REVISING ALLEN NEWELL'S CONCEPTION OF REPRESENTATION

Petros A.M. Gelepithis

Kingston University, England

Abstract

Representation is one of the fundamental notions of Cognitive Science and it is closely related to several others (e.g., knowledge, symbol, information). Today the most elaborate and integrated treatment of representation is that of Allen Newell (1990). Yet even his analysis does not go into the depth required by such a fundamental concept. The aim of this paper is to present a deeper analysis of representation and accordingly set the basis for revising Newell's conception. In particular: First, following some remarks on the nature of human representations, a wider view of representation as a particular type of relation rather than Newell's conception of one–to–one function is presented and two fundamentally different kinds of representational systems are distinguished. Second, we suggest that the closely linked processes of understanding and communication are a suitable basis for exploring the mechanisms materialising our representational capacity, an issue entirely ignored by Newell.

Finally, five closely related issues are briefly introduced which provide a preliminary framework for the further study of representation.

Key Words: Human representations, computer representations, representation, representational relation, representation law, representational capacity, representational system.

1. Outline of Newell's View of Representation and a Summary Critique

The earliest treatment of the nature of representation, within an overall theoretical framework concerning the foundations of cognitive science, is that of Allen Newell (1990). An outline of his views follows. In his Unified Theories of Cognition Newell adopts an approach reminiscent of induction. He introduces three different representations of a simple blocks world and on that basis he proceeds to describe abstractly what is going on in the three representations

Revised version of a paper presented at the 12th Annual Workshop of the European Society for the Study of Cognitive Systems, Camogli (Genoa, Italy), 12–14 April 1994.

chosen. He summarises that description in what he calls "the representation law", namely:

"decode[encode(T)(encode(X))] = $T(X)$ where X is the original external situation and T is the external transformation". (ibid. p. 59). (Note 1)

which is subsequently assumed, without argumentation, that it captures the essence of representation (Newell 1990 p. 59; and 1992 p. 426).

He then proceeds to identify two major types of representational systems: (i) an analogical (or specific) representation system like a map; and (ii) a composable (or general purpose) representation system like logic or Lisp. His treatment of representation ends with a brief introduction of five requirements that a good composable representation system should have. Of these, four are well known like expressive power and completeness of the representing medium. The fifth property, called *executability* by Newell, is a direct consequence (Note 2) of the need, eventually, of using the representations. Here it is how Newell describes medium executability:

"The medium itself can be passive, and then it must be interpretable by some other process, which is the familiar situation with programming languages. Alternatively, the medium can be active, and then it must be evocable. In either case, there is a dual aspect in which the medium is passive while being composed and then shifts at some controllable moment to being active." (Newell 1990, p. 64).

In our view, medium executability brings to the fore a very important point, namely, the fact that representation and reasoning, or representation and computation, or representation and inference, or, finally and most generally, concepts/thoughts and thinking are inseparably linked. Naturally, these pairs are most likely considerably different from each other even with respect to 'representation' but this is, regardless of how important it may be, besides the point in our current discussion (Note 3).

My pivotal point of criticism is that Newell's analysis of the notion of representation does not go into the depth required by such a fundamental concept. Specifically: First, his claim, that the representation law constitutes the essence of representation, presupposes that the key objective of representation (which is really what the representation law is), is an adequate specification of the nature of representation. This conception may be not surprising in view of his theory of knowledge (ibid.) but it needs to be supported and Newell provides no support at all. Second, his examination of composable representation systems is constrained to those designed by humans; but such human-made composable representation systems do not have to constitute the representational basis of non-human intelligent systems. This design constraint is unnecessary in intelligent systems development. Finally, he takes for granted the very capacity

enabling a representation user to appropriately associate the represented and representing domains. What is involved in this fundamentally important ability is far from clear and Newell neither considers it nor even acknowledges it.

In correspondence to the three points stated above, this paper aims to contribute to a deeper analysis of the notion of representation by: (i) presenting a wider view of representation as a particular type of relation rather than Newell's conception of one–to–one function; (ii) introducing two fundamentally different types of computer representations, the existence of which shows that the class of human–made composable representation systems is indeed a genuine subset of the class of composable representation systems and hence the requirements of the former class do not constitute an adequate basis for the development of the latter; and (iii) suggesting that the processes of understanding and communication enable (human) representation creators and/or users to choose whatever appropriate representations may be required.

2. A Deeper Analysis of the Notion of Representation

We start our conceptual analysis by remarking that human representations are everywhere. A human:

a) may observe them in designs of all sorts and in small scale models like those used to represent a major urban development, a spacecraft or a teddy bear,

b) uses them when he/she employs logic, camera images, physical laws, geometry, computer programs, equations, or thought processes to specify a rocket's orbit, or automate a 'repetitive' job,

c) may admire them, like or dislike them etc., in their artistic form in theatre, painting, sculpture, or cinema to name the most obvious ones,

d) may deduce them from behavioural acts of, say, a dog, a computer, or, presumably a Martian. And finally,

e) may simply possess them like perceptual 'images'.

Although the empirical evidence presented above has been readily available for anyone to see it was always somehow ignored. Even in cases where the significance and centrality of representations were stressed, and the difficulty of choosing appropriate representations was illustrated by examples (see, e.g., Anderson 1990) no further consideration was given and, of course, no characteristics of representation were (sought to be) specified. Below we do just that.

In all cases they are characterised by the following properties: (i) they simplify a situation, say, S_1; (ii) they preserve the essential characteristics of S_1; (iii) they can be processed by humans; and (iv) they are part of a physical material,

say, M (Note 4). On this basis we define: a representation of a situation, say, S_1, is another situation, say, S_2, characterised by the properties (i) and (ii) above. In general, a representational relation is a situational relation R: S - >R(S), such that:

- $S_2 \in R(S)$ simplifies $S_1 \in S$.

- $S_2 \in R(S)$ preserves the essential characteristics of $S_1 \in S$.

Naturally, we call the domain of R the represented class of situations (equivalently the represented world), and the range of R the representing class of situations (equivalently the representing world).

The first point to be made is that in contrast to Newell's notions of encoding and decoding which are assumed to be functions (actually decoding should be the inverse of encoding), different representations may well be images of a single represented situation. In other words, R is a much wider notion than that of encoding.

Seen from this perspective, at least five, more or less related, classes of issues emerge. First, the problem of the genesis and storage of mental representations. Second, the question of our representational capacity. Third, the issue of representational system(s). Fourth, the relation between mental and external representations. Finally, the interdependence of a cluster of notions which seem to be inseparably linked to that of representation (either seen as relation, or as a special function a la Newell). In what follows we continue our analysis of the nature of representation as a prelude to the issue of representational system(s). The next section will touch upon the rest as a brief introduction to further work.

Consider. Representations in (a) are not in human language or any other form of notational convention. Representations in (b) are in a 'language'. Those in (c) may employ cultural and/or linguistic conventions as well as expressions of feelings, therefore, they may be in both linguistic and non–linguistic materials. Representations in (d) are in the behavioural material. Those in (e) are brain representations. Essence–preserving simplifications are of two kinds (Note 5): those created/found/modified in CNS; and those expressed in non–CNS terms (e.g., natural language terms, toy models, painting, etc.). Still their considerable, and in some cases remarkable and even extreme, differences are differences of degree not of kind; specifically they differ in the degree of their formalisability and interpretability. On the other hand, they all have something of fundamental importance in common: they are all human–dependent. Talking of the nature of representations though (not of human representations only), an important difference of kind does exist and is due to the existence, in principle, of two fundamentally different types of computer representations. A summary of the point follows.

Currently, work in AI has concentrated on developing theories and techniques that enable us to build machines with cognitive, perceptual or motor capabilities

to carry out jobs we assign to them. All such developments are characterised by representing aspects of human knowledge, problems, reasoning processes, sensory information, etc. They require from us humans to express our problems in a way a computer, (or a computer–based machine) will be able to process. Such machines employ computer representations of the first type. This type of representation is of course of fundamental importance since: (i) it is closely linked to the issue of the formalisability of human knowledge and reasoning; and, consequently, (ii) it imposes a constraint on the range of 'intelligent machines' that humans, on their own, can develop. This is because computer representations of the first type are human–dependent and, subsequently, so are the developable 'intelligent machines'. The emergence of computational machines with their own representational systems would free them from any human dependence (an entity E possesses its own representational system, Re, if and only if, Re is independent of the language of another kind of entity E^*). Computational machines with their own representational systems define computer representations of the second type. This second type, when achieved, will represent states of affairs which will not be constrained to be related to human knowledge, problems, reasoning processes, etc. For the reasoning supporting these points and discussion of some of the consequences see Gelepithis (1991).

So, all types of human representation and the computer representation of the first type differ only in degree; they are all, eventually, describable in terms of human primitives. The second type of computer representation is on a class of its own; representations within this class would, eventually, be describable in terms of machine primitives. So, there are at least two fundamentally different kinds of representational systems: those based upon: (i) human primitives (HP–based); and upon (ii) machine primitives (MP–based).

3. An Outline of Some Important Issues and Related Discussion

The following paragraphs introduce some of the important representational issues and in a couple of instances a hunch for further exploration is offered.

First, given (or assuming) that the nature of representation sketched in this paper is true, what are the mechanisms for the genesis and storage of such representations? Would such mechanisms be related to the ones postulated (Rolls 1987) for 'information' representation and processing at the single neuron level? Alternatively, would a Darwinian framework of explanation for mental representations (Changeux 1983*1985; Edelman 1987; Edelman and Mountcastle 1978) be applicable to such mechanisms?

Second, whatever representational mechanisms may be found, how do we, (or any entity capable of representations), come about selecting essence–preserving simplifications in the first place? What are the mechanisms and scope of such a representational capacity? To what extent can a situation be simplified

so that its essential characteristics are not lost? Let us consider the following problems. An expert automobile engineer is asked to explain the function(s) of an automobile to:

a) a learner driver; and
b) a 2–3 years old child.

Alternatively, an explanation of electricity is sought for:

a) a 2–3 years old child;
b) an electrician; and
c) a theoretical physicist.

Clearly, the child is unable to understand notions like that of the clutch pedal or internal combustion-engine principles. On the other hand, it can easily understand the (potential) danger to his existence by a moving car. Similarly, a learner driver although able to understand at least some of such descriptions he is simply not interested in learning about all of these things. Most likely, he is just interested in the ones enabling him to drive a car and pass his exams. A description of an automobile's functions in terms of the steering wheel and the gas, brake, and clutch pedals, may well be adequate for a learner driver. Similarly, an explanation of electricity will vary from a description in terms of, say, the switch and the appearance/disappearance of the light, through a technical grasp of Ohms' law, to that of quantum electrodynamics. We conclude that knowledge of the internal cognitive state of a recipient is a fundamental prerequisite for successful explanations.

On the other hand, assuming that the internal cognitive state of a recipient is sufficiently known for our purposes, we shall have to decide on: (i) what is to be represented (technically known as the scope of a representation); and (ii) the amount of detail of whatever is represented (technically known as the grain size of a representation). These are the two major types of simplifying conditions for representations. The reader may just look at any picture for a while and subsequently consider a stick figure representation of it to see these two simplifying conditions at play. Our two problems above can provide one with substantially more elaborate examples.

Assuming that an explanation of something, say S, for a particular person, say X, is a description of S in terms that X can understand; and using a definition given before (Gelepithis, 1991) definition of understanding (roughly reducibility to one's own primitives) we have that an explanation of S for X is a description of S in terms of X's premises. On this basis we may reasonably propose that the processes of understanding and communication enable (human) representation users to choose whatever appropriate representations may be required. In particular, they would enable representation users to specify the purpose of representation which in turn guides such a user to specify: (i) the scope of a representation; and (ii) the grain size of a representation.

We now come to the notion of a 'representational system'. In Artificial Intelligence, a representational system (or knowledge representation (KR) scheme, or representation for short) is equated to the notion of a model (Note 6) (see, e.g., Winston 1992). Newell's remark on "the great move" (Newell 1990, 1992) reflects exactly this conception. In what follows we use the term KR scheme for this conception of a 'representational' system. To what extent can existing KR schemes (e.g., logic, semantic networks) serve the requirements of representational systems as outlined in our analysis? What are the consequences of our conception for the restricted language and restricted classification theses debate (see, e.g., Doyle and Patil 1991)? Our hunch is that it will make the efficiency requirement supporting the two theses (e.g., Brachman et al 1983; Levesque and Brachman 1987) unnecessary. How could one reconcile the abstract, uninterpreted nature of the vocabulary elements of a composable representation system with the characteristic properties of a representation?

Let us see Newell's conception, and some of the problems it brings with it, from still another angle. In general, assuming the existence of an initial situation, S_i, a final (e.g., goal) situation, S_f, and a class of transformations T linking the two, there are three types of well-known problems. First, S_i and T given, and S_f required to be specified (i.e., the problem of prediction). Second, S_f (e.g., present situation) and T given, and S_i (e.g., earlier situation) required to be specified (i.e., the problem of explanation). Third, S_i and S_f are given and T required to be specified (i.e., the problem of specifying the mechanisms involved).

As an illustration consider the third class of problems above. How the sought, internal transformations are specified? Trial and error must be ruled out in view of the virtually infinite number of possible representational transformations. Accumulated experience and training, and the resulting expertise, might suggest an alternative answer but that faces similar problems. For example, even within a composable representation system like logic which is essentially pre-moulded and thus of a significantly reduced expressive power, logicians' creativity does occur. How?

Fourth, there is the issue of the relation between mental (sometimes referred to as internal) and external representations. For example, what are the mechanisms for producing external representations suitable, as an explanation, to person X? Could natural-language generation techniques be used as a first approximation to such mechanisms?

Finally, a large class of problems concerns the compatibility of major work, in all the distinct disciplines involved, on the related notions of knowledge/meaning, symbol, information, and memory. Two of the most important, and closely related, questions to be asked are : To what extent are these notions independent or reducible to each other? Is the human CNS a 'symbolic' system? A convincing case to such questions would go a very long way towards developing a genuine theory of (both human and machine) cognition.

Acknowledgements: Thanks to Alberto Greco for reviewing this paper.

Notes

* Some of the ideas in this paper were published before in Gelepithis (1993).

(1) It should be noted that encoding is meant to be seen as creating internal situation(s) and transformation(s) from the 'corresponding' external ones, and decoding is meant to be seen as creating an external situation from an internal one. In other words, 'encoding' and 'decoding' are assumed to be inverse functions of each other.

(2) This claimed consequence is how we see it, not Newell.

(3) The above two paragraphs have been taken from Gelepithis (1995).

(4) Under certain ontological assumptions condition (iv) becomes redundant.

(5) The relation between the two is extremely important but beyond the scope of this paper.

(6) According to standard definitions a model consists of: (i) a vocabulary specifying the set of allowable symbols; (ii) a syntactic subsystem specifying the arrangeability of symbols and the createability and modifiability of symbolic expressions; and (iii) a semantic subsystem specifying a way of associating meaning with the symbolic expressions.

References

Anderson, J.R. (1990), *Cognitive psychology and its implications*, 3rd ed. Freeman, San Francisco.

Brachman, R.J., Fikes R.E., and Levesque, H.J. (1983), Krypton: a functional approach to knowledge representation, *IEEE Computer*, **16**, 67–73.

Changeux, J–P. (1983*1985), *Neuronal man: The biology of mind*, Oxford Univ. Press (Originally published in France as *L' Homme Neuronal* by Fayard).

Doyle, J. and Patil, R.S. (1991), Two theses of knowledge representation: language restrictions, taxonomic classification, and the utility of representation services, *Artificial Intelligence*, **48**, 261–297.

Edelman, G.M. (1987), *Neural Darwinism*, Basic Books, New York.

Edelman, G.M., and Mountcastle, V. (Eds.) (1978), *The mindful brain: Cortical organisation and the group–selective theory of higher brain function*, MIT Press, Cambridge, Mass.

Gelepithis, P.A.M. (1991), The possibility of machine intelligence and the impossibility of human–machine communication, *Cybernetica*, 34(4), 255–268.

Gelepithis, P.A.M. (1993), *On the foundations of cognitive science: The nature of representation*, Working paper '93–1, School of Information Systems, Kingston University.

Gelepithis, P.A.M. (1995, in press), *Artificial Intelligence: An integrated approach*, McGraw–Hill–Europe.

Levesque, H.J., and Brachman, R.J. (1987), Expressiveness and tractability in knowledge representation and reasoning, *Computational Intelligence*, **3**, 78–93.

Newell, A. (1990), *Unified theories of cognition*, Harvard University Press.

Newell, A. (1992), Précis of unified theories of cognition, *Behavioral and Brain Sciences*, **15**, 425–437.

Rolls, E. (1987), Information representation, processing and storage in the brain: Analysis at the single neuron level, In: Changeux, J–P., and Konishi, M. (Eds.), *The neural and molecular bases of learning*, John Wiley.

Winston, P.H. (1992), *Artificial Intelligence*, 3rd ed, Addison–Wesley.

Manuscript received: 6–6–1995

Address author:

University of Kingston, School of Information Systems,
Kingston upon Thames, KT1 2EE, England.

(E-mail: petros@kingston.ac.uk).