Croatian Journal of Philosophy
Vol. V, No. 15, 2005

How to Eliminate Computational Eliminativism

DAVOR PEĆNJAK
Institute of Philosophy, Zagreb

Concerning the question about consciousness, Georges Rey argues that it does not exist from the success of computational theory of human mind. Everything that such a theory requires can be fulfilled by machines which do not have consciousness. So, according to theoretical parsimony, we do not have to attribute consciousness even to human beings. I wish to offer reasons why we should not doubt the existence of consciousness by showing that computational explanations can be explanations of just one part of an aspect of the human mind. Consciousness is also an explanandum rather than an explanans, and the possible reference of "what it is like" expression. Epistemic situation regarding possible accesses to consciousness is also considered.

A question about consciousness could be at least twofold—we can ask what is consciousness or we can ask even more fundamental question, namely whether it exists at all. Although the latter part of the question is rather surprising on its own account, what is even more surprising is the answer that Georges Rey [1997] gives. His answer is *no*! How is such a conclusion drawn?

Let me first expose Rey's [1997] theory and argument, assuming that not everyone is familiar with these matters, and especially because I shall focus on a particular version of it. Then, in this short paper, I would like to show that, despite its merits, this is not as convincing as it may seem. I admit that I will not present some very strong arguments in favour of the existence of consciousness but shall instead try to provide certain appeals to our intuitions how to retain the notion of consciousness for further research.

The plan of Rey's elegant argument is as follows. After describing various mental phenomena as candidates for consciousness and theories which aim to explain them, he asks whether these mental phenomena can be non-conscious. In virtue of their explanations, the surprising answer is, according to Rey, *yes*. These theories and explanations could be completely applicable to systems for which we would not be inclined to say that they are conscious or that they have consciousness; and they are

applicable to building machines—computers—which perform what they perform without the need for consciousness. So, because these machines and computational devices do not have consciousness, and since basically the same theory and explanation can serve to explain human mentality and behaviour, we have a very good reason for saying that consciousness does not in fact exist!

We can call this kind of eliminativism about consciousness computational eliminativism, or explanatory eliminativism, or functional eliminativism. Perhaps computational eliminativism is the most appropriate. It is appropriate to flesh out some important details of Rey's text.

Thinking is *par excellence* an example of mental process. It is a good candidate for consciousness. What components a certain system must have in order to perform thinking? Rey ([1997], 464, 467, 468, 471) says that the following suffices:

1. the alphabet, formation, and transformation rules for quantified modal logic with indexicals (...);
2. the axioms for a system of inductive logic, and an abductive system of hypotheses, with a "reasonable" function for selecting among them for a given input;
3. the axioms for decision theory, with some set of basic preferences;
4. (...) various transducers (...) for supplying inputs to 2;
5. devices that permit (...) realizing outputs (...);
6. the recursive "believer system";
7. a fragment of English adequate to describe/express the mental states entered in executing 1–6, descriptions which are produced as a reliable consequences of being in those states;
8. the Cartesian intuition.

I slightly modified some of the expressions, but they are equivalent to those originally devised by Rey.

These eight conditions do not provide only for thinking, but for sensing as well, and even for self-reflection. How that system operates? Rey ([1997], 464) says: "The input supplied by 4 would produce 'observation' sentences that would be checked against comparable deductive consequences of the hypotheses provided by 2; hypotheses would be selected whose consequences 'best matched' the observation sentences; those hypotheses in turn would serve as input to 3 where, on the basis of them, the given preferences, and the decision-theoretic functions, a 'most preferred' basic act description would be generated, and then executed by 5." It is now relatively easy to see that 6 enables the system to monitor its own states, report about them, and, possibly, modify them. Condition 7 enables the system to communicate in natural language (there is no need, of course that the English language be implemented, any natural language would do) and by condition 8, the system can entertain the thought "'I see clearly that there is nothing easier for me to know than my own mind' and proceed to insist that 'no matter what your theory or instruments might say, they can never give me reason to think that I am not conscious here, now'" (see Rey [1997], 471).

Let me now state a few hints and reasons why it seems to me that, despite elegantly and persuasively conceived arguments, they are not as convincing as they may seem.

Look at the following mathematical, computational expression:

$$X = o\, p_1 p_2 / r^2$$

What is it? Without a context we cannot tell. It could be a formula for calculating gravitational force between two bodies, or else it could be a formula for calculating electrostatic attraction force between two electrically charged objects. The expression is the same. It is the same in the abstract computational sense. But we would not thus say that gravitational force is just the same force as electrostatic force. No one would be content to make these two forces equal or say that it is in fact one and the same. Two things for which we can have the same computational expression need not be the same! Of course, this formula is very simple and differences between gravitational and electrostatic force become evident very soon. (Variables (p_1, p_2) in the expression differ to what they in fact refer: to masses in gravitational formula and to electric charges in Coulomb formula. Constant (o) is constant of gravitational attraction or electrostatic constant.)

It is not so simple, as in the case just above, if we regard complexities of conditions 1–8 above. But still, since mental processes are complex it can be the case that we can have very complex computational description or explanation which is the same for different things, for humans and for the machines; but things which fall under these at some point can start to differ. In many respects their computational characteristics can be the same but not in all, or interpretation of these abstract computational rules can at some point start to differ.

Here we come to a more general possible objection. Namely, though nicely and precisely specified, conditions 1–8 are, logically speaking, very general. It is true that many computational devices or machines would satisfy these conditions. Many computers can perform the same things. So, the explanation or description of these performings can be the same. But, sometimes, they are the same only at *some* level of explanation or description. It is very well known that the same operations can be performed differently by various computers. Though they are the same at one, more general level, they perform tasks differently when looked at some lower level. Even such simple operations as adding, can be performed in various ways. If we come down to electric currents running inside computers, we will see many differently running currents, but it could be that they are different implementations of the same program specified computa-tionally at a more abstract level. So differences begin somewhere. Perhaps we have not yet reached sufficiently refined theory and explanation which would account for consciousness and which will explain how humans do in a conscious way something which can be done also in an unconscious way.

It is also probable that present theories, however good they are, are still not mature or complete theories regarding human mind. Let me oversimplify a bit. History of science teaches us that theories develop, or better theories replace older theories, or one kind of theories are replaced by new but incommensurable theories. Whatever stance we adopt towards the development of science and concerning explanations and theories—cumulative or revolutionary—future theories could perhaps accommodate consciousness. According to the cumulative stance—new theories explain what the older ones explained and more—so they would perhaps include consciousness. According to the revolutionary stance, new theories replace the old ones after a crisis, but they are then incommensurable. New theories have other concepts and references to objects completely different from the old theories. So, perhaps new theories will accommodate consciousness. Of course, it does not have to be that way, but it's possible.

I would also like to apply, in this discussion, some of Hacking's [1988] insights on the relation between theories, experiments and interpretation in science. Instruments, experiments and theories are interrelated and even show interdependency relations. As instruments develop, they will create new data domains which new theories will explain, but old theories were also good, explanatory and stable because they fit with other kind of instruments which has created different sets of data. Sometimes, when we see something as a limitation of a theory, it has not been perceived even as data previously. Hacking ([1988], 512, italics mine) advises us, because of this, to "think of the theories as being *different* representations of *several aspects of the same reality*". He further agrees with Duhem "that nature, and even 'mechanics' or 'optics', is too complex to admit of a single unified description. One can best aim at characterizing *an aspect of parts of nature*..." (Hacking [1988], 513). Human mind and advanced computers are surely very complex things. So, in light of these facts it may be that computational theory characterizes only one aspect, or part, or even *an aspect of a part*, of the complete nature of human mind. A part of its nature can be the same as a part, or even the whole of computers or computational machines, but the whole of the human mind, in the ontological sense, may be different.

Among the other questions, Rey ([1997], 473) poses the following one: "What phenomena are unexplained without (consciousness)?". But it is not that consciousness is that which explains thinking or sensing. What we want to explain is why thinking or sensing or experiencing or perceiving is (at least sometimes) such that it is conscious. Consciousness itself is something we would like to have explained, not other mental phenomena in terms of consciousness. Consciousness is not an *explanans* but an *explanandum*.

It is interesting that various mental phenomena could be conscious: sensations, perceptions and propositional thinking. If we can devise one kind of theory which is computational, which in turn encompass processing all these mental phenomena, then perhaps we can have a common cause or

common property which makes various mental phenomena conscious—if they are conscious.

Perhaps there are no different causes why perception on the one side and, for example, some occurent belief or propositional thinking on the other, are conscious (if they are conscious). So, that could be what is missing in the complete explanation or description of human sensing or thinking.

Then, we can conceive of consciousness at least in a "minimal" sense. It is evident, at least for me, and in introspection, the difference between, for example, dreamless sleep and the waking state of walking through a gallery of fine arts. Dreamless sleep is like nothing, whereas the walk through the gallery is very vivid in subjectivity. I could not express it differently from a computer equipped with 8 above, namely, Cartesian intuition, but metaphysically speaking, difference may lie in what I refer to when I say that some of my thinking or sensing is conscious. I may be referring not only to mechanical computational process which may consist only of syntactical arrangements of elements or sequences of them and which may, no matter what it does, be like nothing (*for* the system which computes), but to some qualitative feel (which may be even *within* this computational process in cases of humans or animals)—qualia in the case of sensation.[1] This is the usual meaning—"what it is like" feature, inner quality of that qualitative feel—of the term as "consciousness", or "qualia" or "qualitative experience" or so forth.

It seems that there is no difference *for* a car when it is parked in a garage with its engine off and when on highway with the engine's computational and other vital processes running. Both states (parked state/driving state) are (*for* the car) like nothing. It seems that, at least in some cases, both the presence and the absence of computational processes can be like nothing.

But for human beings it could be different. Even if our mental states and processes are computational states and processes, there may be a difference between them—some of them being like nothing and some of them being like something to the person that undergoes them. So, computational theory of mental states and processes can be *instrumentally* satisfying explanation, but *metaphysically* speaking, that would not be a description of everything that is the case.

So we said that the usual meaning of the terms such as "consciousness", or "qualia" or "qualitative experience" is the "what it is like" feature, the inner quality of the relevant feel. But if we say that machines lack "qualia" or "qualitative experience", but nevertheless can say or report the same as humans report about inner states using the Cartesian intuition, then they do not refer to the same object as I do. Reference of machine's ut-

[1] It is much harder to say even initially what would consciously entering some propositional attitude consist of, or what would be a "conscious" component of propositional thinking. I think that there is also some qualitative notion involved in this, but I leave this matter for another occasion.

tered Cartesian intuition does not have the same referent as the Cartesian intuition uttered by me, though the same sentence is produced by both. I would refer also to these subjective qualitative feels, but machine would refer to some string of elements with the property of only syntactical arrangement which functionally or computationally constitute its inner state. *Ex hypothesi*, this string of elements is not conscious in the machine, in the usual meaning of the term "conscious".

So, since there is a possibility that there is a (metaphysically speaking) difference in reference for sentential expression (which is the same) of the Cartesian intuition told by me and by the machine, it is up to us to check whether this is so. But how can we do that? Even if consciousness does not exist, we have some conception of what it would be. Namely, it seems that it has, among others, a peculiar property that it is immediately accessible only from the first-person point of view. Only the being whose consciousness it is has immediate introspective access to it. It would not be possible to subjectively introspect consciousness of another being. All we would be able to observe, scientifically or otherwise, of other creatures, is not consciousness itself. We would not be able to observe from the third-person standpoint what it is like to be a bat from its own first-person standpoint.

So, the subject has the same epistemological position regarding possible consciousness toward the fellows of its own species and towards the machines. This is where I locate the plausibility of computational eliminativism. From the objective standpoint that is currently one of the best in our effort to explain the mind it is hard to see what other features except those non-conscious ones have a role to play. But still, from the possibility of a difference in reference of expressions of the Cartesian intuition we can have a good reason not to eliminate our efforts to retain consciousness and efforts toward its explanation.

Let me add just one other brief remark. It seems to me that perhaps we can argue the other way round too. With only a slightly reversed course of argument we can come to a significantly different conclusion. We can start from our desire to describe and explain human mentality and behaviour. Then we consider what can do that in a good and useful way. So we specify some computational theory. It can consist of conditions 1–8. With these, we complete our task satisfactorily. One surprising conclusion suddenly comes to us. Since we, humans, are no doubt conscious, it follows that electrical or electronical devices that can also perform and operate according to 1–8 have to be conscious things too.

In conclusion I would say that Rey has nicely showed how important computational theories are for understanding the mind. But I hope that the hints I have just given point out that the conclusion that consciousness does not exist is too strong and perhaps a little bit premature if we take into account the scope and development of science and scientific theories and explanations, especially with regard to computational theories. I am aware that what I say here does not establish the existence of consciousness

beyond reasonable doubt, but I hope that it provides reasons for doubting the idea that consciousness does not exist.[2]

References

Block, N., Flanagan, O., Güzeldere, G. (eds.) [1997], *The Nature of Consciousness* (Cambridge, Mass.: A Bradford Book).

Hacking, I. [1988], "On the Stability of the Laboratory Sciences", *Journal of Philosophy*, 85, 507–514.

Rey, G. [1997], "A Question About Consciousness", in Block, N., Flanagan, O., Güzeldere, G. (eds.) [1997], 461–482.

[2] I would like to thank Georges Rey, Snježana Prijić-Samaržija, Elvio Baccarini, Aleksandra Golubović, Nenad Miščević and the other participants in the Rijeka conference 2004 as well as Pavel Gregorić.