized as epistemically strongly emergent. This definition will be adopted here as well.

3.2 Emergence and understanding

Emergence precludes reductionistic understanding. In an era where physicists talk about 'theories for everything', emergence tends not to be a welcome concept. It may thus be of interest to consider whether limits for understanding the world manifest themselves in other ways. Epistemically, we may consider at least four categories of phenomena and properties in nature in a physicalist perspective. The first two categories are at a basic level:

I. *Brute facts*. These are indeed also referred to as 'facts without explanation'. To this category belong elementary concepts like matter, time, space, charge, particle spin and even the physical constants of nature. Of the latter, some 20 are presently believed to be independent of each other.

II. *Laws of nature*. Examples are Newton's laws of motion, the law of gravitation, Coulomb's law, relativity theory and the Schrödinger equation.

The low-level phenomena associated with these two categories cannot be understood in a traditional reductionist manner; they simply are. There are no simpler entities that could aid in an explanation of them. According to the debated Anthropic principle, the constants of nature should be tuned to some extent for there to be a universe at all where conscious minds can appear to discuss these matters. It has been shown, however, that there is an allowed window of variation for most constants of nature and thus their precise values, as they appear in nature, cannot be motivated or understood.

From a theoretical standpoint categories I and II are perfectly consistent with that any scientific reductionistic (non-circular) theory requires a basic set of unprovable axioms.

Turning to the two high-level categories, we have:

III. *Phenomena that are reducible to brute facts and laws of nature*. Most phenomena belongs to this category, by virtue of causality.

IV. *Phenomena that are not brute facts or laws of nature, nor reducible to these*. These are the emergent phenomena.

In light of categories I and II, emergence is only one of several obstacles for understanding the world. Emergence is sometimes criticized as an irrelevant construction. To a large extent this appears to be related to supervenience. By assuming supervenience, the notion that all processes of the world including consciousness have one-to-one physical counterparts, emergence may seem like a contradiction. A strong focus of the present work is to show how emergence can arise even when assuming supervenience, as well as to present evidence for instances of both epistemical and ontological emergence.

Arguments for emergence are predominantly related to *complexity*, but there are exceptions. See for example Silberstein and McGeever (1999) and Gambini et al (2015) for a discussion of ontological emergence and non-reducibility related to *basic* phenomena in quantum mechanics. It is however questionable whether this approach is fruitful. If, for example, two entangled particles in a quantum mechanical interpretation

should be assigned emergent properties, in consequence also Newton's third law should be regarded as an emergent property of nature. The latter law states that for every action, there is an equal and opposite reaction in terms of forces. The law cannot be assigned to individual particles or bodies (low-level); it is manifested only when two or more of these are interconnected (high-level). But we do not refer to Newton's third law as proof of an emergent law or property; we simply call it a law of nature. Emergence is mainly related to complexity.

Categories I-III are directly *observable* in nature. We accept the reality of brute facts and natural laws and we can usually identify combinations of these as category III phenomena. A falling snowflake, temporarily caught by the wind, exemplifies the latter. But category IV phenomena are not identified this way. We are usually accustomed to trying to interpret and understand the phenomena we encounter to the extent that occurences of category IV phenomena are typically regarded as potential category III phenomena. Existence of category IV phenomena is hard to comprehend because of our natural insistence to interpret and understand on the basis of brute facts and natural laws. Consequently the fact that we have developed an advanced level of natural science understanding without introducing the concepts of emergence has sometimes lead to the erroneous conclusion that emergent phenomena and properties are of no relevance.

To summarize, whereas we speak of explanation and understanding of category III phenomena, these rely on acceptance of category I and II phenomena as mere facts. The latter do not have reductionistic explanations. In this perspective, emergent phenomena of category IV are not the only obstacles for our understanding of the world.

The main focus of this work is on emergence related to complexity. The brain features about 80 billion nerve cells (neurons), each connected to thousands of other nerve cells via synapses. A reductionistic model of the mind must be able to handle a corresponding complexity. As we have just discussed, algorithmic information theory implies that 'models' or 'theories' that cannot be algorithmically compressed to a complexity lower than that of the data they describe do not measure up. It is however not entirely clear how emergent properties arise. It would be of great help if we could actually point to a relevant example. Our approach will be to, using a thought experiment, provide an example of a system featuring emergent properties, being related to neural networks of the brain but with lower complexity. We will subsequently proceed to address emergence in relation to consciousness.

3.3 The Jumping robot

Our thought experiment is the following. Let us imagine a number of robots that are deployed on an isolated island. All robots are designed in the same way. They are programmed to be able to freely walk around the island and perform certain tasks. The robots can communicate with each other and are also instructed to carry out their duties as effectively as possible. If a robot becomes more efficient by performing a certain action, it should 'memorize' it and 'teach' the other robots the same skill. Let us concentrate on the behaviour of one of these robots and call the thought experiment 'the Jumping robot'.

In order to support the robots to move about freely, their movement patterns are partially determined by the discrete logistic equation just described. The iterative equation $x_{n+1} = \lambda x_n(1-x_n)$ generates new numbers x_{n+1} in the interval [0,1] when x_0 (also in the range [0,1]), and λ are set. These numbers affect how the robot should coordinate its joints, muscles and body parts, but the robot is programmed only to use information leading to safe motion without falling. Let us put $\lambda = \lambda_0$, where λ_0 is a number slightly less than 4. It can then be shown mathematically that, for almost any choice of x_0, a chaotic sequence in

the interval [0,1] is generated already for moderately large n. If we consider consecutive x_k, x_{k+1}, x_{k+2} and so on, these numbers will seem completely random. And, as discussed in section 2, there is no algorithmically compressible explicit expression $x_k = u(k)$ that would provide a theory and understanding of the behaviour. An interesting and important fact is, however, that the sequence of numbers actually is deterministic; each number in the sequence is unambiguously defined by the former and so on in a long chain.

Now assume that it would be of great value if the robots could perform jumps without falling. An attempt is thus made to provide a robot with this property. From a large number of x_0 values different sequences of numbers are generated, using the discrete logistic equation, in the hope that one of these sequences would correspond to movements which when combined would result in a controlled jump by the robot. We ignore here that the procedure is obviously cumbersome; the complexity is partly caused by the fact that the robot consists of a large amount of joints, muscles and other bodyparts that should be coordinated, partly by our ignorance as to what movements the robot would need to perform for a successful jump and partly by that the discrete logistic equation does not allow control of the movements. After numerous unsuccessful attempts the task is thus given up; the robots cannot be taught to jump.

Instead now initiate robots with random x_0 and leave them to themselves for some time on the island, after which we return. To our surprise, we now find that several of the robots make their way not only by walking, but also by jumping over obstacles. We cannot explain how one or more of the robots aquired the new property; no theory is to be found. This would entail finding a relation $x_k = u(k)$ for the logistic equation, which is excluded. We could neither simulate the behaviour. If so, this would have been an example of epistemically weak emergence. Thus the theories that describe the low-level robot phenomena cannot predict behaviour at high-level. The robot's ability to jump is an epistemically strongly emergent property. A main point here is that the emergent ability to jump *per se* is both fully comprehensible to us as well as fully plausible in the sense that we can imagine that a certain sequential use of joints, muscles and bodyparts indeed may accomplish this behaviour, at the same time realizing that some kind of chance or evolution beyond our modelling capacity was required in the light of the complexity involved. There is no magic involved in the process, rather the behaviour is similar to that of random mutations in the genome of an individual organism, producing improved characteristics through evolution. The behaviour in this thought experiment, however, may not be ontologically emergent since the robot's capacity to jump would appear to be reducible to the motions of its finite number of parts. Similar conclusions about the emergent properties of nonlinear systems have been reached by other authors (Silberstein and McGeever, 1999).

No account of precisely how the robots acquire the skill to jump is given in this thought experiment. Actually, whether evolution or chance is involved is not relevant for the fact that a well known property, to jump, has emerged among these particular robots. We cannot compute or design this property, and still it emerges. We could, of course, design other types of robots, differently built and wired without a connection to the logistic equation, that indeed can jump. But emergence should always relate to *specific* systems; just as water molecules are much less likely to appear in a mixture of hydrogen and nitrogen than in a hydrogen and oxygen mixture.

3.4 The epistemically emergent character of consciousness

What then is the relevance of this epistemically strongly emergent system for the mind-body problem? It could be argued that we can make detailed studies of a jumping robot, simply ignoring how it reached its emergent state, in order to understand its functions and

presumably build copies that perform the same movement patterns. We would simply map and reconstruct all the detailed states of the robot involved in the dynamics. Maybe we could also build consciousness in a similar manner?

Building a full robot copy, including its complete physical design and built-in software, would not solve the problem, however. The copy would feature the same complexity as the original robot, including the irreducible logistic equation. As we have seen, a procedure of this kind does not satisfy the criteria for a theory and does not constitute a path to understanding. The same conclusion holds if we make a simulated copy of the robot on a computer; the iterative use of the logistic equation in the simulations would amount to a one-to-one copy of the full robot system. A second possibility would entail identifying some reduced pattern of robot movements (’reverse engineering’) that still would lead to stable jumping performance for this particular type of robots. This procedure, however, is not likely to succeed for several reasons. The main obstacle resides in the logistic equation itself. Since, in spite of extensive and sophisticated efforts, we were unable to design a jumping robot, is quite unlikely that there exists a sequence of robot movements, providing stable jumping, simpler than the one generated by the logistic equation. A second difficulty is that it is quite conceivable that the patterns for jumping are non-intuitive, being difficult to reveal for this reason. When the computer program AlphaGo beat the world number one ranked player in the game Go in 2017, an analysis of the game showed that the computer often chose to use non-intuitive and seemingly questionable unorthodox moves. AlphaGo reaches its excellence through engaging its neural networks in machine learning techniques, foremost by playing an extensive set of games against other instances of itself (Silver et al 2016). A parallel can be drawn to evolution which does not design but rather ’tries’ different possibilities which are then measured in a survival context.

In conclusion, the Jumping robot provides an example of the epistemical thesis that what can be built cannot always be understood. A proper theory for the properties of the Jumping robot, being algorithmically compressed in relation to what it explains, stands little chance to be developed.

The roundworm *Caenorhabditis elegans* features only 302 neurons. The interconnections of all its neurons have been mapped. This mapping is an interesting first step towards understanding more complex neuronal networks. A bird’s brain has some 100 million neurons. Some argue that in a brain of this complexity, there are signs of basic characteristics of consciousness. The smallest primate brains feature about 500 millions and monkeys about 10 billion. The human brain features 80 billion neurons with some 16 billion interconnected in the cerebral cortex, being the primary area associated with consciousness. The question arises whether the conscious properties of the brain, such as thoughts and emotions, can be understood from a theoretical mapping of these neurons.

We thus turn to investigate the potentially emergent character of consciousness. The human brain works along vastly more complex paths than the discrete logistic equation, controlling the Jumping robot. Its neurons communicate, in brief, as follows. Via so-called dendrites, each neuron can obtain electrochemical signals from tens to tens of thousands (on average 7000) neighbouring neurons. The contributions from these signals are weighted in the neuron’s cell body to an electrical potential; the so-called membrane potential. When this reaches a certain threshold, the neuron sends out a pulse, the action potential, along a nerve fibre termed axon, which in turn connects via synapses and dendrites to other neurons. The outgoing signal from a neuron has the form of a spike rather than a continuous, nonlinear function of the incoming signal. Thus our choice of the discrete logistic equation rather than its continuous counterpart for the robot thought experiment. Neurons fire typically in the range of 1-100 signals per second (Maimon and Assad 2009) but also at higher frequencies (Gittis et al 2010), with signal lengths of at

most a few ms and with speeds of up to 100 meters per second. The behaviour varies between neurons. For networks of neurons, functions called sigmoids, with *S*-shaped dependence on the input signals, provide realistic activation function models of the relation between neuron firing and membrane potential.

Communication within the neural network of the brain thus occurs nonlinearly and discretely with a complexity vastly exceeding that of the simple logistic equation. Furthermore, evidence has been presented that even the activity of individual neurons play a role for conscious experiences (Houweling and Brecht 2007). Consequently it may be assumed that a reductionistic theory for consciousness should take into account firing of individual, or small clusters of, neurons. In the example of the Jumping robot it was the functional value, generated by the logistic equation, that was of interest. For consciousness, it is mainly the interspike intervals and patterns of neuronal action potentials that are of significance rather than the amplitudes of the action potentials.

A determining factor for the neuronal firing behaviour is the membrane potential in relation to the treshhold for firing. This threshhold is individual for each neuron and sensitively determined by the weighted contribution from thousands of other neurons through its dendrites. We saw, in the Jumping robot thought experiment, that the behaviour of the simple logistic equation is algoritmically incompressible. A proper, algorithmically compressed, theory for phenomenal consciousness involving thousands of networking neurons, obeying the behaviour outlined above, thus certainly seems out of reach. This argument will be strengthened in the next section, as ontological emergence is considered.

Summarizing, we have compared the problem of understanding consciousness, that is predicting it from its low-level neural components, with the problem of understanding the Jumping robot. We have argued that the robot's ability to jump is an epistemically emergent property; it was impossible to construct a proper theory that explained how it could jump. The major problem lies in that its behaviour is partly attributed to an iterative nonlinear function; the discrete logistic function. There is no possibility to construct an algorithmically compressed theory for the logistic equation except for some singular special cases. Consciousness as a property of the mind is, as we just discussed, grounded in the low-level behaviour of a huge network of neurons with interspike intervals that can be modelled with a similar iterative theory as for the logistic equation. The robot can jump, and we know this ability stems from the coordination of its low-level components. The brain can be conscious, and we know consciousness stems from the activity of its low-level neurons. But since there is no middle ground, no possibility to reduce collective neural activity, generating consciousness, into a compressed theory, we are facing an explanatory gap between individual neuronal activity and consciousness.

Thus there is strong evidence that consciousness, and similarly subconsciousness, is *epistemically strongly emergent*. In the same way that the Jumping robot's behaviour cannot be described reductionistically, the properties of consciousness cannot be epistemically related to the behaviour of its low-level neurons, it cannot be represented in a reductionistic theory. In consequence, the mind-body problem is reductionistically unsolvable.

It may finally be noted that mental processes involve an additional, well known, complexity, not necessarily related to emergence: they cannot be scientifically related to measurable properties in the same manner as movements of the robot parts are linked to its *externally* measurable ability to jump. The phenomenological, or subjective, conscious properties of the mind are predominantly accessible *internally* or subjectively, not from externally distinguishable physical states. Our focus is here on emergence, so we will not dwell further on this difficulty.

3.5 The ontologically emergent character of consciousness

We may now ask whether consciousness is also *ontologically* emergent; are the properties of consciousness irreducible to the lower level states and processes that form the basis of consciousness, the ones that consciousness supervene on? The meaning of 'reducible to' for this question needs to be illuminated. Let us return to the example of the Jumping robot. The property to be able to jump was not deemed ontologically emergent for the reason that in an objective meaning this property was an option that was reducible to the system, although its details were unknown to us. By 'objective' we refer to that the various possible sequences of numbers being generated by the discrete logistic equation, of which at least one potentially lead to jumping behaviour, correspond to an amount of information that is manageable *in principle*. This latter statement demands clarification, since we now have made contact with the consequences of quantum mechanics and information theory for ontological properties.

It has been shown (Lloyd 2002, Davies, 2004) that the information storage capacity of the universe is limited by the available quantum states of matter inside the causal horizon. The latter is the distance, limited by the finite speed of light, outside which no events may be causally influential. It is found that the order of 10^{120} bits of digital information may be contained within this horizon. The fact that this 'ontological information limit' is an estimate is not essential; what matters here is that information storage capacity is universally limited to a magnitude of this nature. A property that is associated with a complexity transcending some 10^{120} bits of digital information can be characterized as ontologically emergent since then there is no possibility, even in principle, to 'reduce' it to the low-level phenomena on which it is based. This property is physically irreducible. The point made here is that quantum mechanics, which provides the basis for physicalism, limits the number of achievable states in nature and thus also implies ontological restrictions. This circumstance is usually absent in discussions of ontological emergence. It should be noted that 'ontology' is used in this work in the traditional, philosophical sense and *not* as a reference to properties or interrelationships between entities used in computer science and information science.

An ontological information limit may be hard to digest and a natural reaction would be to claim that real processes and properties simply develop in the world, without any relation to its computational capacity. Responding to this, we must realize that our quest, the topic of this paper, is an epistemologic one although we are investigating the ontological behaviour of the world. This entails using theoretical concepts like computability. Applying these to the world, we find that certain phenomena feature a complexity to the degree that their appearance comes as a 'surprise', their complex behaviour is not immediately given by the state of the world.

The situation is analogous to that of the world of mathematics. Here Zermelo–Fraenkel set theory constitutes an axiomatic system for generating the truths (theorems) of standard mathematics including algebra and analysis. The system of axioms is quite limited, but the number of theorems that can be deduced is vast, covering most of mathematics. However, in 1931 Gödel showed that there exist true propositions of this system that cannot be proven inside the system. These propositions can, however, be proven by adding axioms, that is by stepping outside the system. The epistemical point we are making here is that the physical states of processes and properties in the world may imply subsequent states, the occurrence (truth) of which are not given by present states. We cannot step outside the world to decide whether the former will appear or not, but we can however ascertain that their potential appearance is undecided. In mathematics, Gödel-undecidable propositions usually feature a complexity related to infinite sets that, although they may be dealt with, have a power extending outside of the system. Turning to the real world, its limitations in

representing properties come not from dealing with infinity but from its discrete character, governed by quantum mechanical laws.

It may be helpful to discuss a specific example of an ontologically emergent system in nature. To this end, we note that emergence rarely is associated with the results of human activities, with *design*, but rather with *evolution*; the development of nature. Evolution has through natural selection access to a tremendous diversity of degrees of freedom and features a huge potential to generate emergent systems. An example from chemistry is myoglobin, an important oxygen binding molecule found in muscle tissue (Luisi, 2002). Here 153 amino acids are interconnected in a so-called polypeptide chain. Since there are 20 different amino acids, the number of possible combinations of chains amounts to the enormous number $20^{153} \approx 10^{199}$, which corresponds to a number of digital bits much larger than 10^{120}. Myoglobin thus features an ontologically emergent property; the molecule is, in terms of its optimized high oxygen affinity, not reducible to its low-level constituents. It could only evolve, it could not be designed.

It could possibly be argued that in order to free memory, by employing some efficient algorithm, only the most relevant data relating to each computation need be stored. This is, however, not a successful path. In nature, changes do not come and forces do not act instantaneously. All effects of interactions in nature, that is changes of state, are in fact due to combinations of the four basic forms of interaction through exchange of particles called bosons. This interaction indeed takes a finite time; the lower limit is given by a key relationship in quantum mechanics called Heisenberg's uncertainty principle (Lloyd, 2002). The limiting time is proportional to Planck's constant and inversely proportional to the system's average energy above the ground state, which for a one kilo system means that no more than $5 \cdot 10^{50}$ changes of state are possible per second; a theoretical limit for a quantum computer. In the entire visible universe, which dates back some 14 billion years and has a mass of about 10^{53} kg, not more than about 10^{121} changes of quantum states have occurred. Although this is a huge number it is not infinite. The universe's 'capacity to act and compute' is thus limited (Wolpert 2008). Since several quantum states are involved in each computation of the oxygen affinity of a polypeptide chain, it is clear that quantum mechanics sets a universal limit, prohibiting reduction to the amino acid low-level components.

We will now argue that consciousness is ontologically emergent. The line of reasoning is the following. First we specify what it takes for consciousness to be ontologically emergent. Next we specify physical assumptions made for the neural network of the brain. It is subsequently argued that the information associated with conscious states, in relation to low-level neural states, exceeds the ontological information limit discussed above. Consciousness is thus found to be ontologically detached from its low-level neural states, whereupon ontological emergence follows.

The argument proceeds as follows. Referring to the previously stated definition, consciousness is ontologically emergent if it cannot be reduced to the properties or behaviour of its low-level states. This, in turn, means that no explicit relation can be established, not even in principle, between consciousness and the activity of the neural network that generates it. Hence we want to find out whether such a relation, that reduces consciousness to its low-level states, can be expressed or not.

At this point we need to specify the details of the neural network that we assume as the basis for consciousness, out of which some will be used for our argument. Causality and supervenience, in the sense that the properties of consciousness correspond to certain configurations of low-level neural states, are both assumed. Whereas the human brain contains some 80 billion neurons, a lower number of interacting neurons appears sufficient for consciousness, perhaps of the order of one billion neurons. We also assume

that it is only particular configurations of these, in terms of their interrelations, and certain temporal neuronal activity that generate phenomenal consciousness. Each neuron is, on average, connected to about 7000 other neurons, affecting its behaviour. Individual neuronal activity is assumed important (Houweling and Brecht 2007) for consciousness. Furthermore we assume that there is a lower time limit for collective neural activity where consciousness cannot be upheld. It has been shown that after stimulation of neural brain activity there is a delay before the individual becomes conscious of it. Experiments (Libet 1993) suggest that this limit is of the order 0.5 seconds.

We will now consider the amount of information associated with a conscious state, in terms of its relation to its low-level components, the neurons. Our approach will be to make a lower estimate by employing a very crude and simplified model and determine its implications. Thus we start by assessing the information processes associated with an individual neuron *k*. We assume it is, through its axon (output) and dendrites (input), connected to *K* neurons. Since the primary action of the neuron in its network contribution to consciousness is to fire action potentials at certain rates and in certain patterns, it is natural to focus on information relating to whether the neuron, upon integration of its input, reaches the threshold potential for firing or not. The associated membrane potential is found from adding its present electric potential to the integrated contributions from other connected neurons. The threshold potential for firing is as mentioned earlier nonlinear, often assumed as *S*-shaped, function of the cell potential. For simplicity we here model this function as a third order polynomial. Mathematically, the state Z_{n+1}^k of a neuron k at time T_{n+1} can be symbolically modelled as $Z_{n+1}^k = a^k + b^k s + c^k s^2 + d^k s^3$, where a^k, b^k, c^k and d^k are constants, unique for each neuron, and $s_n = \sum_{i=1}^{K} Z_n^i$, a sum over the contributions from neighbouring neurons. Its ”state” is decided by whether it has fired ($Z_{n+1}^k = 1$) in the time interval $[T_n, T_{n+1}]$ or not ($Z_{n+1}^k = 0$). Here ”n” is an integer, where $n = 0$ denotes the initial time T_0 and the maximum number of time intervals of interest is denoted by $n = N$. Note that self-dependence on the previous state is included.

Of primary information theoretical interest is the number of characters that are needed for expressing Z_{n+1}^k in terms of its dependence on signals from neighbouring neurons. Using a computer math program (in our case Maple), it can be shown that the number of characters required to express Z_{n+1}^k scales approximately as $10\cdot(3K)^{n+1}$. Transformed to binary code (extended ASCII, for example), each character corresponds to 8 digital bits of information. Thus for $K = 7000$, the information content associated with state Z_{n+1}^k entails some $8\cdot10\cdot(21000)^{n+1}$ digital bits. Assuming an *average* interspike interval of 0.1 s, this is an amount of information that does not reach the ontological limit within the N intervals that need be accounted for. This is concluded from estimating $n_{max} = N = 0.5/0.1 = 5$, where 0.5 s is assumed for the total time T_c of neuronal activity required for conscious mind processes to take place. However, and importantly, account need also be taken for the significant number of neurons that fire at interspike intervals towards the estimated *minimal* interval of about 0.001 s (Softky and Koch 1992, Paré and Gaudreau 1996). We now find that even at average interspike intervals of 0.018 s (thus for $N = 0.5/0.018 \approx 27$) the number of bits representing Z_{n+1}^k becomes $8\cdot10\cdot(21000)^{28} \approx 8\cdot10^{122}$, exceeding the ontological limit 10^{120}. The result, in terms of n_{max}, is relatively insensitive to K and choice of model for Z_{n+1}^k. Choosing $K = 2000$ and 12000, for example, yields $n_{max} = 31$ and 25, respectively, with corresponding interspike intervals 0.016 and 0.020 s. Instead using a second order model for Z_{n+1}^k results in the similar character scaling $12\cdot(2K)^{n+1}$. Again $T_c = 0.5$ s is assumed. Thus, since a part of the spectrum of action potential firing cannot be represented within the ontological information limit, we find that conscious high-level processes cannot be reduced to, or related to, neuronal low-level processes.

It should be remarked that the above estimate clearly *underrepresents* the information content associated with the dynamics of a single neuron. For example, its dendrites are not identical, the biological and chemical modelling of which would substantially increase the number of bits needed to represent Z_{n+1}^{k}. Furthermore, we have only studied a *single* of the 16 billion neurons of the cerebral cortex. Thus we may safely argue that the information content associated with neural processes for consciousness exceeds the ontological limit; consciousness is ontologically emergent.

Summarizing, complex systems of the world, like the Jumping robot and myoglobin, develop or evolve. The properties of these systems, such as ability to jump and oxygen affinity, supervene on the systems. Some properties come as surprises in the sense that they cannot be reduced to anything less than the behaviour of the full system itself. These are the emergent properties. *Epistemically*, we consider the associated system (like the bits and parts of the Jumping robot), *ontologically* we consider the physical universe. The notion of algorithmic incompressibility as a diagnostic for emergence is applicable in both the epistemical and the ontological cases. If properties of a complex system, being acquired through for example long term evolution, can only be represented by the system itself, that is if nature, because of the limited quantum mechanical information capacity of the world, cannot accommodate a compressed representation of its properties, then the system features ontologically emergent properties. The neural system relating to consciousness features an incompressible character due to its nonlinear complexity as described above; thus it is ontologically emergent. As a consequence of the definitions in section 3.1 it follows that consciousness is also epistemically strongly emergent.

3.6 Neuroscience

The neural networks of the brain communicate in *discrete* nonlinear processes to generate cognitive functions such as the abilities to feel pain, think, make choices, experience feelings and introspect. If these basic neural processes were linear in their physical character, their behaviour could possibly be reduced to a theory. This theory would have lower complexity than what it describes since it would be algorithmically compressible. Nonlinear systems like the neural network of the brain, however, generally feature higher, second order complexity. Since the neural network associated with consciousness thus is nonlinear and discrete to its nature, we argue that a theory cannot be produced for it; consciousness is emergent and cannot be understood in a reductionistic framework, regardless whether we seek a computational theory of mind or some other formally reductionistic theory of mind.

It could be of interest here to briefly discuss a quite different obstacle for understanding consciousness. Abandoning efforts for finding theories of consciousness, we may be inclined to instead turn to the possibility of artificially designing consciousness. In neuroscience there is a search for 'neural correlates of consciousness' (NCC), which form the neural processes in the brain that are directly linked to the individual's current mind activities.

Let us say that NCC:s indeed can be identified to an extent that serious attempts to create conscious processes in artificial brains can be made. On each such experimental attempt, the function must be ensured - the system must be diagnosed. Otherwise there is the possibility that we have designed an advanced system that externally behaves like a consciousness but actually lacks mental processes. But a problem with this approach is that essentially no limit exists for *non-cognitive* 'intelligence' of advanced computer programs. These would then, properly designed, be able to pass any kind of Turing test. In these tests, where the respondent is hidden so that the person performing the test does not

know whether it communicates with a human or a machine, any machine producing similar responses as humans are deemed intelligent on the level of a human.

The Turing test is valuable for testing intelligence, but is obviously unreliable for testing consciousness. But what would then be an adequate diagnosis? Current definitions of phenomenal consciousness provide an answer: we must ensure that the system can have subjective experiences. But since all measurement of the functions of consciousness must be done externally, that is by laboratory personnel using diagnostic equipment, the system's internal cognitive functions cannot be measured directly. There is simply no information externally available from the system that would be indistinguishable from that which can be produced by an advanced, but unconscious, computer program. We could be facing an intelligent robot, without ability for conscious behaviour. This is, as mentioned elsewhere, therefore not a viable route for solution of the mind-body problem.

In short: understanding of a system implies the possibility of constructing it, with all of its functions. But construction does not imply understanding; since the intended functions, like generation of conscious thoughts, cannot be experimentally verified we thus cannot say with certainty that they are in place nor that we understand them.

4 Conclusions

In this work we have argued for non-reductive physicalism; mental states supervene on physical states but cannot be reduced to them. In a physicalist analysis of the mind-body problem, resting on results from mathematics and physics, the concepts of algorithmic information theory and emergence are used to argue that the problem is unsolvable. The vast neural complexity of the brain is the basic obstacle; from a thought experiment it is shown that even a much simpler but related nonlinear system may exhibit *epistemically* strongly emergent properties. Reductionistic understanding of consciousness is thus not possible. Neuroscience will continue to make progress - we will almost certainly find, for example, the cognitive centra that are active at certain stimuli or thought processes, and we may even be able to construct conscious systems - but emergent cognitive phenomena like qualia, feelings or introspection are not likely to be expressed in a theory. The 'explanatory gap' cannot be bridged.

We furthermore argue that consciousness is *ontologically* emergent; there is no possibility, even in principle, to reduce its characteristic properties to the low-level phenomena on which it is based. The limited quantum mechanical information and computational capacity of the world presents an unsurmountable obstacle. A basic example of ontological emergence, featuring less complexity than the brain, is discussed, namely the oxygen affinity of the protein myoglobin. The main argument is that if properties of a complex system, being the result of for example long term evolution, can only be represented by the evolution of the system itself - that is if nature cannot accommodate a representation of the system - then the system features ontologically emergent properties. Without an expressible relation to its constituting low-level components, consciousness in a way comes as a surprise to nature.

Interestingly, the problem of finding a true theory for consciousness is related to the problem of free will. If a theory for consciousness could be designed the mind would not be emergent, a prerequisite for free will. This follows from that if an individual's behaviour would be epistemically computable or could be simulated, its behaviour would be predictable and thus not free.

Several of the topics touched upon in this work would benefit from a more thorough analysis. The ambition here has been to sketch some of the consequences for the mind-body problem when analyzed using the tools of algorithmic information theory, emergence and quantum mechanics.

Acknowledgements

Many thanks go to Mr Keith Elkin, for sharing his knowledge in neuroscience and several related concepts as well as for comments during many philosophical discussions. Thanks also professor Erik J. Olsson for insightful and constructive discussions on several aspects of the work.

References

Block, N. (1995). On a confusion about the function of consciousness. *Behavioral and Brain Sciences*, 18, 227-247

Carnap, R. (1950). Logical Foundations of Probability. (University of Chicago Press)

Chaitin, G. J. (1987). Algorithmic Information Theory. (Cambridge University Press)

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2, 200–219

Chalmers, D. J. (2006). Strong and weak emergence. *The Re-Emergence of Emergence*. (Oxford University Press)

Corning, P. A. (2012). The re-emergence of emergence, and the causal role of synergy in emergent evolution. *Synthese*, 185, 295-317

Davies, P. C. W. (2004). Emergent biological principles and the computational properties of the universe. *Complexity* 10,11-15

Francescotti, R. M. (2007). Emergence. *Erkenntnis*, 67, 47-63

Gambini, R., Lewowicz, L. & Pullin, J. (2015). Quantum mechanics, strong emergence and ontological non-reducibility. *Found Chem* 17:117–127

Gittis, A. H., Moghadam, S. H., du Lac, S. (2010). Mechanisms of Sustained High Firing Rates in Two Classes of Vestibular Nucleus Neurons: Differential Contributions of Resurgent Na, Kv3, and BK Currents, *J Neurophysiol* 104: 1625–1634

Goldstein, J. (2013). Re-imagining emergence: Part 1. *Emergence: Complexity & Organization*, 15, 77-103

Houweling, A. R. & Brecht, M. (2007). Behavioural report of single neuron stimulation in somatosensory cortex. *Nature*, 451, 65.

Kim, J. (1999). Making sense of emergence. *Philosophical Studies*, 95, 3-36

Kim, J. (2006). Emergence: Core ideas and issues. *Synthese*, 151, 47-559

Levine, J. (1983). Materialism and qualia: the explanatory gap. *Pacific Philosophical Quarterly*, 64, 354-361

Libet, B. (1993). The neural time factor in conscious and unconscious events. *Novartis Foundation Symposia,* Ed. by Beck, G. R and Marsh, J, Ciba Foundation, 123-146.

Lloyd, S. (2002). Computational Capacity of the Universe. *Physical Review Letters*, 88, 237901-1-4

Locke, J. (1690). An Essay Concerning Human Understanding

Luisi, P. L. (2002). Emergence in chemistry: chemistry as the embodiment of emergence. *Foundations of Chemistry*, 4, 183-200

Maimon, G., & Assad, J. A. (2009). Beyond Poisson: Increased Spike-Time Regularity across Primate Parietal Cortex, *Neuron* 62, 426–440

McGinn, C. (1989). Can we solve the mind-body problem? *Mind*, 98, 349-366

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review,* 83, 435-450

Paré, D. & Gaudreau, H. (1996). Projection Cells and Interneurons of the Lateral and Basolateral Amygdala: Distinct Firing Patterns and Differential Relation to Theta and Delta Rhythms in Conscious Cats, *The Journal of Neuroscience*, 16, 3334-3350

PhilPapers, https://philpapers.org/browse/philosophy-of-mind

Schröder, J. (1998). Emergence: Non-Deducibility or Downwards Causation?, *The Philosophical Quarterly*, 48, 433-452

Silberstein, M. & McGeever, J. (1999). The search for ontological emergence. *The Philosophical Quarterly*, 49, 182-200

Silver, D. et al, (2016). Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484–489

Softky, W. & Koch, C. (1992). Cortical Cells Should Fire Regularly, But Do Not, *Neural Computation* 4, 643-646
Van Gulick, Robert, "Consciousness", *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition), Edward N. Zalta (ed.), http://plato.stanford.edu/archives/spr2014/entries/consciousness
Wolpert, D. H. (2008). Physical limits of inference. *Physica D* 237, 1257–1281