Is Consciousness not a Computational Property? — Response to Caplain

Damjan Bojadžiev
Department of Intelligent Systems, Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
E-mail: damjan.bojadziev@ijs.si, WWW: http://nl.ijs.si/~damjan/me.html

Keywords: knowledge, reflexivity, consciousness, computation, automata

Edited by: Rudi Murn
Received: October 15, 1998 **Revised:** February 12, 1999 **Accepted:** December 2, 1999

Caplain's argument that a conscious automaton would violate a certain principle of cognition is inconclusive. Its central part has the non-demonstrative form: X is sufficient for Y because Z is not and nothing else could be. The argument and the principle are also not specific to automata.

1 Introduction

Caplain has recently argued — first in a special issue of this journal and later in book form — that 'consciousness cannot be adequately described as a computational structure and (or) process' because a conscious automaton would violate a certain general principle of cognition (Caplain, 1997, p. 190). The argument seems original, interesting and quite intricate, to the point of becoming slippery in its decisive steps. Since it also runs against my own views (Bojadžiev, 1997), I was strongly motivated to analyze it (and reanalyze, and ...) in order to find its weak or missing link(s). I point out these links in section 3 and analyze the principle on which the argument is based in section 4.

2 Consciousness and knowledge

Caplain sets up his argument against automatic consciousness by recalling the distinction between knowledge and belief — knowledge as justified true belief — and connecting human consciousness with the capacity for knowledge (p. 190–2). Even this initial step already appears puzzling, since it is not obvious that the capacity for belief is any less characteristic of human consciousness. The connection seems even more puzzling when Caplain qualifies the knowledge in question as generally partial, approximate and subject to improvement (p. 190). The connection becomes clear only when Caplain moves to a different kind of knowledge and connects consciousness in general with the capacity for *self*-knowledge:

> Any conscious being, whose consciousness is active at some moment, is able to know something for sure at that moment: the fact that "there is conscious impression there" (p. 191).

Caplain then introduces apparently yet another kind of knowledge, namely

> truths which are "basic", or "primordial", in the sense that we rightfully consider them as *self-evident*, without having any clear idea of how we got to know them. Examples of such statements are: our own existence, the real existence of the world external to ourselves, the ability of our senses to provide us with some reflection of that external world (p. 191).

This kind of knowledge is supposed to illustrate what Caplain calls the reflexivity of consciousness, which is the key concept in his main argument. Caplain says that the capacity for knowledge entails the capacity for self-checking, which he calls reflexivity (p. 191). He does not spell out more exactly what reflexivity or self-checking is, so that it remains unclear how the kind of knowledge he cites illustrates this concept. Going by ordinary meaning, this third kind of knowledge is better described as consisting in *evident* (rather than *self*-evident) truths, and reflexivity is better illustrated by the previous kind of knowledge (active consciousness knowing itself). But Caplain quotes Putnam's brain in a vat scenario (Putnam, 1981), though not his argument, and adds that 'unless we fall into absolute skepticism, we are compelled to admit that strange property of reflexivity' (p. 191). This could be taken as an oblique reference to the reflexive, "counter-performative" nature of Putnam's argument. Put simply, Putnam's argument is that entertaining the notion that we are brains in a vat refutes it, a negative twist on Descartes' dictum: I think (in a vat), therefore I'm not (in it). Similarly, going again by ordinary meaning, self-checking or reflexivity might literally be the tendency of knowledge to somehow check itself, keep itself in check by preferring what is apparent (the "basic" truths above) to what may be conceivable (the pickled brain hypothesis).

3 The argument against conscious automata

Caplain's argument revolves — the verb is carefully chosen — around the question how could an automaton check or

verify its knowledge. He argues that a conscious automaton could verify that all it knows is true — or, more precisely, establish the truth of the statement that all it knows is true — merely on the basis of containing the formula of that knowledge i.e. the statement expressing that all it knows is true. This would contradict what Caplain calls the cognitive separation principle, which means that the premise of a conscious automaton has to be rejected. However, in this clash of principle and particular case, it is doubtful that Caplain actually establishes the particular case. Since this is the central part of the argument, with words and clauses in it under considerable inferential stress, an extended quotation is appropriate. Caplain uses the following notation: E is the hypothetical conscious automaton, $\Sigma(E, C)$ is the set of informations of which E is certain, recorded in it through some method C, and $T(E, C)$ is the statement expressing that all E knows through C is true, i.e. that all informations in $\Sigma(E, C)$ are true (p. 192). This statement expresses the 'true' part of the definition of the automaton's knowledge i.e. justified true beliefs, and it is itself included in $\Sigma(E, C)$, which Caplain refers to as condition 2:

$$T(E, C) \in \Sigma(E, C) \tag{2}$$

This condition now expresses the 'true' part of the definition of the automaton's knowledge for the automaton itself, and the argument now centers on the way in which the automaton verifies its knowledge, the 'justified' part:

> The realization of condition 2 would be sufficient to validly warrant to E that $T(E, C)$ is true. This is why. We could imagine, for a moment, that E makes sure of $T(E, C)$ [...] by another means: by [...] But in this case, it would be necessary that [...] In other words, *to infer* $T(E, C)$, *it would be necessary to already know* $T(E, C)$*!* Finally, the realization of condition 2 would be the only means for the automaton to get such a guarantee. Hence our condition 3: *The realization of condition 2 i.e. the recording of* $T(E, C)$ *through C, is sufficient for guaranteeing to E that* $T(E, C)$ *is true* (p. 193–4).

The reasoning here is implicit enough to invite the impression that it is merely a roundabout way of restating, rather than proving, the sufficiency of (2), the detour being what looks like an argument for its *necessity*. A clearer way of putting the argument would be this: an automaton must, as such, have sufficient reason for knowing that all it knows through C is true, and its only way of knowing is (again) through C; in particular, an alternative way which may come to mind is not available, because it would be circular. So, if the automaton knows that $T(E, C)$ is true, the sufficient reason for this knowledge can only be the recording of $T(E, C)$ through C. But what this comes down to is that $T(E, C)$ can only be known in the same way as any other $S \in \Sigma(E, C)$, namely by being recorded through C, and that was already sufficiently clear beforehand.

The major weakness of the argument quoted above is its non-demonstrative, eliminative form: X is sufficient for Y because Z is not, and nothing else could be. A stronger, positive argument would show directly that (2) *is* sufficient by showing *how* it is sufficient. Such a demonstration might take into account the special, partly self-referential character of $T(E, C)$: it says that all of $\Sigma(E, C)$ are true, and is itself included in $\Sigma(E, C)$; so, $T(E, C)$ includes itself in its statement of what is true. The inclusion of $T(E, C)$ in $\Sigma(E, C)$ could be compared to saying 'I can speak', thereby establishing the truth of what I'm saying. This self-affirming character of $T(E, C)$ might even provide a better illustration of what Caplain calls self-checking or reflexivity than the ones he offers.

Another weakness of Caplain's argument is its content: the argument is supposed to be about automata, but it does not rely in any way on their defining concepts. No mention or use is made of characteristic restrictions on structure and function, e.g. the fixed number of internal states or state transitions. Thus, the argument is not specific to (finite) automata, and it is hard to see why it would not go through for any kind of being which records information, including humans, though it would not be any more persuasive for them.

A similar point can be noted by returning to the top level of Caplain's argument. Its punchline is that condition 3, which says that (2) is sufficient guarantee for $T(E, C)$, contradicts a certain cognitive principle. Since there is much doubt as to how firmly Caplain actually establishes condition 3, the outcome could also be that a conscious automaton can know, and know that what it knows is true, without sufficient guarantee, "any clear idea of how it got to know it" (the "basic" or "primordial" knowledge above). Thus, automata might be in much the same situation as humans with respect to guarantees of knowledge, tentatively settling for evident or simplest explanations and revising them as they go along, if they must.

4 The principle of cognitive separation

Caplain formulates what he calls the cognitive separation principle for an automaton A and a conscious being E observing A. If I is the method of recording information in A and $\Sigma(A, I)$ and $T(A, I)$ are defined as above, the principle says:

> The inclusion of $T(A, I)$ in $\Sigma(A, I)$ cannot be sufficient to validly guarantee to E that $T(A, I)$ is indeed true, i.e. that all informations in $\Sigma(A, I)$ are true (p. 193).

In his main argument, Caplain uses only the special case in which $I = C$ and $E = A$. But the principle itself is easier to formulate than this special case and it also seems important in itself. In explaining the principle, Caplain says that

> it expresses that any kind of information recording in an automaton cannot contain in itself a sufficient validation of these informations (p. 193).

This formulation invites the comment that it expresses the principle better than the statement of the principle itself, with all its attendant definitions. Indeed, the whole argument with its painful details also appears superfluous, obviously decided in advance by the stipulation of the principle for automata only. Caplain does not go so far as to formulate a principle of cognitive *non*-separation for conscious beings

> The inclusion of $T(E, I)$ in $\Sigma(E, I)$ can be sufficient to validly guarantee to E that $T(E, I)$ is indeed true, i.e. that all informations in $\Sigma(E, I)$ are true

or to say that "some kind of information recording in a conscious being contains in itself a sufficient validation of these informations", but he says that 'the cognitive separation between a field of reality and recorded informations supposed to describe it does not extend to consciousness' (p. 194). By itself, this could mean either that there is no cognitive separation if the information is recorded by a conscious being, whatever the field of reality, or it could mean that there is no separation if the field is consciousness itself. Since Caplain adds that 'a conscious being builds its knowledge of reality only from conscious impressions' (p. 194), he apparently means the former, but the problem is that only the latter clearly supports his claim. That is, there is clearly no cognitive separation, or indeed much difference, between (the content(s) of) consciousness and our informations about it. But in less self-referential cases it is less obvious that cognitive separation is absent, and why it should or might be.

On the other hand, cognitive separation in humans or automata can be reduced or eliminated to the extent that the process of recording information is self-referential, providing information either about the entity in which it functions or about itself. These kinds of information amend what Caplain says in support of the principle of cognitive separation, namely that verifying that the recording process information requires 'an observation of A, I *and* the domain of reality being considered' (p. 193). If the domain is A itself, so that the automaton records information about itself, observation of A and I is sufficient for verifying these informations, but the principle of cognitive separation remains in force: it is not enough to consider what I says about A, even if I says that it is. Similarly, even if someone only talks about himself, it is no guarantee that he tells the truth if he says that he does.

At the next level of self-involvement, the recording process could turn upon itself, though this would not in itself guarantee that it provides only true informations about itself. But checking whether I provides true informations about itself would then require only an observation of I itself. Furthermore, it seems possible to construct an I which would only provide true (though possibly not complete) informations about itself, "a kind of information recording in an automaton that can contain in itself a sufficient validation of these informations". This recalls the second kind of knowledge Caplain mentions, active consciousness registering its own effects:

> having some conscious sensation at some moment entails the knowledge that, at least, there is that conscious sensation (p. 191).

This kind of self-referential knowledge would correspond to a process of automatic self-observation registering its own effects, similar to what Perlis calls self-noting (Perlis, 1997, p. 518); put this way, this kind of self- knowledge may not be that far out of automatic reach (Webb, 1980).

5 Conclusion

Caplain does not prove that consciousness is not a computational property. I do not prove that it is, much less show *how* it could be, but I indicate *why* it might be: by agreeing with Caplain's initial observation about consciousness knowing itself and noting that self-reference is something which formal systems are very good at.

Acknowledgement

I am grateful to the first referee for his detailed comments which prompted me to express myself more precisely, leave out some points which appeared less important, incorrect or in doubt, and make clearer my own position.

References

[1] Bojadžiev, Damjan (1997), Mind versus Gödel, in Gams et al (1997), pp. 202–10; HTML at http://nl.ijs.si/~damjan/g-m-c.html

[2] Caplain, Gilbert (1997), Is Consciousness a Computational Property?, in Gams et al (1997), pp. 190–4

[3] Gams, Matjaz, Paprzycki, Marcin, Wu, Xindong (ed.), (1997), *Mind Versus Computer* (Amsterdam: IOS Press)

[4] Perlis, Donald (1997), Consciousness as Self-Function, Journal of Consciousness Studies, 4 (5–6), pp. 509–25

[5] Putnam, Hilary (1981), *Reason, Truth and History* (Cambridge: Cambridge University Press)

[6] Webb, Judson (1980), *Mechanism, Mentalism and Metamathematics - An Essay on Finitism* (Dordrecht: D. Reidel Publishing Company)