This article is part of a loosely connected series of review articles located at www.marcelkvassay.net

NEW: In February 2019, a sequel to this article, titled *The meta-problem and the transfer of knowledge between theories of consciousness*, was submitted to the *Journal of Consciousness Studies*. Its preprint is available from the open access internet archive philPapers.org.

Machines, Intelligence, Consciousness

Marcel Kvassay

What is consciousness? Can machines have it? Research community is deeply divided when it comes to such questions. In order to demonstrate "an astonishing lack of consensus" among researchers, Sloman and Chrisley (2003) compiled a collection of putative definitions. While some researchers thought consciousness could be "located in specific regions or processes in brains," others asserted that any "talk about a location for consciousness is a 'category mistake'." Some said it was missing when we sleep; others claimed it was present when we dream, and so on. This "Babel of views" signals that our preliminary notions of "consciousness" often do not amount to more than an inchoate pre-theoretical concept.

This article offers an informal comparison of two candidate frameworks for the study of consciousness: "virtual machine functionalism" formulated by Sloman and Chrisley (2003), and "nonreductive functionalism" proposed by David Chalmers (1995, 1997).

The study presented below was born of a lucky coincidence and an overlap between my personal and professional interests. I have recently joined a sub-field of applied computer science called "Intelligent and Knowledge-based Technologies" (IKT). In order to acclimatize myself to the new context, I started to explore its links with Artificial Intelligence (AI). My primary focus was the concept of "intelligence" and whether both of the fields defined it in the same way. Among the papers I found on the web, one had an intriguing title: "Virtual Machines and Consciousness" *(Sloman & Chrisley 2003)*. It set off in a no less intriguing vein:

> *The study of consciousness, a long-established part of philosophy of mind and of metaphysics, was banished from science for many years, but has recently re-entered (some would say by a back door that should have been kept locked). Most AI researchers ignore the topic, though people discussing the scope and limits of AI do not. We claim that much of the discussion of consciousness is confused because what is being referred to is not clear. That is partly because "consciousness" is a cluster concept. (Sloman & Chrisley 2003, p. 3)*

A *cluster concept*, the authors clarify,

> *is worse than merely being a vague concept (e.g. "large", "yellow"), for which a boundary cannot be precisely specified along some continuum. It is also worse than being a mongrel concept (Block 1995), which merely confuses a collection of concepts that are individually supposed to be well-defined (e.g., "phenomenal consciousness" and "access consciousness"). It is true that even in the case of vague, mongrel, and even well-behaved concepts, people often disagree on whether or how the concept is to*

> *be applied in a particular case. But what is particularly problematic about the concept "consciousness" is that such disagreement is not merely empirical, but (also) deeply conceptual, since it is often the case that disputants cannot even agree on what sorts of evidence would settle the question. (Sloman & Chrisley 2003, pp. 4-5)*

These passages struck a chord: the question of consciousness had fascinated me since high school. At a time when most of my classmates had scarcely heard of it, an unexpected encounter with yoga shattered my incipient world-view. That world-view, to be frank, hardly went beyond the usual stuff pumped into the young brains in the then still intact Communist regimes of Central and Eastern Europe. Quite naturally, I started veering towards Indian spirituality and soon embraced it fully. At first I was ignorant of the fact that yoga and spirituality could also be used for brainwashing and domination. During Communism yoga was watched with suspicion and only some of its varieties (notably hathayoga and rajayoga) were tolerated. They were typically presented as methods for improving physical health and mental capacities. On rare occasions it might be hinted – or just cautiously whispered in the ear – that they could also activate deeper and more potent layers of consciousness. Overall, this approach gave the impression of an open-ended repertoire of techniques that individuals might use on their own – of course, after a proper training by a competent instructor, just as we attend a driving course in order to get a driving licence. After the fall of Communism, however, a different picture of yoga began to emerge. Regardless of whether or not such a liberal approach was feasible, it was definitely *not* the way yoga was traditionally practised in India. And if I was not scared away at this point, it was only because I chose a system that in this respect (for me all-important) departed from the established tradition.

[redacted] The aspect that I consider the most relevant is the idea of layers of consciousness, which include but do not stop at our surface mental awareness. I tried briefly to sketch these in the context of the system I knew in the opening section of (Kvassay 2011), and I employ the same sort of perspective here.

It goes without saying that people with spiritual inclinations tend to dismiss any talk about "machine consciousness" as hopelessly reductive. I too felt the same about Sloman and Chrisley's paper on my first, superficial reading. I had the impression as if they were claiming that if a calculator could manipulate numbers, then it had to be in some sense "conscious" of them. As I re-read and cross-checked their paper with others, I gradually realised I had missed several subtle but important points, which led to my misreading of their real thesis. I shall now try to elucidate them, as well as explain why I think it is wrong to dismiss reductive approaches off-hand. To this end, I will compare Sloman and Chrisley's approach with a "minimalist" non-reductive alternative proposed by David Chalmers (1995).

* * *

The opening sections of Sloman and Chrisley's paper deal with conflicting views on the nature of consciousness. In order to demonstrate "an astonishing lack of consensus" in the research community, they compile an illustrative collection of putative definitions and present it in a two-column table. Each row stands for a particular "bone of contention," such as localizability of consciousness, while the two columns provide clashing views in that respect. Thus, for example, some researchers think that consciousness can be "located in specific regions or processes in brains," while others assert that any "talk about a location for consciousness is a 'category mistake'." Some say it is missing when we sleep; others claim it is present when we dream, and so on. This "Babel of views" signals that our preliminary notions of "consciousness" often do not amount to more than an inchoate pre-theoretical concept.

Sloman and Chrisley propose a biologically inspired and architecture-based approach that they believe would help us supplant these vague notions with more precise and empirically tractable ones. They "start with the tentative hypothesis that although the word 'consciousness' has no well-defined meaning, it is used to refer to aspects of human and animal information-processing." Their basic working assumption is:

> *The phenomena labelled "conscious" involve no magic; they result from the operation of very complex biological information-processing machines which we do not yet understand. (Sloman & Chrisley 2003, p. 9)*

They admit that while the first part of their assumption is "uncontentious to anyone of a naturalist bent," the second half is "notoriously controversial." The rest of their paper is essentially a defence of this thesis.

Since the authors try to explain consciousness as a biological phenomenon, they first need to examine the existing "biological information-processing architectures." They start with reactive architectures, the oldest and simplest group. A reactive mechanism, they write, "is one that produces outputs or makes internal changes … without … explicitly representing and comparing alternatives." It would be a mistake, however, to underestimate reactive architectures. They can adapt and learn (e.g. through weight changes in neural nets), and give rise to some of the most robust and successful biological communities:

> *Some purely reactive species have a social architecture enabling large numbers of purely reactive individuals to give the appearance of considerable intelligence, e.g. termites building "cathedrals". The main feature of reactive systems is that they lack the core ability of deliberative systems … namely, the ability to represent and reason about nonexistent or unperceived phenomena (e.g., future possible actions or hidden objects). (Sloman & Chrisley 2003, p. 22)*

"In principle," the authors note, "a reactive system can produce any external behaviour that more sophisticated systems can produce. However, to do so in practice it might require a larger memory for pre-stored reactive behaviours than could fit into the whole universe."

Next come deliberative architectures: these can "represent possibilities (e.g. possible actions, possible explanations for what is perceived) in some explicit form, enabling alternative possibilities to be compared and one selected." Examples of man-made deliberative systems include "theorem provers, planners, programs for playing board games, natural language systems, and expert systems of various sorts." Here I got an indirect answer to my original query: it is mainly deliberative algorithms that tend to get picked by other fields, such as the "Intelligent and Knowledge-based Technologies" (IKT), and employed there for special purposes.

In the early phase of Artificial Intelligence it seemed that deliberative algorithms were the key to human-like intelligence. Ultimately, however, human "common sense" turned out to be too elusive for deliberative approaches. It was a big surprise – and a great disappointment – that ordinary common sense should be so much more difficult to implement than automated theorem-proving, for example. It took the researchers some time to digest, and even today people unacquainted with AI tend to regard the latter as more difficult. AI research has made it clear, however, that this is due to an error of perspective: the complexity of the brain-processes by which we interpret our physical surroundings and perform even trivial actions simply eludes us.

The root problem regarding the early AI systems, the authors explain, was their tendency to "get stuck in loops or repeat the same unsuccessful attempt to solve a sub-problem." A more recent trend in tackling this difficulty "is to have a parallel subsystem monitoring and evaluating the deliberative processes. If it detects something bad happening, then it may be able to interrupt and re-direct the processing." The authors call this monitoring function *meta-management.*

Architectures with meta-management (i.e. with reflective and self-reflective mechanisms) form the highest rung in the biological hierarchy. "The richest versions of this evolved very recently, and may be restricted to humans," the authors remark, "though there are certain kinds of self-awareness in other primates (Hauser 2001)." They believe architectures with meta-management are sufficient for the construction of robots with human-like common sense, and present their own candidate: a "human-like architecture for cognition and affect" ("H-CogAff").

The authors ultimately zero in on the aspect widely considered the chief obstacle to replicating consciousness in a machine: the privacy and ineffability of our conscious experience, the problem of *qualia.* They go about it in a rather roundabout way, and for good reason. In fact, quite early in the article they drop a hint though I suspect that most readers (just like me) would miss the implications on their first reading:

> *Specifically, we hope to explain how an interest in questions about consciousness in general and qualia in particular arises naturally in intelligent machines with a certain sort of architecture that includes a certain sort of "meta-management" layer. Explaining the possibility (or near-inevitability) of such developments illustrates the notion of "architecture-driven" concepts (concepts likely to be generated within an architecture) and gives insightful new explanations of human philosophical questions, and confusions, about consciousness. (Sloman & Chrisley 2003, pp.3-4)*

Such claims typically signal extremely reductive views on the nature of consciousness. I will try to show that this does not necessarily apply to the framework proposed by Sloman and Chrisley, at least in the sense that it can fit in with less reductive approaches as well. With that end in view, I will now introduce a minimalist non-reductive alternative proposed by David Chalmers (1995).

* * *

While Aaron Sloman is a recognised old-timer in the field of Artificial Intelligence (and originally a philosopher), David Chalmers entered the field of philosophy of mind relatively recently (in 1990s) and is originally a mathematician. Their "nearly inverted" professional trajectories help explain, I believe, the curious complementariness that I find in their views. Although one is predominantly reductive and the other non-reductive, there is plenty of common ground. The first similarity is actually not even surprising. Given the overwhelming variety of approaches to the study of consciousness, it is but natural that Chalmers in his seminal article too starts by acknowledging the inherent ambiguity of the task:

> *There is not just one problem of consciousness. "Consciousness" is an ambiguous term, referring to many different phenomena. Each of these phenomena needs to be explained, but some are easier to explain than others. At the start, it is useful to divide the associated problems of consciousness into "hard" and "easy" problems. The easy problems of consciousness are those that seem directly susceptible to the standard methods of cognitive science, whereby a phenomenon is explained in terms of computational or neural mechanisms. The hard problems are those that seem to resist those methods. (Chalmers 1995)*

The problems Chalmers tags as "easy" include the ability of a cognitive system to discriminate, categorize, and react to environmental stimuli, to integrate information coming from different sensory channels, to access its own internal states, and to control its behaviour through deliberation. "There is no real issue about whether *these* phenomena can be explained scientifically," he writes. By tagging them as "easy," however, he does not mean they are trivial: "Getting the details right will probably take a century or two of difficult empirical work. Still, there is every reason to believe that the methods of cognitive science and neuroscience will succeed." The problem he considers "really hard" is the problem of experience:

> *When we think and perceive, there is a whir of information-processing, but there is also a subjective aspect. As Nagel (1974) has put it, there is something it is like to be a conscious organism. This*

> *subjective aspect is experience. When we see, for example, we experience visual sensations: the felt quality of redness, the experience of dark and light, the quality of depth in a visual field. Other experiences go along with perception in different modalities: the sound of a clarinet, the smell of mothballs. Then there are bodily sensations, from pains to orgasms; mental images that are conjured up internally; the felt quality of emotion, and the experience of a stream of conscious thought. What unites all of these states is that there is something it is like to be in them. All of them are states of experience. (Chalmers 1995)*

"The hard problem is hard," he continues, "precisely because it is not a problem about the performance of functions":

> *Even when we have explained the performance of all the cognitive and behavioral functions in the vicinity of experience—perceptual discrimination, categorization, internal access, verbal report—there may still remain a further unanswered question: Why is the performance of these functions accompanied by experience? A simple explanation of the functions leaves this question open. (Chalmers 1995)*

For Chalmers, this is the key question regarding consciousness: "Why doesn't all this information-processing go on 'in the dark,' free of any inner feel?"

> *We know that conscious experience does arise when these functions are performed, but the very fact that it arises is the central mystery. There is an explanatory gap (a term due to Levine 1983) between the functions and experience, and we need an explanatory bridge to cross it. A mere account of the functions stays on one side of the gap, so the materials for the bridge must be found elsewhere. (Chalmers 1995)*

Chalmers admits that "a remarkable number of phenomena have turned out to be explicable wholly in terms of entities simpler than themselves," but points out that this is not universal. He likens the situation to earlier developments in physics, where "it occasionally happens that an entity has to be taken as *fundamental*." Such entities

> *are not explained in terms of anything simpler. Instead, one takes them as basic, and gives a theory of how they relate to everything else in the world. For example, in the nineteenth century it turned out that electromagnetic processes could not be explained in terms of the wholly mechanical processes that previous physical theories appealed to, so Maxwell and others introduced electromagnetic charge and electromagnetic forces as new fundamental components of a physical theory. To explain electromagnetism, the ontology of physics had to be expanded. New basic properties and basic laws were needed to give a satisfactory account of the phenomena. (Chalmers 1995)*

The fact that we do not try to explain entities like mass, charge and space-time in terms of anything simpler, Chalmers argues,

> *does not rule out the possibility of a theory of mass or of space-time. There is an intricate theory of how these features interrelate, and of the basic laws they enter into. These basic principles are used to explain many familiar phenomena concerning mass, space, and time at a higher level. (Chalmers 1995)*

Chalmers proposes to accept experience as a fundamental concept. "We know that a theory of consciousness requires the addition of something fundamental to our ontology," he maintains,

> *as everything in physical theory is compatible with the absence of consciousness. We might add some entirely new nonphysical feature, from which experience can be derived, but it is hard to see what such a feature would be like. More likely, we will take experience itself as a fundamental feature of*

> *the world, alongside mass, charge, and space-time. If we take experience as fundamental, then we can go about the business of constructing a theory of experience. (Chalmers 1995)*

"A nonreductive theory of experience," he clarifies,

> *will add new principles to the furniture of the basic laws of nature. These basic principles will ultimately carry the explanatory burden in a theory of consciousness. Just as we explain familiar high-level phenomena involving mass in terms of more basic principles involving mass and other entities, we might explain familiar phenomena involving experience in terms of more basic principles involving experience and other entities. (Chalmers 1995)*

Chalmers terms these new principles "psychophysical," since their role is to bridge the "explanatory gap" between the functions and experience. This implies that they should not interfere with the laws of physics, "as it seems that physical laws already form a closed system. Rather, they will be a supplement to a physical theory." Chalmers concedes that this position "qualifies as a variety of dualism, as it postulates basic properties over and above the properties invoked by physics." "But it is an innocent version of dualism," he maintains, "entirely compatible with the scientific view of the world":

> *Nothing in this approach contradicts anything in physical theory; we simply need to add further bridging principles to explain how experience arises from physical processes. There is nothing particularly spiritual or mystical about this theory—its overall shape is like that of a physical theory, with a few fundamental entities connected by fundamental laws. It expands the ontology slightly, to be sure, but Maxwell did the same thing. Indeed, the overall structure of this position is entirely naturalistic, allowing that ultimately the universe comes down to a network of basic entities obeying simple laws, and allowing that there may ultimately be a theory of consciousness cast in terms of such laws. If the position is to have a name, a good choice might be naturalistic dualism. (Chalmers 1995)*

Chalmers' 1995 paper elicited a number of responses, which he analysed and answered in (Chalmers 1997). In the analysis he first confronted reductive (or "deflationary") critiques and distinguished two types of materialism:

> *The type-A materialist denies that there is a "hard problem" distinct from the "easy" problems; the type-B materialist accepts (explicitly or implicitly) that there is a distinct problem, but argues that it can be accommodated within a materialist framework all the same. (Chalmers 1997)*

"Type-A materialism," Chalmers explains,

> *is not merely the view that consciousness is identical to some function, or that it plays a functional role, or that explaining the functions will help us explain consciousness. It is the much stronger view that there is not even a distinct question of consciousness: once we know about the functions that a system performs, we thereby know everything interesting there is to know. (Chalmers 1997)*

"This is an extremely counterintuitive position," he remarks.

> *At first glance, it seems to simply deny a manifest fact about us. But it deserves to be taken seriously: after all, counterintuitive theories are not unknown in science and philosophy. On the other hand, to establish a counterintuitive position, strong arguments are needed. And to establish this position – that there is really nothing else to explain – one might think that extraordinarily strong arguments are needed. So what arguments do its proponents provide? (Chalmers 1997)*

"Perhaps the most common strategy for a type-A materialist," he observes,

> *is to deflate the "hard problem" by using analogies to other domains, where talk of such a problem would be misguided. Thus Dennett imagines a vitalist arguing about the hard problem of "life", or a neuroscientist arguing about the hard problem of "perception". Similarly, Paul Churchland (1996) imagines a nineteenth century philosopher worrying about the hard problem of "light", and Patricia Churchland brings up an analogy involving "heat". In all these cases, we are to suppose, someone might once have thought that more needed explaining than structure and function; but in each case, science has proved them wrong. So perhaps the argument about consciousness is no better. (Chalmers 1997)*

For Chalmers, such arguments do not carry much weight. "There is a disanalogy between the problem of consciousness and problems in other domains," he says, since in the latter case "it is more or less *obvious* that structure and function are what need explaining, at least once any experiential aspects are left aside."

> *When it comes to the problem of life, for example, it is just obvious that what needs explaining is structure and function: How does a living system self-organize? How does it adapt to its environment? How does it reproduce? Even the vitalists recognized this central point: their driving question was always "How could a mere physical system perform these complex functions?", not "Why are these functions accompanied by life?" It is no accident that Dennett's version of a vitalist is "imaginary". There is no distinct "hard problem" of life, and there never was one, even for vitalists. (Chalmers 1997)*

In the case of consciousness, however, "the manifest phenomena that need explaining" include not only structure and functions, but also *subjective experience*. So, for Chalmers, "this analogy does not even get off the ground." The same line of reasoning applies to "*any* phenomenon that we observe in the external world":

> *When we observe external objects, we observe their structure and function; that's all. Such observations give no reason to postulate any new class of properties, except insofar as they explain structure and function; so there can be no analog of a "hard problem" here. Even if further properties of these objects existed, we could have no access to them, as our external access is physically mediated: such properties would lie on the other side of an unbridgeable epistemic divide. Consciousness uniquely escapes these arguments by lying at the center of our epistemic universe, rather than at a distance. In this case alone, we can have access to something other than structure and function. (Chalmers 1997)*

"To have any chance of making the case," Chalmers concludes, "a type-A materialist needs to *argue* that for consciousness, as for life, the functions are all that need explaining."

> *Often, a proponent will simply assert that functions are all that need explaining, or will argue in a way that subtly assumes this position at some point. But that is clearly unsatisfactory. Prima facie, there is very good reason to believe that the phenomena a theory of consciousness must account for include not just discrimination, integration, report, and such functions, but also experience, and prima facie, there is good reason to believe that the question of explaining experience is distinct from the questions about explaining the various functions. Such prima facie intuitions can be overturned, but to do so requires very solid and substantial argument. (Chalmers 1997)*

"Such arguments," Chalmers observes, "are surprisingly hard to find." Among the contributors to the symposium, he notes, Daniel Dennett seems to be the only one openly endorsing the view that "in the case of consciousness, the functions are all that need explaining." The key to Dennett's position, Chalmers suggests,

> *lies in what Dennett has elsewhere described as the foundation of his philosophy: "third-person absolutism". If one takes the third-person perspective on oneself – viewing oneself from the outside,*

> *so to speak – these reactions and abilities are no doubt the main focus of what one sees. But the hard problem is about explaining the view from the first-person perspective. So to shift perspectives like this – even to shift to a third-person perspective on one's first-person perspective, which is one of Dennett's favorite moves – is again to assume that what needs explaining are such functional matters as reactions and reports, and so is again to argue in a circle. (Chalmers 1997)*

Chalmers is not flustered by the Dennett's argument, "subtract the functions and nothing is left":

> *An analogy suggested by Gregg Rosenberg is useful here. Color has properties of hue, saturation, and brightness. It is plausible that if one "subtracts" hue from a color, nothing phenomenologically significant is left, but this certainly doesn't imply that color is nothing but hue. So even if Dennett could argue that function was somehow required for experience (in the same way that hue is required for color), this would fall a long way short of showing that function is all that has to be explained. (Chalmers 1997)*

When challenged "to provide 'independent' evidence … for the 'postulation' of experience," Chalmers replies: "But this is to miss the point: conscious experience is not 'postulated' to explain other phenomena in turn; rather, it is a phenomenon to be explained in its own right." And he quips:

> *I would be interested to see Dennett's version of the "independent" evidence that leads physicists to "introduce" the fundamental categories of space and time. It seems to me that the relevant evidence is spatiotemporal through and through, just as the evidence for experience is experiential through and through. (Chalmers 1997)*

Proponents of type-A materialism "sometimes like to suggest that their view is supported by the results of modern science," Chalmers remarks, "but all the science that I know is quite neutral here: I have never seen any experimental result that implies that functions are all that need to be explained."

Given the overall tenor of Chalmers' argument, his concluding words rather surprised me:

> *This is not to say that type-A materialism cannot be argued for at all. There are a few sophisticated arguments for such a position in the literature (for example, Shoemaker 1975 and White 1986), but even these ultimately come down to "consider the alternatives", isolating the difficulties that one gets into if one accepts that there is a further phenomenon that needs explaining. There is no doubt that these difficulties (both ontological and epistemological) are considerable; life would be a lot easier if the hard problem did not exist. But I think these difficulties are solvable; and in any case, to deny the problem because of the difficulties has the flavor of solution by decree. (Chalmers 1997)*

Chalmers does not think the results achieved by type-A materialists are worthless, or that type-A materialism is going to disappear any time soon:

> *We will probably just have to get used to the fact that there is a basic division in the field: that between those who think the "easy" problems are the only problems, and those who think that subjective experience needs to be explained as well. We can therefore expect two quite distinct sorts of theories of consciousness: those which explain the functions and then say "that's all", and those which take on an extra burden. In the end, the most progress will probably come from internal advances in the respective research programs, rather [than] from the endless battle between the two. So beyond a certain point in the argument, theorists in these camps might just agree to disagree and get on with their respective projects. This way, everyone can move forward. (Chalmers 1997)*

Chalmers then proceeds with equally illuminating analyses of type-B materialism and nonreductive critiques, which I will take up later.

* * *

Overall, it is quite obvious that Sloman and Chrisley's views are close to Dennett's. After all, Sloman himself admits that much in a note on Dennett's book *The Intentional Stance* (Sloman 1988), even as he argues that the "intentional stance" is not enough: a "design stance" alone, in his opinion, can provide "real insight into the nature of intelligence." This was in fact the main reason why Sloman shifted from academic philosophy to artificial intelligence. I will now briefly review some of the results and implications of this approach.

Sloman and Chrisley's "design stance" is based on "a new kind of functionalist analysis of mental concepts" that they have developed. They call it "*virtual machine functionalism*" (VMF), and claim it to be immune from "a number of standard objections" to functionalism among philosophers. "Most philosophers and cognitive scientists," they remark, "write as if 'functionalism' were a well-defined, generally understood concept." As an example, they quote (Block 1996):

> *According to functionalism, the nature of a mental state is just like the nature of an automaton state: constituted by its relations to other states and to inputs and outputs. All there is to [a mental state] S1 is that being in it and getting a [certain] input results in such and such, etc. According to functionalism, all there is to being in pain is that it disposes you to say 'ouch', wonder whether you are ill, it distracts you, etc. (Block 1996)*

Block's summary, according to Sloman and Chrisley, "has (at least) two different interpretations." The first, in which "an entity can have only one, indivisible, mental state at a time," they propose to call "*atomic state functionalism*" and dismiss on the grounds that such indivisible mental states "could not be states like human hunger, thirst, puzzlement or anger, since these can coexist and start and stop independently."

In the second interpretation, an entity can have "several coexisting, independently varying, interacting mental states" at the same time:

> *It is possible that Block did not realise that his examples, as ordinarily understood, were of this kind: for instance, the same pain can both make you wonder whether you are ill and distract you from a task, so that having a pain, wondering whether you are ill, having the desire or intention to do the task, and being distracted from it are four coexisting states which need not start and end at the same time. If the pain subsides, you may continue to wonder whether you are ill, and while working on the task (e.g. digging the garden) you might form a new intention to see a doctor later. Coexistence of interacting sub-states is a feature of how people normally view mental states, for instance when they talk about conflicting desires or attitudes. (Sloman & Chrisley 2003, p. 16)*

It is this "parallelized" version of functionalism that the authors call "*virtual machine functionalism*" (VMF). They distinguish between two forms: *restricted* VMF, which "requires that every sub-state be causally connected, possibly indirectly, to inputs and outputs of the whole system," and *unrestricted* VMF, which is free from this constraint. The distinction is important. It is generally accepted that the space of design options for intelligent systems is so wide as to be practically unmanageable. Consequently, we need meaningful criteria to constrain it. Biologically inspired approaches often invoke the notion of evolution to do the filtering: ideally, no new feature should be introduced in the design until it has been plausibly demonstrated how it could have helped structurally similar organisms survive or gain evolutionary advantage. Since VMF is biologically inspired, it needs to prove itself in this respect too. While *restricted* VMF fares well on that score, *unrestricted* VMF is a bit problematic. I will skip the details and rather focus on *why* the authors put effort into legitimizing *unrestricted* VMF. The reason is simple: it turns out it can elegantly explain certain elusive features of human psychology, such as the problem of *qualia*. The authors demonstrate this through a graded series of examples:

> *Unrestricted VMF allows that some sub-state S or continuing sub-process is constantly altered by other sub-states that are connected with the environment even though none of the changes that occur in S affect anything that can affect the environment. An example in a computer might be some*

> *process that records statistics about events occurring in other processes, but does not transmit any of its records to any other part of the system, and no other part can access them. Unrestricted VMF even allows sub-systems that beaver away indefinitely without getting any inputs from other parts of the system. For instance, a sub-system might be forever playing games of chess with itself, or forever searching for proofs of Goldbach's conjecture. (Sloman & Chrisley 2003, p.18)*

These initial examples are not terribly impressive. While the statistics may have a value for the *end-users* of the system, most system administrators are likely to kill any run-away process the moment they spot one. And in most cases it is indeed the right thing to do, except for autonomously evolving systems, where each process may need the freedom to disconnect and reconnect back on its own.

We get a first glimpse of what the authors are up to a bit later, as they pass on to an interesting case of a semi-detached process,

> *that causally interacts with other processes, e.g. by sending them instructions or answers to questions, but whose internal details do not influence other processes, e.g. if conclusions of reasoning are transmitted, but none of the reasons justifying those conclusions. Then other parts of the system may know what was inferred, but be totally unable to find out why. (Sloman & Chrisley 2003, pp. 18-19)*

This category includes an important special case of a semi-detached process in the meta-management layer that monitors and evaluates other processes:

> *This internal self-observation process might have no causal links to external motors, so that its information cannot be externally reported. If it also modifies the processes it observes … then it may have external effects. However it could be the case that the internal monitoring states are too complex and change too rapidly to be fully reflected in any externally detectable behaviour: a bandwidth limitation. For such a system experience might be partly ineffable. (Sloman & Chrisley 2003, p.19)*

The authors seem to imply that partial ineffability at meta-management level amounts to a kind of proto-*qualia*, or perhaps even a sort of *qualia* instantiated in a machine. They buttress up their argument through an elaborate discussion of concept formation in self-organising systems, which, as they show, provide scope for more advanced forms of *qualia*. The discussion is rather technical, so I won't go into details. Suffice it to say that they managed to convince me that intelligent machines can indeed possess various sorts of *qualia* (more precisely, I felt we should rather speak of proto-*qualia*, structural and functional preconditions that may lead to full-blown *qualia* in the presence of consciousness). Henceforth, I would hesitate to accept the mere existence of *qualia* in humans as an insuperable objection against machine consciousness. Given that till now I have been a staunch opponent of the very idea of machine consciousness, it is quite a feat. It may still turn out there are kinds of *qualia* that only humans can have, but that is a different story. And the proponents of machine consciousness are certainly entitled to ask: "Could you be more specific as to what kinds of *qualia* are those, and what makes you think intelligent machines should forever be incapable of them?" I admit I could not give a fitting reply to such a question yet.

At the same time, I had reservations about Sloman and Chrisley's argument, mainly because they try to pull it off the Dennettian way, that is, without bringing in "consciousness" in the non-reductive sense of the word. And without it I could not see how the proto-*qualia* they so ingeniously devised could bring about something that non-reductionists too would accept as full-blown *qualia*. Presumably, the authors are of the opinion that no such non-reductive full-blown *qualia* exist, and that that is precisely the error of non-reductionists, as evidenced by the amusing collection of inchoate "pre-theoretical" notions of consciousness that they started with. Grant that as yet there is no agreed nonreductive definition of the term, and that most people indeed use it incoherently. I do not see how this entails the non-existence of consciousness in the nonreductive

sense, when it may simply be (due to its unique properties) just exceedingly difficult to capture in a consistent intellectual framework.

The moral of the story goes along with Chalmers' earlier suggestion: wherever we seem to be dealing with something persistently eluding conceptualization due to principled reasons, there is legitimate scope for those who believe in its reality and want to continue trying, as well as for those who doubt its reality and prefer to do without it. In the case of consciousness it is evident that both directions yield interesting insights and spur each other to more intense efforts towards a solution. And this, in my opinion, would hold even if it came to pass that neither could provide a definitive conceptual framework for dealing with the problem.

* * *

When I first read Chalmers' 1995 article, I took his *naturalistic dualism* to imply that consciousness *cannot* be instantiated in a machine even in principle – an exact opposite of Sloman and Chrisley's position. You can imagine my consternation when I later discovered that the Wikipedia entry on Artificial Consciousness credited Chalmers for "one of the most explicit arguments for the plausibility of artificial consciousness"! How come? Wikipedia referred to Chalmers' article "A Computational Foundation for the Study of Cognition" (Chalmers 2011), which was based on an early unpublished draft of 1993. As I read through it, I had the impression of an almost exact match of Chalmers' ideas with those of Sloman and Chrisley. "Hm. Looks like he wrote this while he was a type-B materialist," I thought. This interpretation seemed to be supported e.g. by "Mind and Consciousness: Five Questions" (Chalmers 2009), where he says that while working on his 1996 book he became "convinced … contrary to my initial inclination, that a materialist approach to consciousness cannot succeed." And later he adds:

> *I don't think that a successful science of consciousness can be a wholly reductive science of consciousness, cast in terms of neuroscience or computation alone. Rather, I think it will be a nonreductive science, one that does not try to reduce consciousness to a physical process, but rather studies consciousness in its own right and tries to find connections to brain, behavior, and other cognitive processes. (Chalmers 2009)*

At the same time, in his recent introductory note to his 1993 draft (Chalmers 2011) he says he is "still largely sympathetic with the views expressed here, in broad outline if not in every detail." "So, is he a computationalist or not?" I muttered to myself as I hurriedly scanned his articles for any clue. Eventually I realised I had made a similar mistake as with Sloman and Chrisley earlier. The clues, in this case, were as follows. First, although in his 1995 article Chalmers mentions he is no longer a materialist, he still keeps the "dancing qualia" argument in favour of the thesis of computability of mind. This was in fact my primary source of confusion: while I took computationalism to imply both functionalism and reductionism (i.e. at most type-B materialism), Chalmers explicitly rejected reductionism in that very same article!

The second clue was Chalmers' opening words from the section dealing with non-reductive critiques in (Chalmers 1997), where he remarks that in his 1995 article he seems to have "staked out some middle ground":

> *The intermediate nature of my position may stem from an inclination toward simplicity and toward science. Reductive materialism yields a compellingly simple view of the world in many ways, and even if it does not work in the case of consciousness, I have at least tried to preserve as many of its benefits as possible. (Chalmers 1997)*

The third and final clue was the fact that consciousness has two aspects: phenomenal (roughly, what it *is* or how it *feels* to be something) and functional (what consciousness *does*, exemplified by its participation in the mental processes like verbal reporting, discrimination, categorization, deliberation, etc.). The first is usually termed "phenomenal consciousness," the second "access

consciousness." Chalmers proposes a slightly different terminology, calling the first "consciousness" and the second "awareness."

Because I missed the import of the third clue, I wrongly believed that abandoning reductionism implies abandoning computationalism altogether. In fact it is possible to make a half-step only, and abandon computationalism with respect to one aspect of consciousness while still keeping it with regard to the other. This, then, seems to be Chalmers' real position, and it might be summed up in the motto: "We cannot compute what consciousness *is*, but we can still compute what it *does*." The first half ensures the view is nonreductive, while the second keeps it within the ambit of functionalism, broadly conceived. Chalmers himself calls his position "nonreductive functionalism."

Is such a position philosophically feasible? Chalmers believes it is and puts forward a candidate framework inspired by Russellian monism. This view builds on the fact that

> *physics characterizes its basic entities only extrinsically, in terms of their causes and effects, and leaves their intrinsic nature unspecified. For everything that physics tells us about a particle, for example, it might as well just be a bundle of causal dispositions; we know nothing of the entity that carries those dispositions. The same goes for fundamental properties, such as mass and charge: ultimately, these are complex dispositional properties (to have mass is to resist acceleration in a certain way, and so on). But whenever one has a causal disposition, one can ask about the categorical basis of that disposition: that is, what is the entity that is doing the causing? (Chalmers 1997)*

If we dodge this question by stipulating that the world consists only of dispositions, Chalmers observes, we are left with

> *a vast amount of causation and no entities for all this causation to relate! It seems to make the fundamental properties and particles into empty placeholders ... and thus seems to free the world of any substance at all. (Chalmers 1997)*

"The idea of a world of 'pure structure' or of 'pure causation' has a certain attraction," he admits, "but it is not at all clear that it is coherent." This naturally leads to two questions:

> *(1) what are the intrinsic properties underlying physical reality?; and (2) where do the intrinsic properties of experience fit into the natural order? Russell's insight, developed by Maxwell and Lockwood, is that these two questions fit with each other remarkably well. Perhaps the intrinsic properties underlying physical dispositions are themselves experiential properties, or perhaps they are some sort of proto-experiential properties that together constitute conscious experience. This way, we locate experience inside the causal network that physics describes, rather than outside it as a dangler; and we locate it in a role that one might argue urgently needed to be filled. And importantly, we do this without violating the causal closure of the physical. The causal network itself has the same shape as ever; we have just colored in its nodes. (Chalmers 1997)*

An interesting exposition of why the intrinsic properties of the physical should have anything to do with consciousness is given in (Rosenberg 1999). A closely related question is whether Russellian monism, by embracing the causal closure of the physical, does or does not lead to epiphenomenalism (the view that consciousness has no effect on the physical world). In this respect, Chalmers maintains that by placing experience inside the causal network we give it a causal role:

> *Indeed, fundamental experiences or proto-experiences will be the basis of causation at the lowest levels, and high-level experiences such as ours will presumably inherit causal relevance from the (proto)-experiences from which they are constituted. So we will have a much more integrated picture of the place of consciousness in the natural order. (Chalmers 1997)*

Of course, Russellian monism has its own problems. Chalmers mentions, among others, "the threat of panpsychism" and "the problem of how fundamental experiential or proto-experiential properties at the microscopic level somehow together *constitute* the sort of complex, unified experience that we possess." In addition, many nonreductive researchers criticized Chalmers for remaining too close to functionalism. In order to illustrate their concerns, I will briefly review (Lowe 1995).

Lowe commends Chalmers for "challenging the complacent assumptions of reductive physicalism" but fears that Chalmers' approach "plays into the hands of physicalists by suggesting that the only problem with functionalism is its apparent inability to say anything about 'qualia'." He finds "Chalmers' notions of experience and consciousness … seriously inadequate," particularly in missing how deeply and inextricably is "the intentional content of a perceptual experience … grounded in its phenomenal character." He also objects to Chalmers' reliance on the Shannonian notion of information (which he considers "wholly inappropriate for characterizing [human] cognitive states"), as well as to his "terminological proposal regarding the use of the words 'consciousness' and 'awareness'":

> *In Chalmers' proposed sense of 'awareness', it seems fair to say, there could be nothing in principle wrong in speaking of a computer, or even a thermostat, as being 'aware' — but then to suggest that human beings are only 'aware' in this attenuated sense is completely to misrepresent the capacities involved in our being 'aware' of our selves and of our own thoughts and experiences. (Lowe 1995)*

Lowe's main charge goes against Chalmers' view of "human thought and cognition in general [as] just a matter of 'information-processing' … which could in principle go on in a mindless computer." This, according to Lowe, might lead to the idea

> *that all that is really distinctive about consciousness is its qualitative or phenomenal aspects (the 'what it is like', or 'inner feel'). And then it begins to look like a strange mystery or quirk of evolution that creatures like us should possess this sort of consciousness in addition to all our capacities for thought and understanding — these capacities being, for Chalmers, simply capacities for certain sorts of information-processing and storage. (Lowe 1995)*

Lowe's answer to this "strange mystery" is

> *that consciousness has only been put in this queer position by Chalmers (and, to be fair, by many others) because he has mistakenly denied it any role in his account of the nature of human thought and understanding. In short, it is the reductive, and wholly inadequate, information-processing conception of human cognition which is responsible for the misperception that 'consciousness' (in the form of 'qualia' and the like) occupies what threatens to be a merely epiphenomenal role as a peculiar additional feature of human mentation that is in no way essential to our basic intellectual capacities. (Lowe 1995)*

Answering Lowe's criticism, Chalmers clarifies that in his 1995 article he did not intend to claim that humans were "aware" *only* in the same (attenuated) sense as machines. He also did not intend to address "the exact relationship between consciousness and 'intentional' (or semantic) mental states" as this raised "deep and subtle questions" that were beyond the scope of that paper. "I am torn on the question of intentionality," Chalmers writes,

> *being impressed on one hand by its phenomenological aspects, and on the other hand being struck by the potential for functional analyses of specific intentional contents…. Over time I am becoming more sympathetic with the [idea] … that consciousness is the primary source of meaning, so that intentional content may be grounded in phenomenal content, as Lowe puts it. (Chalmers 1997)*

This reply, given fifteen years back, seemed to hint at a possibility of modification of his computationalist view of "awareness" (functional side of consciousness). Of course, his recent

comment that he is "still largely sympathetic with the views expressed" in his 1993 draft (Chalmers 2011) rules out any substantial change. Further details are unlikely to transpire before the ongoing discussion of his draft in the *Journal of Cognitive Science* concludes with the publication of his final response and analysis.

Let me wrap up on a personal note. I am not in a position to pass authoritative judgements on these debates, but I found it deeply rewarding to formulate my own hypotheses and then watch how they would fare with respect to the debated points (often not too well, as you could see). Being a person with spiritual inclinations, it should not come as a surprise that Lowe's views resonated with me the most. At the same time I found Chalmers' "nonreductive functionalism" invaluable as the "minimalist" version of non-reductionism, and perhaps even an ideal communication point for the flow of research results between the reductionist and the non-reductionist camps.

From my point of view, Chalmers' most significant result in the articles I dealt with might be summed up as a theorem: "Even if we grant that the functional aspects of human consciousness can be subsumed by machine-level information-processing, yet consciousness in its phenomenal aspect remains irreducible to structure and function." Viewed from this perspective, his "selling out to functionalism," for which he was criticized by Lowe and others, becomes in fact his greatest asset. As in mathematics, where the theorem with the weakest assumptions is the strongest and the most generally applicable, I find Chalmers' result extremely potent with regard to those who find themselves near the borderline between the reductive and the nonreductive camps.

Last but not least, I consider Sloman and Chrisley's architectural ideas regarding intelligent systems very inspiring and, paradoxically, these might be the ones that I end up using in my work in the field of IKT. Although Sloman and Chrisley lean towards Dennett, my initial impression of their thesis – that e.g. calculators, simply by virtue of their capability to manipulate numbers, must be in some way conscious of them – was plain wrong. They stake this claim on behalf of highly intelligent machines with sophisticated meta-management layer that can reflectively turn upon itself and inspect and deliberate about its own functioning. The fact that non-reductionists disagree with this does not diminish the practical value of such designs. If we are going to have robots with human-like intelligence around in the foreseeable future (and I for one believe that we are), then it seems to me their design is more than likely to draw on the results achieved by Sloman and his colleagues since the late 1970s. Lowe's contention (shared by other nonreductive contributors) that this is not how human cognition actually works, does not impair, in my opinion, the utility of these ideas for building intelligent machines. Besides, I have seen several intriguing and noteworthy defences of functionalism even with respect to human cognition, so the dispute is far from settled.

Even more important is the fact that Sloman and Chrisley's architectural ideas can be straightforwardly "co-opted" onto the nonreductive Russellian framework and so (perhaps) avail themselves of the minimal element of subjective experience (or proto-experience) required for their functional proto-*qualia* to turn into genuine *qualia*. This makes a compelling case for the plausibility of artificial consciousness – an idea I was unwilling to entertain before. Now I find it almost past dispute. Its practical realisation, however, is not going to be easy. It won't do simply to say, "Well, you see, here is a robot pretty close to humans in terms of intelligence, and since it is agreed now that the intrinsic properties of the physical *do* provide for an element of subjective experience, isn't it about time to accept that robots like this are conscious?" This won't wash because there is no guarantee that, as we go on constructing more and more intelligent machines, the intrinsic proto-experiential properties of their material components follow suit. They may not do so at all, in which case we just end up with intelligent machines without consciousness or, more precisely, with only so much diffused proto-consciousness as Russellian monism concedes to any lump of amorphous material. For me, the question of consciousness is distinct from that of intelligence, and I am not entirely convinced by Chalmers' "dancing qualia" argument, because it makes the functional specification of the system fully determinative of its subjective experience. My uneasiness stems from the fact that functional specifications are expressed in formal (extrinsic) terms, while the intrinsic properties are supposed to be richer and *carry* the extrinsic ones, so it is

far from certain that functional specifications suffice for consciousness (as opposed to intelligence). Chalmers is well aware of these problems and admits that experiential composition is not likely to follow the rules of physical composition, though he still hopes it might follow those of *informational* composition. These are complicated issues, which I can't take up now, so let me just stick to the main point: while the road to artificial consciousness is at present unclear, Russellian monism clearly admits of such a possibility, at least in principle.

The upshot of it all is that while Chalmers' ideas make a compelling case for non-reductionism, their side-effect – an equally compelling case for artificial consciousness – is something that most non-reductionists shy away from. There is here an element of surprise for both camps, which I take for a sign that genuine and substantial progress has been made.

Let me conclude by reiterating my starting point: this article was conceived as a preliminary and informal exploration of a fascinating area lying at the intersection of my personal and professional interests. As far as possible, I tried to rely on open-access sources and provide a web link for each article in the reference section. As for the responses to (Chalmers 1995), these were originally published in the *Journal of Consciousness Studies*, which (unfortunately), is not open-access, but (Lowe 1995), for example, has been subsequently reprinted in *Antimatters* and can be downloaded from there. Some other responses are available online as well; links to them are given in (Chalmers 1997). I look forward to studying the much more that has happened in the fields of artificial intelligence, cognitive science and philosophy of mind since these debates took place. I am sure pondering over all the new material will be no less rewarding than over the ideas presented here.

References:

Block, N. (1996). What is functionalism?
http://www.nyu.edu/gsas/dept/philo/faculty/block/papers/functionalism.html
(Originally in *The Encyclopedia of Philosophy Supplement*, Macmillan, 1996)

Chalmers, D. J. (1995). Facing up to the problem of consciousness.
http://consc.net/papers/facing.html
(Originally in the *Journal of Consciousness Studies* 2:200-19)

Chalmers, D. J. (1997). Moving forward on the problem of consciousness.
http://consc.net/papers/moving.html
(Originally in the *Journal of Consciousness Studies* 4:3-46)

Chalmers, D. J. (2009). Mind and Consciousness: Five Questions.
http://consc.net/papers/five.pdf
(Originally in Patrick Grim, ed. *Mind and Consciousness: Five Questions*. Automatic Press, 2009.)

Chalmers, D. J. (2011). A Computational Foundation for the Study of Cognition.
http://consc.net/papers/computation.html
(Originally in the *Journal of Cognitive Science* 12:323-57.)

Kvassay, M. (2011). Psychological Foundations of Sri Aurobindo's Philosophy and His Approach to the Problem of Evil.
http://marcelkvassay.net/article.php?id=psychological

Lowe, E. J. (1995). There are no easy problems of consciousness.
http://anti-matters.org/articles/46/public/46-41-1-PB.pdf
(Originally in the *Journal of Consciousness Studies* 2:266-71.)

Rosenberg, G. H. (1999). On the Intrinsic Nature of the Physical.
http://cognet.mit.edu/posters/TUCSON3/Rosenberg.html
(Originally in *Toward a Science of Consciousness III, The Third Tucson Discussions and Debates*, 1999.)

Sloman, A. (1988). Why Philosophers Should be Designers.
http://www.cs.bham.ac.uk/research/projects/cogaff/sloman-dennett-bbs-1987.pdf
(Commentary submitted to *The Behavioral and Brain Sciences Journal*.)

Sloman, A. & Chrisley, R. (2003). Virtual Machines and Consciousness.
http://www.cs.bham.ac.uk/research/projects/cogaff/sloman-chrisley-jcs.pdf
(Originally in *Journal of Consciousness Studies*, 10:113–172)