Are Computational Explanations Vacuous?

Vinod Goel
Institute of Cognitive Science
University of California, Berkeley
goel@cogsci.berkeley.edu

Abstract

There is a certain worry about computational information processing explanations which occasionally arises. It takes the following general form: The informational content of computational systems is not genuine. It is ascribed to the system by an external observer. But if this is the case, why can't it be ascribed to any system? And if it can be ascribed to any system, then surely it is a vacuous notion for explanatory purposes. I respond to this worry by arguing that not every system can be accurately described as a computational information processing system.

Introduction

Computers play a very different role in cognitive science than they do in other disciplines, such as meteorology, city planning, or physics. No one claims that traffic patterns, thunder storms, and galaxies *are* computational systems. The claim is simply that many aspects of their "behavior" — in fact any aspect to which we can give an algorithmic description — can be simulated on a computer (given sufficient resources).

The cognitive science claim is of a very different nature. We do not simply claim that cognitive processes can be simulated on computational systems, but rather that cognitive processes *are* computational processes. The interesting versions of the cognitive claim (Fodor, 1975; Fodor, 1981; Fodor, 1987; Johnson-Laird, 1983; Newell, 1980; Pylyshyn, 1984) make it clear that this is not hand waving or a metaphorical way of speaking. It is meant to be taken very literally.[1]

Thus, cognitive science has a very special claim on the notion of computation not shared by other disciplines. Computation is not simply a modeling tool for us. It lies at the center of our theoretical/ explanatory apparatus. Our theories of cognition quantify over the notion of *computational information processing* and use this notion in the explanation of cognitive behavior.

There is, however, a worrisome problem with the notion of computational information processing which is often raised. On most accounts, computational information processing is *as-if*. It is a matter of ascription. But — so the objection goes — anything can be described *as-if* it is doing computational information processing. If anything can be described *as-if* it is doing computational information processing, then to explain cognition as computational information processing is not to advance a substantive thesis.

A number of researchers have noted the problem and have been worried by it to different degrees (Chomsky, 1980; Cummins, 1989; Dietrich, 1990; Fodor, 1975, p.74, footnote 15; Searle, 1984; Searle, 1990). However, no satisfactory response has been forthcoming.

In this paper I would like to respond to the form of this problem raised by Searle (1990). I will argue that, from the fact that computational information processing is *as-if*, and thus ascribed to a system, it does not follow that it can be ascribed to *any* system. In fact, there are some very stringent constraints that systems have to meet before they can be described as doing computational information processing. Both my discussion of the problem, and my response, shall be restricted to "classical" computational systems and explanations (i.e., the Language of Thought and Physical Symbol Systems type accounts).[2]

The Vacuousness Objection

Searle's (1990) specific objection takes the following form: Computation is defined syntactically. But syntax is not intrinsic to the physics. It is *assigned* to the physics by an outside observer. In fact it can be assigned to any physical system. This is disastrous because we want "to know how the brain works." It is no help to be told that "the brain is a digital computer in the sense in which the stomach, liver, heart, solar system, and the state of Kansas are all digital computers." We want to know what fact about brains makes them digital computers. "It does not answer that question to be told, yes, brains are

[1] The cognitivist claim is not the very strong claim that computation is both necessary and sufficient for cognition. It is the more interesting claim that computation is necessary for cognition.

[2] A very brief discussion of connectionist information processing claims can be found in Goel (1991).

digital computers because everything is a digital computer" (Searle, 1990, p.26).

The logical structure of Searle's argument is the following:

P1) Computation is defined syntactically.
P2) Syntax is not intrinsic to the physics.
P3) Syntax can be assigned to any system.
Therefore: Any system can be described as a computational system.

I will argue that premise P1 is false, and thus the conclusion does not follow.

I think that Searle's intuition about syntax (premise P2) is correct. The syntax of external symbol systems is not intrinsic to the physics. Perhaps the way to construe syntax is as *an arbitrary property of the world that we use to individuate elements to which we will assign a semantical interpretation.* But the notion of syntax, while necessary, is not sufficient for the notion of computation that we use in classical cognitive science (contrary to P1). We also need notions of causation and interpretability. And if there is more to computation than syntax, then from the facts that syntax is ascribed to a system (P2), and that it can be assigned to any system (P3), nothing interesting about computation follows.

The burden of my response will be to argue that there is more to computation than syntax — specifically, causation and interpretability — and these notions in turn place stringent restrictions on the assignment of syntax to physical systems for purposes of describing them as computational systems.

Structure of Classical Computational Explanations[3]

I have argued elsewhere (Goel, 1991) that what cognitive science wants/needs from computer science is a notion of information processing, where information processing requires (i) quantification over the *content* of states of the system, and (ii) a *causal implication* of that content in the behavior of the system. Such a notion of information processing is derived from our folk psychology and seems to be the the one desired by an number of writers (Dretske, 1989; Fodor, 1975; Fodor, 1987; Newell, 1980; Newell & Simon, 1981; Pylyshyn, 1984). I will call any notion of information processing which satisfies these two criteria, a notion of *cognitive information processing.*

Such a notion of cognitive information processing does not, on most accounts, satisfy the requirements of "respectable scientific explanations". It uses mental or intentional predicates, which themselves require explanation. What we need is a mechanistic account of cognitive information processing which cashes out the intentional predicates. It is for this that we turn to computer science.

As it turns out, computer science can not currently deliver such an account (Goel, 1991; Searle, 1980; Searle, 1984). It can, however, deliver an epistemic counterpart to it in the notion of *computational information processing.* The notion of computational information processing which we get from computer science involves (i) the systematic individuation of physical states into computational states, (ii) the assignment of content to those states, and (iii) the systematic recoverability of computational states and contents at each step in the trajectory of the system over time.

Minimally, such assignment and interpretability involves the following:

A) One needs to be able to (i) assign (at the initial state of the system, t=0) a subset of the physical states of the system to equivalence classes of physical states (i.e computational states); (ii) correlate a subset of the computational states with reference-classes; and (iii) once the assignments and correlations have been set up, one must be able to look at the physical states of the system and systematically recover the computational states and reference-classes. To recover computational states means, minimally, that it is possible to identify equivalence-classes of physical states and "read off" their content. In certain cases this content will be an address of another computational state or device. To recover reference-classes means, minimally, to trace through the pointers to the actual computational state or device being referred to.

B) One must be able to (i) maintain the assignment and interpretation of the system as it evolves through time; i.e., given any instantaneous description of the system one should be able to recover the computational states, the reference-classes (as above), and a pointer to the next instantaneous description in the sequence; (ii) given a temporal sequence of instantaneous descriptions, it must be the case that some set of the computational states of the instantaneous description at t *cause* the computational states and/or device activations at instantaneous description t+1, and do so by virtue of the very property which gained them membership into that equivalence class of states; and (iii) the computational story one tells of the system must parallel the causal story.

We can consider these necessary criteria for a notion of computational information processing. Any system which can satisfy these criteria may be called a CIP system.

The relationship between cognitive and computational information processing is the following: In the

[3]Parts of this and the following sections are adopted from Goel (1991).

case of cognitive information processing there is an ontic fact of the matter as to the content of a mental state, independent of assignability and interpretability. But since there can't be such a fact without genuine reference/content, the systematic assignability and interpretability of computational information processing gives us an epistemic fact, or at least that's the intuition. Similarly, in the cognitive case, we have the content of mental states causally implicated in behavior, but again there can be no such ontic fact without genuine reference/content. But again, being able to trace through the evolution of the system (by maintaining the assignability and interpretability of the instantaneous descriptions), and discovering a parallelism between the causal and logical levels in the computational case, gives us an epistemic fact; or again, that's the intuition. In going from cognitive information processing to computational information processing we are in effect trading in some ontology for epistemology, a move that is not without precedent.

To summarize, the form that I am suggesting that classical computational explanations take is depicted in Figure 1. We have a notion of cognitive information processing, derived from folk psychology, that we appeal to in explaining cognitive behavior. However, it contains mental predicates which need to be cashed out. We turn to classical computational mechanisms for this purpose. However, these mechanisms cannot directly satisfy the criteria of cognitive information processing. They can, however, give us a related notion of computational information processing, which is underwritten by a reasonably well-understood mechanism. So the strategy is to map the notion of cognitive information processing onto the notion of computational information processing and to explain the latter notion with a classical computational mechanism.

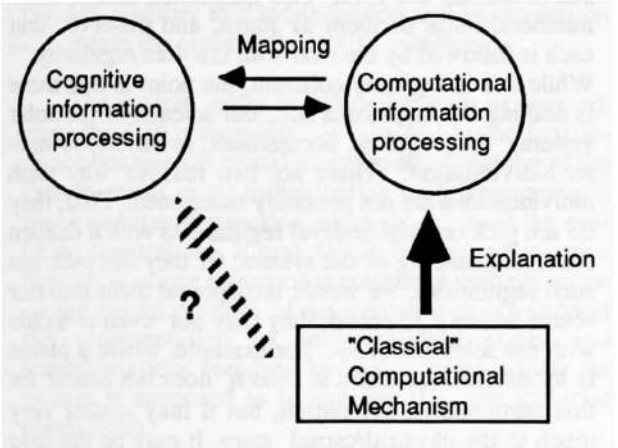

Figure 1. Structure of classical computational explanation. See text.

This, of course, leaves us with some deep questions about what, if anything, is gained by accounting for computational information processing, when our real interest is cognitive information processing. However, this question will not be pursued here. The focus of this paper is to show that the notion of computational information processing we get is not vacuous. Its relationship to cognitive information processing will be considered elsewhere.

Constraints on CIP Systems

Satisfying the criteria for computational information processing places rather severe constraints on any system. In fact, it is the case that only a system which meets the following constraints can be interpreted as a CIP system:

1) Equivalence classes of physical states of the system must be specified in terms of some function of causally efficacious characteristics such as shape or size.
2) These equivalence classes of physical states must be disjoint.
3) Membership of physical states in equivalence classes must be effectively differentiable, where differentiability is ultimately limited by physical possibilities.
4) Each state in the trajectory of the system must be "causally connected in the right way". While the specification of "causally connected in the right way" is obviously problematic, the intuition is something like the following: Certain physical states in the instantaneous description at t_n must have a direct causal connection to certain physical states in instantaneous descriptions at t_{n-1} and t_{n+1}. The connection must be such that certain physical states at t_{n-1} cause or bring about certain physical states at t_n, which in turn bring about certain states at t_{n+1}, and so on. Furthermore the transformation of the computational state CS_n at t_n into CS_{n+1} at t_{n+1} must be *realized as* the causal transformation of physical state PS_n at t_n into PS_{n+1} at t_{n+1}, where PS_n at t_n and PS_{n+1} at t_{n+1} are a subset of physical states of the system which are to be mapped onto computational states.
5) The correlation of equivalence classes of physical states with contents and/or reference-classes — within each instantaneous description of the trajectory — must be unambiguous in the sense that each member of an equivalence class of physical states must pick out the same, single, content and/or reference-class.
6) The membership of entities in reference-classes must be effectively differentiable.
7) The transformation of the system from one instantaneous description to the next instantaneous

description must be such that the above six criteria are preserved.

These are necessary constraints on CIP systems and may be called CIP constraints. It is worth noting that some are constraints on the individuation of syntactic elements, while others are constraints on semantic interpretation. If any of these constraints are violated, then some criteria on computational information processing will also be violated. For example:

- If equivalence classes of physical states are not specified in terms of some causally efficacious property, then B(ii) will be violated.
- If the individuation of equivalence classes of physical states is not disjoint, there will be not be a fact of the matter as to which computational state some physical state belongs to, thus thwarting the assignment of computational states to physical states. This would be a violation of A & B(i).
- If the individuation of computational states of the system is not effectively differentiable, then — whether they are disjoint or not — no procedure will be able to effectively make the assignment of physical states to computational states. For example, if the individuation of computational states is dense, then in the assignment of physical states to computational states, there will always be two computational states such that one cannot be ruled out as not belonging to a given physical state. This would also violate A and B(i).
- If the correlation of computational states with reference-classes is ambiguous, then there will be no fact of the matter as to the referent of any given computational state, and the systematic interpretability of the system will be impaired. This would violate A(iii) and B(i).
- If membership in reference-classes is not effectively differentiable, then no effective procedure will be able to specify which object any given computational state refers to. For example, if the reference classes are densely ordered, then in the assignment of objects to classes, there will be two classes for any object O, such that it is not effectively possible to say that O does not belong to one. This would violate A(ii, iii) & B(i)
- If the causal constraint is violated, we will not get an isomorphism between the physical and computational story and violate B(ii, iii). Furthermore, we will get the absurd results that time-slice sequences of arbitrary, unconnected patterns (e.g. the conjunction of the physical states consisting of craters on the moon at t1, the meteor shower on Neptune at t2, the food on my plate at t3, the traffic pattern on the Bay Bridge at t4, etc.) qualify as computational systems.
- If at any instantaneous description of the system, any of the above constraints are violated, then at that point some constraint on computational information processing will be violated.

Not Every System is a CIP System

The final step in the argument is to show that not every physical system is a CIP system, and that there is indeed a fact of the matter as to whether some system is, or is not, a CIP system. Given the nature of CIP constraints, determination of CIP systems can be made at just the syntactic level, or the syntactic and semantic levels. Both situations are discussed below.

Syntactic Individuation

Let's take a particular dynamical system — for example, the solar system — and ask whether it is a CIP system. If we accept the physical/causal story given by Newtonian mechanics — which recognizes things like planets, gravitational force, the shape of orbits, etc. — and use it to individuate the states and transformations of the systems (which are mapped onto computational states and transformations), our question becomes something like, "do the orbits of the planets around the sun constitute a CIP system?". I think one can unproblematically say they do not. For one thing, the instantaneous descriptions of the system will be densely ordered and thus violate the effective differentiability constraints.

Of course, it is possible to take the solar system and individuate components and relations in such a way that the CIP constraints are met. For example, a colleague suggested the following individuation: "we can divide up the orbit into quadrants, assign them numbers, think of them as states, and observe that each is followed by the next with law-like regularity."[4] While this is logically coherent, the point is that there is nothing in our physics (i.e., our science of the solar system) that requires, necessitates, or sanctions such an individuation. There are two reasons why such individuations are not generally sanctioned. First, they do not pick out higher-level regularities which deepen our understanding of the system. (If they did pick out such regularities, we would incorporate them into our scientific story.) Second, they may not even coincide with our scientific story. For example, where a planet is located in a quadrant at time t_i does not matter for this particular individuation, but it may matter very much to the physical/causal story. It may be the case that particular locations in the quadrant are associated with varying degrees and types of causal interactions with other heavenly bodies. If this is the case, this

[4] Kirk Ludwig

individuation does not coincide with our physics and can be dismissed on that basis.

Semantic Individuation

Can we make the same claims about the semantic constraints? Given an arbitrary dynamical system, can there be a fact of the matter as to whether it does, or does not, satisfy the semantic constraints on CIP systems? If one chooses not to interpret the system semantically, clearly there can be no such fact. The question will never arise. However, the important point is that, if one does choose to interpret the system, then *relative to a specific individuation of states and transformations* (ie. a particular syntactic individuation) *and a specific semantic interpretation*, there is a matter of fact as to whether the system is a CIP system or not. To get this matter of fact, one proceeds as follows:

(i) Decide on the system and phenomenon you are interested in and the level at which it occurs.
(ii) Understand the system/phenomenon on its own physical/causal terms; i.e., explicate the structure and dynamics of the system which are causally relevant in the production of the phenomenon under investigation.
(iii) Use the physical/causal structure to individuate equivalence classes of physical states and transformations which are to be assigned to computational states and transformations (i.e., the syntactic interpretation).
(iv) Specify the program the system is supposed to be running (i.e., the semantic interpretation) and again use the causal structure and dynamics of the system to interpret the computational states and transformations.
(v) Ask whether this individuation and interpretation meets the constraints on CIP systems.

The system under investigation may or may not meet the CIP constraints. It may fail in the first instance because the causal structure and dynamics of the system result in an individuation of (computational) states and transformations which do not meet the syntactic constraints. It may fail in the second instance because — since reference is correlated with causation — the causal network of the system may not support the interpretation of computational states and transformations required by the program which the system is supposed to be running (i.e., the semantic constraints).

Is our stomach — as a processor of food — a CIP system with respect to a certain individuation and interpretation of computational states and transformations? It is an empirical question. There is no *a priori* answer independent of the causal structure and dynamics of the system and a specific semantic interpretation. One needs to proceed as above and discover the answer. Is our brain a CIP system under the relevant individuation and interpretation of computational states and transformations? That is, do the structure and dynamics of the brain which are causally relevant in the production of mental life satisfy the CIP constraints? Maybe they do; maybe they don't. It is, as cognitive science claims, an empirical question.

Since the facts about CIP systems are relative to some individuation and interpretation of computational states and transformations, they need not be unique facts. A system may turn out to be a CIP system with respect to several individuations and interpretations. But there is no reason to believe that it will turn out to be a CIP system with respect to every individuation and interpretation because the CIP constraints tie the individuation and interpretation into the physical/causal structure of the system.

Conclusion

If it is indeed the case that (i) we appeal to computational systems for a notion of computational information processing, (ii) only CIP systems can satisfy the criteria on computational information processing, and (iii) not every system is a CIP system, then from the (correct) premise that syntax is not intrinsic to the physics, it does not follow that the notion of computation as used by cognitive science is vacuous.

Indeed, to say the brain is a computer is to make a very substantial empirical claim. What cognitive science is doing by appealing to computation — and claiming it is a necessary condition for cognition — is putting forward the empirical hypotheses that the mechanism that underwrites computational information processing is the very same mechanism which underwrites cognitive information processing. This mechanism is a dynamical system that satisfies the CIP constraints. Thus the cognitive system on this view is accurately described as a CIP system. This claim is not vacuous, nor harbors an homunculus. It may of course be false, but that is a separate question which can be determined only by empirical enquiry.

Acknowledgements

The author is indebted to John R. Searle and Brian C. Smith for both inspiration and assistance in the course of developing this argument. They are of course not responsible for its shortcomings. This work has been supported by a Gale Fellowship, a Canada Mortgage and Housing Corporation

Fellowship, and a research internship at System Sciences Lab at Xerox PARC and CSLI at Stanford University.

References

Chomsky, N. (1980). Rules and Representations. *Behavioral and Brain Sciences, 3*, 1-61.

Cummins, R. (1989). *Meaning and Mental Representation*. Cambridge, Massachusetts: The MIT Press.

Dietrich, E. (1990). Computationalism. *Social Epistemology, 4* (2), 135-154.

Dretske, F. (1989). Putting Information to Work. In P. P. Hanson (Eds.), *Vancouver Studies in Cognitive Science*. Vancouver, Canada: University of British Columbia Press.

Fodor, J. A. (1975). *The Language of Thought*. Cambridge, Massachusetts: Harvard University Press.

Fodor, J. A. (1981). Methodological Solipsism Considered as a Research Strategy for Cognitive Psychology. In J. Haugeland (Eds.), *Mind Design*. Cambridge Mass.: MIT Press.

Fodor, J. A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, Massachusetts: The MIT Press.

Goel, V. (1991). Notationality and the Information Processing Mind. *Minds and Machines, 1* (2), 129-165.

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge, Mass.: Harvard University Press.

Newell, A. (1980). Physical Symbol Systems. *Cognitive Science, 4*, 135-183.

Newell, A., & Simon, H. A. (1981). Computer Science as Empirical Inquiry: Symbols and Search. In J. Haugeland (Eds.), *Mind Design*. Cambridge, Mass.: MIT Press.

Pylyshyn, Z. W. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science*. Cambridge, Massachusetts: A Bradford Book, The MIT Press.

Searle, J. R. (1980). Minds, Brains and Programs. *The Behavioral and Brain Sciences, 3*, 417-457.

Searle, J. R. (1984). *Minds, Brains and Science*. Cambridge, Mass.: Harvard University Press.

Searle, J. R. (1990). Is the Brain a Digital Computer? In *Sixty-fourth Annual Pacific Division Meeting for the American Philosophical Association*. Los Angeles, CA, March 30, 1990.