The Grandmother-Sensor Fallacy

Chris Thornton
Cognitive and Computing Sciences
University of Sussex
Brighton BN1 9QN
Email: Chris.Thornton@cogs.susx.ac.uk
Tel: (44)273 678856

December 14, 1994

CSRP 360

Abstract

The view that cognizers make use of detailed, internal, world models is not as widely accepted as it once was. However, its rejection turns out to entail absurd assumptions about the way in which cognizers implement behavioural contingencies. The implication seems to be that theories of world-model use should not be dispensed with prematurely.

1 Introduction: strong representationalism in the dock

'Strong representationalism' is a hypothetical doctrine whose central claim is that cognizers are able to do the things they do because they have detailed, internal world models. This claim is, indisputably, a strong one. It implies that a cognizer has an abstraction for every object, and for every static, dynamic, and relational property of its environment. It also implies that, if that environment exhibits structure (i.e., hierarchies in which objects and properties are both constituted of, and constituents of lower and higher-level objects and properties), the cognizer will have internal structures of abstractions which precisely mirror the structures of the environment.

Arguably, the plausibility of strong representationalism has been weakened by the recent developments in Artificial Intelligence and Cognitive Science, (e.g., [1,2]). These have tended to lead the field away from the 'classical position' and its firm emphasis on the importance of representation (see also [3]). But it would be unfortunate if these developments caused us to throw out the baby with the bathwater. Classical AI may not have provided all the answers; but the stress on representation may not have been entirely misplaced.

The present paper argues that representation is, in fact, quite appropriately viewed as a key concept for cognitive science. It shows that strong representationalism, when developed from first principles, is a compelling thesis. Once we look closely at the operational constraints applying to cognizers which produce contingent behaviours, certain implications come to light. These turn out to firmly support the strong representationalist view of cognition and to provide general reinforcement for the epistemologically interesting view that cognition must, in some sense, 'recapitulate physics'.

2 The abstract 'cognizer'

The argument to be developed makes use of a high-level abstraction intended to cover all those biological and computational mechanisms which we typically think of as 'cognizers'. The properties of the 'abstract cognizer', as I will call it, are quite simple. It exists in an environment. Phenomena (events and things) in the environment impact the cognizer via sensory inputs; and the cognizer impacts the environment via motor outputs (behaviour). The conceptualisation is illustrated diagrammatically in Figure 1.

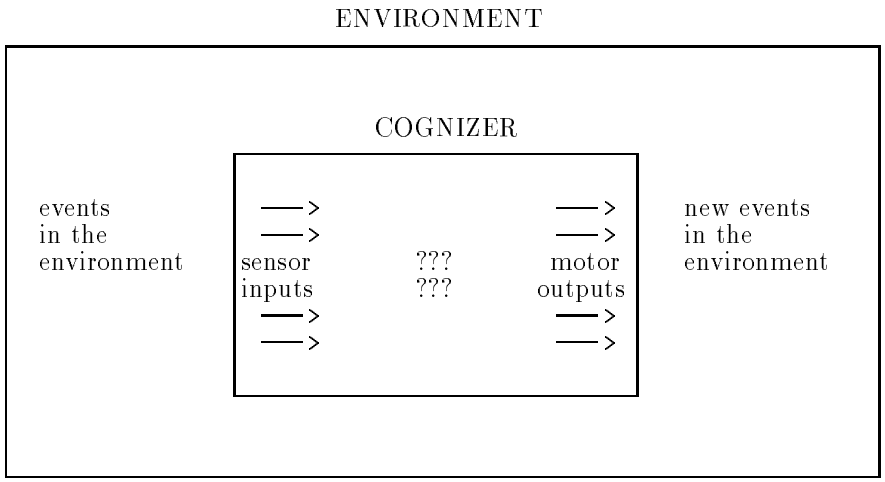

Figure 1:

A behaviour of the abstract cognizer (henceforth just the 'cognizer') is assumed to be either

- contingent on one or more previous sensory inputs, or
- non-contingent, i.e., spontaneous.

I do not rule out the possibility of behaviours having both a contingent aspect and a spontaneous aspect. However, I will assume that any degree of contingency is sufficient to justify calling the relevant behaviour 'contingent'. I will also assume that any realistic cognizer (i.e., any realistic instance of the abstract-cognizer model) must be capable of responding to a large range of environmental scenarios and must thus have a large repertoire of contingent behaviours.

By definition, a cognizer that exhibits a contingent behaviour must *implement* the relevant contingency. This means that it must somehow distinguish the sensory inputs which populate the contingency in question from sensory inputs that populate other contingencies. For present purposes, I will distinguish between two implementation strategies. I will say that the cognizer 'captures' the contingency 'explicitly' if the implementation explicitly identifies (in absolute terms) any of the sensory inputs which feature within it. And I will say that the cognizer captures the contingency 'implicitly' if the implementation does not explicitly identify any of the relevant sensory inputs.

A concrete example may help here. Imagine that we are dealing with a humanoid cognizer and the behaviour of 'rapid finger withdrawal'. This might be produced contingently following sensory inputs

which indicate the close proximity of the finger with an extreme heat source. The contingency is captured 'explicitly' if and only if its implementation in any way explicitly identifies (in absolute terms) one of the sensory inputs that features within the contingency. This might occur, for example, if the implementation involved the explicit identification of a 1-millisecond firing rate for a particular sensory cell as being specifically indicative of the presence of an extreme heat source.

3 Sensory-input spaces

I now turn attention to the 'input' side of the cognizer. My main assumption here is that a cognizer has a number of sensors, each of which produces a range of sensory inputs. This assumption allows us to envisage the combination of inputs received via a set of sensors in terms of points in a 'sensory-input space'. The general idea is illustrated in Figure 2. In the situation shown we have two sensors named 'sensor1' and 'sensor2'. The combination of inputs received simultaneously from these two sensors can be conceptualized in terms of points in the 2-dimensional space. I will refer to such points as 'sensory events'.

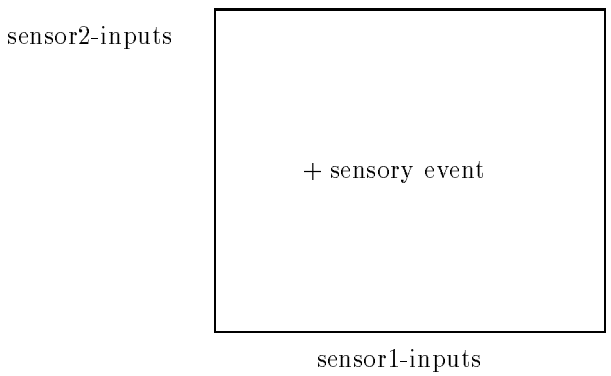

Figure 2:

The minimal contingency for a particular (contingent) behaviour is a single sensory input from a single sensor. However, in general, a contingency can be expected to be populated by *combinations* of sensory inputs. (There are lots of ways a flame can burn your finger but they all cause you to remove it from the flame.) Thus we can say that the contingency for any particular behaviour effectively instantiates a set of points (sensory events) in some particular sensory-input space.

So let us, then, define a 'sensory-event class' as the collection of sensory events which populate the contingency for a particular behaviour. Let us further distinguish two ways in which two or more sensory-event classes might instantiate the same sensory-input space. The simplest situation is illustrated in Figure 3. This shows a 2-dimensional (2 sensor) sensory-input space. The points within the space labeled 'A' are points within the contingency for behaviour A. The points labeled 'B' are within the contingency for behaviour B. Note that points in the two classes occupy distinct localities of the space; they do not mingle together. I will call classes which co-instantiate a sensory-input space in this way 'localized classes'.

In contrast we have the situation illustrated in Figure 4. This shows the same sensory-input space

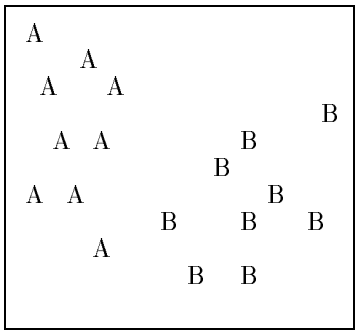

Figure 3:

populated with points from the same two behaviours. However the pattern of instantiation is very different. Points in the two classes now mingle together in a haphazard fashion. I will call classes that co-instantiate a sensory-input space in this way 'distributed classes'.

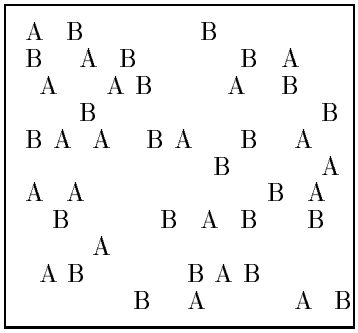

Figure 4:

4 Implementation strategies for localized and distributed classes

The distinction between localized and distributed instantiation classes enables us to draw inferences about the ways in which contingencies can be captured. Note that in the fully localized scenario, absolute sensor values are informative regarding the class-membership of events. For example, a particular sensor value may be only instantiated by members of one class. Or, it may form a 'boundary' between members of one class and members of another. However, as we approach the fully distributed scenario, absolute sensor values become uninformative regarding class-membership. In the extreme case (uniform distribution of points from all relevant classes throughout the space) an absolute value is associated with (or bounds) equal numbers of events from all classes.

From this we can deduce that explicit capture of contingencies for distributed classes is likely to be

impractical. In the fully-distributed scenario, a particular sensory-input can only be informative with respect to the class membership of the single sensory event of which it is a component. Thus, the cost of explicit capture under a distributed scenario is effectively tied to the number of distinct points in the relevant sensory-input space. If we have a space based on 10 sensors, each of which produces a range of 100 values then an explicit capture of all the classes instantiating that space (assuming uniform and exhaustive instantiation of points) would involve the explicit identification of 100^{10} (100 million, million, million) distinct sensory objects. From this we can rapidly infer that, for cognizers with non-negligible numbers of contingent behaviours, explicit capture of distributed classes is implausible.

Having ruled out explicit capture of distributed classes we are left with implicit capture as the only feasible option. It is worth observing in passing that an implicit capture of a contingency is *literally* an abstraction of that contingency since it identifies it in a manner that abstracts away the absolute sensory-inputs in which the contingency is ultimately grounded.

5 Grandmother sensors

The core of the argument can now be unraveled. Its central component is a definition for the term 'grandmother sensor'. For present purposes a grandmother sensor is simply a 'grandmother cell' treated as a sensor. Recall that a grandmother cell [4] is a neuron in the brain which produces activity only in response to some high-level phenomenon (e.g., one's Grandmother), and is thus plausibly regarded as some sort of 'representation' of that phenomenon. In similar fashion, a grandmother sensor is a sensor which produces a given input only in response to some particular high-level phenomenon. However, the definition is a little more restrictive. A grandmother sensor is defined to be a sensor which produces a given input only in response to a particular *behavioural contingency*. Moreover, the grandmother sensor is assumed *not* to have any internal (e.g., representational) structure. For the purposes of the argument, then, a grandmother sensor is a *primitive*.

Having proposed a definition of 'grandmother sensor' it is now possible to implement the underlying aim of the exercise, which is to take note of the fact that, in general, grandmother sensors *do not exist*. For most cognizer/contingency combinations, this assumption seems to me to be virtually beyond dispute so I will not offer any long-winded justification for it. Humanoid cognizers, for example, just *do not have* sensors for news items concerning politicians, broken lawnmowers or foods containing animal products even though these things may well populate behavioural contingencies.

An immediate implication of the assumption is that cognizers cannot capture a particular contingency in terms of absolute sensory inputs since in this case the sensor(s) in question would form grandmother sensors for the contingency. And, clearly, if absolute sensor values cannot play any role in the capture of a contingency then they cannot be informative with respect to the relevant class-membership of sensor events. This tells us that event classes cannot be localized, they must be distributed.

We thus have the following chain of deductions. From our assumption that cognizers do not have grandmother sensors we deduce that sensor-event classes are necessarily distributed. This enables us to conclude that sensory-event classes must have implicit captures, which as we noted above, form literal abstractions of the relevant contingencies. A cognizer without grandmother sensors thus necessarily forms an abstraction for every phenomenon of its world — including every object, and every static, dynamic or relational property of that world — which features in a behavioural contingency. If its world is structured then the cognizer necessarily has internal structures of abstractions which precisely mirror the structures in the world.

These last two implications are, of course, the tenets of 'strong representationalism', introduced at the

outset. The chain of deductions thus supports the view that cognizers are able to do what they do because they have detailed, internal world models.

6 Commentary

The argument developed shows that, under plausible assumptions, cognizers without grandmother sensors must have detailed world models. It provides general support for the notion that representation is a key process within cognition. It also provides some sort of justification for the high-level of attention that has traditionally been given to the topic, particularly in classical AI and connectionist work.

The rather extreme implications of the argument should not be underestimated. The argument tells us that grandmother-sensorless cognizers must have accurate internal models of all behaviour-affecting phenomena in their environment. It tells us that these models will tend to recover the structural, relational and hierarchical properties of that environment. In the case of more developed cognizers, the recovery will include structural and relational properties of the relevant social and cultural context. And certainly, for all imaginable cognizers, it will include all structural and relational properties of the physical context. (Hence the slogan 'cognition recapitulates physics'.)

There seems to be no obvious way to defeat the argument. However, there are various ways in which its implications can be dodged. A simple strategy is to deny the validity of one of the argument's assumptions. For example, one might propose that the 'abstract cognizer' conceptualisation is far too abstract to be a meaningful model of any realistic cognizer. Alternatively, one might propose that cognizers *do* have an abundance of grandmother sensors and thus have no use for representational structure.

The implication that cognizers engage in world-modeling can also be eluded in a number of ways. The argument shows that a cognizer uses abstractions for phenomena within its behavioural contingencies. But since there may be phenomena within a cognizer's environment that do not feature in any contingency, there may also be phenomena of the relevant world that lack any internal abstraction. (As an example, consider the phenomenon of 'compound interest'. Presumably this does not feature within any behavioural contingency of the Panamanian fruit-fly even though it is 'within' that organism's physical universe.) If we insist that to be counted as engaging in world-modeling, the cognizer must be guaranteed to abstract *all* phenomena of the relevant world, then the grandmother-sensor argument does not force us to accept that cognizers do world-modeling.

A further variant of this strategy can be developed. Consider those sophisticated cognizers, namely humans, who engage in explicit scientific enquiry. For such cognizers there is the possibility of being able to effectively identify certain phenomena of their world regardless of whether those phenomena feature within any human, behavioural contingency. The phenomenon of 'quantum non-determinism' forms a plausible example. Humans may be able to identify this higher-order, dynamical property of their world, but it is not obvious that it would feature within any human behavioural contingency. On the basis of this observation one might go on to argue that a human cognizer's world must therefore be much more than the sum total of phenomena featured within its behavioural contingencies, and that having abstractions for those phenomena cannot therefore be counted as world-modeling. But the final effect of this proposal is not exactly satisfactory. It seems to tell us that human beings are the only cognizers who can be guaranteed *not* to do world-modeling, which, as a general conclusion, runs rather counter to intuition.

Acknowledgements

I thank Matthew Elton and Ron Chrisley for drawing my attention to shortcomings in the original presentation of the argument.

References

[1] Brooks, R. (1991). Intelligence without representation. *Artificial Intelligence, 47* (pp. 139-159).

[2] van Gelder, T. (1992). What might cognition be if not computation?. Research Report 75, Bloomington, IN47405: Cognitive Science, Indiana University (Indiana).

[3] Clark, A. and Toribio, J. (1994). Doing without representing?. CSRP 310, Cognitive and Computing Sciences, University of Sussex, UK.

[4] Barlow, H. (1972). Single units and sensation: a neuron doctrine for perceptual psychology?. *Perception, 1* (pp. 371-394).