Interactively converging on context-sensitive representations: A solution to the frame problem

(French, R. M. & Anselme, P. (1999). Interactively converging on context-sensitive representations: A solution to the frame problem. *Revue Internationale de Philosophie, 3,* 365-385.)

]

Robert M. French[1] and Patrick Anselme
Department of Psychology (B33)
University of Liège
4000 Liège, Belgium

Abstract

While we agree that the frame problem, as initially stated by McCarthy and Hayes (1969), is a problem that arises because of the use of representations, we do not accept the anti-representationalist position that the way around the problem is to eliminate representations. We believe that internal representations of the external world are a necessary, perhaps even a defining feature, of higher cognition. We explore the notion of dynamically created context-dependent representations that emerge from a continual interaction between working memory, external input, and long-term memory. We claim that only this kind of representation, necessary for higher cognitive abilities such as counterfactualization, will allow the combinatorial explosion inherent in the frame problem to be avoided

Introduction

You live in a tidy little suburb of America. You call up a friend and invite him to have a drink. He agrees to meet you at seven. But then, just before you hang up, he adds, "Unless my wife has scheduled something else for this evening." You think: that's reasonable. Then he adds, "Unless the bar doesn't exist anymore." You think: that's a bit strange since the bar was there three months before. Then he adds, "Unless I'm killed on the way there." You begin to wonder if you really want to have a drink with him. And finally, he adds, "Unless a meteorite destroys the earth" and you suddenly remember a prior engagement. But what, exactly, makes some of these conditions perfectly reasonable, others crazy? The philosophical problem is that we cannot exclude any of them a priori because contexts do exist in which they would be perfectly appropriate remarks. For example, if the bar in question was a theme bar devoted to punk rock music, or if the conversation had occurred in London in 1942 or in Sarajevo fifty years later, or if on that day the earth happened to be passing through a dense band of meteorites the size of Madagascar.

For humans this problem seems trivial. Under any normal circumstances, *of course* you don't need to mention the extraordinarily low-probability event that a meteor's encounter with the earth might prevent you from appearing somewhere for a drink with a friend. But for an autonomous robot how can this be done when:

- the event in question would be highly relevant under some circumstances, and therefore cannot be excluded a priori

[1] All correspondence should be addressed to Robert M. French, rfrench@ulg.ac.be. For related work, see: http://www.fapse.ulg.ac.be/Lab/Trav/rfrench.html.

- the robot will never have sufficient time to consider all possible events that might influence its course of action.

An initial attempt to sidestep this problem might be to reply, “Well, associate a probability of occurrence with each event that could possibly prevent you from meeting your friend at the bar. Then you order them in terms of their probabilities and only consider those events with probabilities greater than some fixed threshold.” Unfortunately this will not work for a number of reasons. First, there is no way you could enumerate all possible events that would prevent you from meeting your friend (e.g., “...unless all of the oxygen atoms in the atmosphere suddenly lose a proton and become nitrogen atoms”). And, even if you could, the probability of occurrence of an event depends on its context. (I have never checked my car for bombs before starting it. However, had I recently received a death-threat, I might well.) In other words, the probabilities of events are very *context-dependent.* And as a result, we cannot possibly associate each event with all the possible contexts in which it could occur and assign probabilities to all of them.

The conclusion is that if the robot must base its actions on a set of *context-independent* representations of the environment, it will fail, necessarily. In any real-world situation, there are just too many representations to be considered. And no clever algorithm or ingenious design will allow us to sidestep this problem. This is a slightly modified version of the frame problem, a problem that was first explicitly stated explicitly in an article by McCarthy & Hayes (1969). Their claim was that whenever a system uses representations to act (in simulated and real worlds) it was confronted with the frame problem, which can be defined as *the requirement of determining what information must remain unchanged at a representational level between two successive states of its environment after the system has produced a particular action.* McCarthy and Hayes were most concerned with what they referred to as ”side-effects,” in other words, effects of a particular action that were not directly anticipated by the rules used by the system. Since the original article, numerous authors have commented on this problem and, in some cases, presented solutions to it (see, in particular, Lars-Erik Janlert, 1987, 1996; Dreyfus & Dreyfus, 1987; Haugeland, 1987; etc.)

In an example very similar to one given by Dennett (1984), a robot is told that its power pack is in a room in which there is a bomb. It is instructed to retrieve its power pack. Its battery is on a small wagon in the room. So it uses its rule PULLOUT(BATTERY, ROOM) to retrieve the battery. It does this by pulling the battery out of the room. But, unfortunately, in pulling the battery, it also pulls along the wagon. As we might expect (unlike the robot, who hasn’t read many philosophy papers), the bomb in the room turns out to be on the wagon. Consequently, the side-effect of the robot’s retrieving the power pack is that it inadvertently pulls the bomb out of the room along with its battery. The bomb goes off, of course, destroying the robot. So, the robot is designed to deductively check the consequences of its action of removing the battery from the room. Some, like the fact that the wagon will come with the battery, are of no consequence. But it must also check that second-order actions, i.e., actions associated with the actions associated with pulling the battery out of the room, are also of no consequence. This is where it would discover that the bomb would come along with the wagon which came along with the battery. But what determines how far out along this web of possible deductions the robot must go and which of these deductions are relevant to removing its power pack and which are not ? Without some way guiding or focusing the robot’s deductive mechanisms, by the time it had deduced that the removal of the battery would leave unaffected the length of its wiring, the size of the wheels on the wagon, and the quantity and origin of varnish on floor, the bomb would have long since obliterated it.

The source of the problem: Representations

Is it possible to guide a robot through the shoals of the frame problem? We believe that the answer is yes. Exactly how this might be done requires starting with a discussion of what we view to be the source of the problem — namely, the problem of representation. Almost everyone agrees that the frame problem arises because of the use of representations. As a result, there seem to be three ways of dealing with representations when developing intelligent systems. These are:

i) deny the necessity of internal representations (anti-representationalism);
ii) use fixed, "context-independent" representations but somehow achieve the necessary speed-up in their processing (traditional AI and many connectionist models);
iii) build "context-dependent" representations on the fly.

In what follows, we will examine each of these views with respect to its possibility of handling the frame problem.

What is meant by "internal representation"?

The Artificial Intelligence literature is replete with discussions about internal representations, their structure, their origin, how best to manipulate them, their accuracy in reflecting the external world, etc. There seems to be one defining characteristic of representations that is crucial to the ensuing discussion — namely, they are not merely descriptive but causally efficacious. In the traditional view of internal representations may be independent of the presence or absence of any "real" environment. If, for example, the internal representations of a robot produced a "door-opening movement" (i.e., moving its hand upward to a certain position, pushing downward six inches and then pushing forward three feet), it would not be necessary for there to be a real door there to be opened. What we mean by causally efficacious is that the neural representations themselves play a causal role in corporal action. They are, in this sense, "active" representations. This contrasts with "passive" representations which might be likened to a television-monitor image. Assume the system had a television camera attached to it that merely recorded events in its surroundings and displayed them on a an internal monitor. The images on the monitor would "represent" the external world of the system in the sense that there would be a perfect correspondence between events in the world, but they would play no causal role in the corporal activities of the system. These "passive" representations (TV monitor images) could, however, become "active" (i.e., causally efficacious) were the system to observe them and act with respect to them.

This distinction is an important one and will play an important part of the discussion concerning the anti-representationalist solution to the frame problem.

Anti-representationalism: Denying the necessity of representations

The anti-representationalist point of view (Brooks, 1986, 1991; van Gelder, 1995, 1997; Maturana & Varela, 1980; etc.) considers representations, if it considers them at all, in the passive "TV monitor" manner. They would not, of course, deny that there are patterns of neural firings in the brain, but they would presumably claim that these patterns are causally irrelevant. These patterns, they would agree, reflect the characteristics of stimuli coming to the brain from the sensory apparatus but play no causal role in the state of the actuators of the system. But for them, the presence or absence of this type of passive representation is of no importance.

The major problem with this view is it overlooks the fact that *an organism's own patterns of neural activation are also stimuli*. In other words, internal representations are *part of* the environment in which the robot operates. The representations impinge on the robot's consciousness in much the same was as the sight of a tree, the smell of a rose or the sound of a gunshot. Just as the robot must interact with these phenomena in its environment, it must be able to treat its own internal state as input.

Another significant concern of anti-representationalists involves the nature of the representations of the environment that are provided as input to the machine. They feel that, while the these stimuli may be informative and meaningful to us, humans, the same is not true for the machines. This criticism extends not only the atomic nature of the symbolic representations of traditional AI, but also to the micro-featural decomposition of the inputs to connectionist networks (Rumelhart & McClelland, 1986), both of which are selected and organized in a priori manner by a human observer (Harvey, 1992). This remark is unquestionably correct but we claim that it is much less problematic than they would claim, especially in the context of connectionist networks. The sensory receptors from which the robot will gather its information about the environment are indeed chosen ahead of time by human designers. But one can argue that a human being is also born with fixed sensory detectors "designed" by evolution. In both cases the kind and amount of information supplied to the system is predetermined by the sensory equipment. On the other hand, if the input sensors are sufficiently sensitive, the precise manner in which incoming information will organize itself in creating an internal representation of a particular situation cannot be determined in an a priori manner. This is crucial to "unsupervised learning" algorithms that allow the robot a great deal of flexibility in developing its own internal representations of the environment (Thompson, 1997). Thus, to say that the robot is processing information is merely another way of saying that it is developing an understanding of its environment within the framework of a number of constraints imposed on it via its sensory mechanisms. Whether these constraints were artificially established by a human designer or naturally created by the process of evolution is irrelevant.

In addition, while the anti-representationalist position is perhaps adequate for ants or horseshoe crabs, it is far less certain that it is also appropriate for higher animals. Anti-representationalism seems to be tantamount to a strict behaviorist position, one which becomes very hard to defend when we address questions having to with real human cognition. Humans not only act based on their thoughts about their "internal" representations of a not-immediately-present environment — but this ability, perhaps more than anything else, is one of the main differences between humans and lower animals. A number of authors (Hofstadter, 1979; Dennett, 1991) have even gone so far as to argue that the origin of consciousness is precisely this ability to monitor and mentally interact with our own thoughts.

So, at least for human cognition, representations are necessary because they provide the stimuli to a great deal of thought and subsequent action. We therefore find the anti-representationalist position, at least in its strictest guise, to be unacceptable for modeling human cognitive capacities. The conclusions is therefore that internal representations, *active* internal representations that can act as stimuli in their own right, are necessary in modeling (or implementing) higher cognitive function.

So, we are obliged to conclude that the anti-representationalist position, as least insofar as it applies to higher cognitive activity, will not work. In short, representations — active representations — are necessary for human cognition. But if this is so, are we not once again faced with the frame problem? No, not necessarily. In what follows, we accept the necessity of representations. However, we will argue that the standard notion of *context-independent* representations does not make sense because it leads to intractable computational problems. We will then discuss how it might be possible to extract *context-dependent* representations by means of a technique that we will call "gradual convergence" and that requires a continual interaction between the cognitive system and the environment that will allow the on-the-fly construction of representations appropriate to the context in which it is found. Representations developed in this manner will, we claim, allow the system to overcome the frame problem.

Why the idea of context-independent representations does not make sense

The notion that the world could be represented by means of a vast set of symbols designating the objects and actions in the world and by a set of rules with which to manipulate those symbols goes back at least to the work of Frege and Russell (see Frege (1952) and Russell (1924)). This view has been called Objectivism by George Lakoff (Lakoff, 1987) who characterized it as follows: "On the objectivist view, reality comes complete with a unique correct, complete structure in terms of entities, properties, and relations." The application of this principle to the modeling of cognition bears a name: the Physical Symbol System Hypothesis (hereafter, PSSH; Newell & Simon, 1976). This view, one that served as the cornerstone of artificial intelligence for over two decades, posits that thinking occurs through the manipulation of representations composed of atomic symbolic primitives.

This idea of fixed representations for concepts has one overwhelming advantage: it allows a convenient division of labor in building AI programs — namely, some researchers can tackle the problem of representing the world, while others can work on processing the representations developed by the first group. In other words, representation specialists could design representation modules whose output would be fed to processing modules built by representation-processing specialists. It has all the trappings of a great idea. The only problem is that it can never work. One area in which the shortcomings of this approach are very apparent is analogy-making — the ability to see one object or situation in terms of another — in our attempt to demonstrate the impossibility of any scheme that relies on context-independent representations.

Making an analogy requires highlighting various different aspects of a situation, and the aspects that are highlighted are often not the most obvious features. The perception of a situation can change radically, depending on the analogy we are making. Chalmers, French & Hofstadter (1992) provide a good example of this in their discussion of DNA. Consider, they suggest, two different analogies involving DNA. The first is an analogy between DNA and a zipper. When we are presented with this analogy, the image of DNA that comes to mind is that of two strands of paired nucleotides (which can come apart like a zipper for the purposes of replication). The second analogy involves comparing DNA to the source code (i.e., non-executable high-level code) of a computer program. What comes to mind now is the fact that information in the DNA gets "compiled" (via processes of transcription and translation) into enzymes, which correspond to machine code (i.e., executable code). In the latter analogy, the perception of DNA is radically different -- it is represented essentially as an information-bearing entity, whose physical aspects, so important to the first analogy, are of virtually no consequence. But, although they stop at two, focusing on other aspects of DNA give a host of different analogies. For example, it is also like the central staircase of the famous 16th century Chambord Chateau in the French Loire valley. The staircase is a perfect double helix. In neither of the first two analogies was the double-helical form of DNA of much importance, whereas here it is the focal point of the analogy. It would seem that no single, rigid representation can capture what is going on in our heads. It is true that we probably have a single rich representation of DNA sitting passively in long-term memory. However, in the contexts of different analogical mappings, very different facets of this large representational structure are selected out as being relevant, by the pressures of the particular context. Irrespective of the *passive* content of the long-term representation of DNA, the *active* content that is processed at a given time is determined by a flexible representational process.

While this example might seem plausible, it might also seem to be so exotic as to be exceptional. Who, after all, goes around making comparisons with DNA in their day-to-day cognitive existence? But there is a deeper issue that applies not only to exotic analogies but to perfectly ordinary ones as well. What makes analogy-making a central feature of intelligence is that it is involved whenever the thought "That's like..." occurs to us, in other words, whenever we are perceiving one thing in terms of something else. New situations are understood in terms of previously encountered ones, emphasis is placed on particular aspects of one situation by likening it to another, and so on. And, again, this is unquestionably one of humans' most fundamental means of making sense of the world. Central to this ability to perceive the "sameness" in two different objects or situations is the problem of representation. Just as in the previous analogy with DNA, the goal of the exercise that follows is to attempt to demonstrate the necessity of extremely malleable, context-dependent representations. The only difference is that, in this case, we will not be dealing with an exotic, little-used concept like DNA, but rather that most ordinary of all objects, a credit card.

Just as in the case of DNA, whenever we make an analogy between the credit card and something else, we focus on certain features of the card and not others. So, for example, when we say, "A credit card is like money," we are focusing on its pecuniary aspect; the card, like money, can be used to purchase things. It is crucially important to observe how the representation of "credit card" must change with each statement in order to accommodate the analogy. The point is the *context-dependent nature* of representations. Our hope is to convince you that *no* a priori property list for "credit card," short of all of our life experience (i.e., the complete contents of long-term memory), could accommodate all possible analogical utterances of the form, "A credit card is like an X." Consider this short list of examples:

- "A credit card is a like a doorkey." In this case, we are no longer focusing on it's money-providing features — which, in fact, become completely irrelevant — but rather on its very thin shape, size, relative rigidity, and thickness.
- "A credit card is like a Braille book," Here, we are focusing on the raised letters on the front of the card.
- "A credit card is like a ruler." Because you can draw a straight line with it.
- "A credit card is like an autumn leaf." The focus here is on wind resistance. If you dropped both from the Empire State Building, they would have similar falling patterns (although the card would no doubt fall faster).
- "A credit card is like a breeze." Because you can cool yourself off with it if you use it as a little fan.
- "A credit card is like a soup-can label." Both contain encoded information that can be automatically read by a machine (in one case, from a magnetic strip; in the other, from a bar code).
- "A credit card is like fingernails." Both produce goosebumps in listeners who hear them scraped across a blackboard.
- "A credit card is like a bat." Because you'll never know what it's like to be either of them...

Perhaps it is becoming apparent that, with a little imagination, one can explain why a credit card is like absolutely *anything*. Even though your explanation (i.e., the context you create) may be stretched, it will be *understood*. Try it: A credit card is like a rose. A credit card is like a doormat. A credit card is like a horse race. A credit card is like a banana peel. A credit card is like a switch-blade knife. A credit card is like the Spanish Inquisition. The list is endless, but you will always be able to transfer some facet of your long-term memory representation of "credit card" — a representation that, ultimately, consists of everything in your life experience — to working memory in order to be able to say why a credit card is like some other object (French, 1997).

A final, extremely simple example will serve to conclude this discussion. Consider the most ordinary of utterances: "After the Christmas holidays my bathroom scale is my worst enemy," We all know exactly what this sentence means. But what a priori representations of "bathroom scale" and "worst enemy" could allow us to understand this simple expression? It would have to include knowledge about the tradition of big meals and excessive eating at Christmas, about people's concerns about being overweight, about irony, as well as subtle and complex knowledge about battles, enemies and competition in order to make sense of the idea of a hostile encounter between you and your bathroom scale. *All* of this information would have to be included in context-independent representations of "bathroom scale" and "worst enemy". With simple sentences like these (and many, many others), one begins to understand the necessity for context-dependent, process-interactive representations.

We hope that the above examples illustrate the essential impossibility of context-independent fixed representations for concepts. But if we conclude that the representation of any given concept must ultimately consist of the entire (passive) state of the brain, then how, exactly, are the parts necessary to handle a particular situation activated? If any complete representation must really involve the whole brain state, then the ability to extract context-appropriate sub-representations is really a *processing* issue, rather than a representational issue. But while this may shift the focus of the discussion, the problem does not go away.

A "gradual convergence" approach to representation

We hope to have shown in the previous sections some of the difficulties with the notion of a representation module that is separate from processing. Are we then obliged to process only "full" representations of every object or situation — a representation that would have to include virtually everything that we had ever stored in long-term memory — we encounter? This would not seem possible because of size limitations on working-memory (hereafter, WM), at least as this memory is normally construed (Miller, 1956; Atkinson & Shiffrin, 1968; Waugh & Norman, 1965; for a more recent review, see Baddeley, 1986). These limitations would not allow WM to accommodate such unwieldy representations. For this reason, long-term memory representations must be pruned in such a way that they can be used by working memory.

This would seem to strongly argue for an "gradual convergence" approach to representation. This approach has been developed, in particular, in the work of and Chalmers, French, & Hofstadter (1992), Hofstadter (1984), Hofstadter & Mitchell (1991), Mitchell (1993), and French (1995). Others have applied a similar philosophy to areas such as artificial life (Parisi, 1997). The following succinct explanation of this process of gradual representational convergence is from the Chalmers, French & Hofstadter (1992).

> Structures in working memory activate long-term memory items, activation then spreads from these items in long-term memory and activates other related items. Highly active long-term memory items will then be considered for participation in working memory. In this way, the activation in long-term memory influences the contents of working memory. When new structures are introduced into working memory, they may combine with structures already there, which would in turn send activation back to long-term memory, which would activate new long-term memory items, activation would radiate out from these items, and so on. In this way, *contextually appropriate* representations will gradually be built up in working memory.
>
> In this way, the representations in working memory do not have to include every bit of information that could possibly be associated with a particular situation. They include only contextually relevant information, this being determined in large measure by the concept activation levels in long-term memory. It is also the fact that representation-building is largely dependent on concept-activation levels in long-term memory which keeps the process of representing from becoming combinatorially explosive.

“Pengi”: an early attempt at context-sensitive representing

This notion of a permanent interaction between working memory and long-term memory is essential in understanding the incremental creation in working memory of context-dependent representations. This interactive mechanism allows the system to make use not only of “external” environmental stimuli but also to incorporate information previously stored by the system in long-term memory. In short, we suggest that this mechanism is *necessary* to solve the frame problem for high-level cognition in a psychologically plausible manner.

Agre and Chapman (1987) developed a simulation, Pengi, in which an agent — a penguin — makes use of context-dependent representations in order to avoid being attacked, in this case, by a bee. Piles of ice cubes are used to allow Pengi to protect himself from being stung by the bee. In addition, Pengi can fight back against the bee by means of a well-placed kick to any ice cube that the bee happens to be directly behind. As for the bee, it has two means to kill the penguin: either by striking an ice cube that Pengi is hiding behind or by stinging the penguin. Space does not allow a full development here of the strategies that the penguin can adopt, nor of his actual sensori-motor abilities (for a detailed description, see Agre, 1997). It is, however, important to note that the penguin has well-developed attentional capabilities that allow him to focus on the salient factors of the situation in which he finds himself.

In the Pengi simulation, all events are contextually determined. There are not context-independent representations, like “ice-cube wall” or even “bee.” The penguin’s representations are, to use Agre’s term, *deictic*, meaning that they depend on the circumstances in which they are used. Pengi’s representations take the form of the-wall-behind-which-I-will-hide or the-bee-I-want-to-kill (Agre, 1997, p. 267). These representations do not describe context-independent entities; rather they describe some aspect of the environment that is in a particular relation to the agent at a particular moment in time. In other words, rather than a context-independent “bee” representation, the system produces a representation for a particular bee in a particular place at a particular time. The Pengi model avoids the frame problem because the agent is manipulating context-sensitive representations for “bee” that include various salient aspects of situation at hand. However, the overly restrictive notion of context in the Pengi model leaves the question open as to whether this type of architecture could scale up to domains in which broader notions of context applied.

The agent is limited to representing situations in its environment but — and this difference is of crucial importance for cognitive modeling — cannot include in its representations the experience of past learning. In other words, the agent’s ability to contextualize its representations is limited to “external” contextualization. So, for example, the bee might be contextualized by Pengi as moving quickly, being in a certain place, acting in a certain way, but it cannot contextualized the bee with respect to previously learned facts, such as bees are rather like flies, therefore it might be worth applying certain previously acquired fly-avoidance techniques to bee-avoidance. Furthermore, in Pengi information is not processed and stored. Thus, there is not possibility of the program to learn causal relationships. In short, Pengi operates exclusively with respect to the present and has no ability to reason, even in an elementary fashion, about causality (Toth, 1995). We believe that, even though this type of program is on the right track, the constraints on Pengi’s operation and simplicity of its micro-domain are such that, ultimately, it is able to avoid the frame problem more because of the elementary nature of its domain rather than the sophistication of its mechanisms. In Pengi’s environment all factors are external to the agent and all have the same salience. However, if we are considering an agent capable of real high-level cognition, this assumption of equal salience will lead back to the frame problem. An agent in a complex environment *cannot* attach all possible aspects to its representation of a particular concept that must be dealt with. It must have some means of prioritizing concepts with respect to the context at hand: Some are very important, others are less so, and still others can be ignored altogether.

The problem is that the simplicity of the Pengi's environment and its inability to incorporate in its current representations previously stored information of past events and experiences puts Pengi on a par with non-representational systems that can only do low-level (i.e., immediate stimulus-response) cognition. If, as we believe we have established, internal representations of past mental acquisitions are necessary to model cognition, then would the new version of Pengi, operating in a fuller environment that would also include its own internal mental representations in its representation-building, still be unaffected by the frame problem? In other words, would the New Pengi still be able to act based on its immediate sensory perceptions of the external world *and* its previously stored knowledge?

While Pengi is perhaps the problem most closely related to work in the area of robots and animats, other programs have been developed in an attempt to find a way around the frame problem. The most extensive attempts to model context-sensitive representations have been attempted by a number of researchers who, over the years, have developed models based on an architecture first outlined by Hofstadter (1984). This early work led to a number of computer programs working in a variety of micro-domains (Mitchell, 1993; French, 1995; Defays, 1995; McGraw, 1996). In these programs, WM and LTM are presented as distinct, although continually interacting, memory structures. There is certainly a need to integrate these two memory structures in a more direct way. One attempt along these lines has been produced by Kokinov (1994).

One of the serious drawbacks of these programs is their inability to learn over the course a series of interactions with their environment. The long-term memory structures in these programs are hand-coded. However, this is not acceptable for agents that are to become truly autonomous. They must be designed so as to be able to:

- Use their long-term memory structures in combination with external input from their surrounding environment in order to develop working-memory representations needed to act effectively in the world.
- Allow these working memory structures to contribute to the content and structure of long-term memory.

In addition, a satisfactory implementation of "long-term memory" and "working memory" must be found. Is working memory comprised of the activated structures of long-term memory, as some authors have suggested? (Shastri, & Ajjanagadde, 1993; Sougné, 1996; Sougné & French, 1997) Or is it rather physically separate storage? (Baddeley, 1986) Or something else? For the moment, these are open questions, but ones that must ultimately be resolved.

Converging representations and the frame problem

We have now suggested that i) representations are necessary to high-level cognition, but ii) not just any kind of representations, context-dependent representations, that can only be produced iii) by a continual interaction between the environment and long-term memory. We will go on to suggest that this kind of representation is the kind that is needed to overcome the frame problem.

First, in reading much of the literature on the frame problem, one has the impression that this problem is uniquely a problem encountered by robots. Strictly speaking, this is not accurate. Consider the following incident. A six-year old girl came to dinner with her parents at the home of one of the authors. She was unfamiliar with a gas stove, since in her family they use an electric range. When everyone was about to sit down to the table, the food was taken off the stove and moved to the table. One of the small burners on the stove remained on and the little girl, knowing we were about to eat, decided to help by extinguishing the little flame on the stove burner. The only household flames she was familiar with to that point were candle flames and she know how to put those out. So, she blew very hard on the small flame on the stove, coming very close to extinguishing it as she had intended. We had to explain to her that "gas stove flames" and "candle flames" were not the same and that if you blew out the former, gas would continue to come out of the stove and possibly cause an explosion.

Here, then, is an example of a "side effect" — blowing up the house — completely unanticipated by the simple act of blowing out a flame. The little girl could have had no reasonable way of knowing that blowing out a candle flame was the proper way to do things, but blowing out the gas flame on a stove was extremely dangerous. It would clearly be unreasonable to suggest that she should have "thought about the consequences of her actions" (i.e., in robotic terms, deduced the consequences of actions) because there would be too much to check. How close do two flames have to be before you can say one is the "same" as the other and act with the new instance as you did with the old? If no action can be taken until every possibility of every new instance of every concept is checked for safety, then this will lead to paralysis — both in humans and in robots.

But if we adopt the "gradual convergence" approach to representation-building, we can see how this could lead to the selection of representation structures that are context-dependent and that will serve as the basis of subsequent action. In other words, the robot in Dennett's example will not have to check the color of the wallpaper, unless, for some reason, wallpaper is a highly active concept. That may lead to a particular internal representation that may well modify the next things checked by the robot, which may then cause some other feature of the environment to become active in robot's internal representation, which again will influence the robot's actions, and so on. In other words, a continual interaction is necessary with the environment in order to gradually converge on representations that are appropriate to overcoming the frame problem.

The price of this is, as in the case of the young child, that sometimes the robot will blow out a gas flame and cause an explosion, but *in general*, this interactive method of construction of representations will allow us to overcome the frame problem.

Acknowledgments

This work has been supported in part by grant PAI P4/19 from the Belgian government.

References

Agre, P.E. & Chapman, D. (1987). Pengi : An implementation of a theory of activity. *Proceedings of the Sixth National Conference on Artificial Intelligence*, Seattle, 1987 : 196-201.

Agre, P.E. (1997). *Computation and Human Experience.* Cambridge University Press.

Atkinson, R. & Shiffrin, R. (1968). Human Memory: A proposed system and its control processes. In K. Spense and J. Spense (eds.) *The Psychology of Learning and Motivation, Vol. 2*. New York, NY: Academic Press.

Baddeley, A. (1986). *Working Memory*. Oxford: Oxford University Press.

Brooks, R. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation, 2,* 14-23.

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

Chalmers, D. J., French, R. M. and Hofstadter, D. R. (1992). High-level Perception, Representation, and Analogy: A Critique of Artificial Intelligence Methodology. *Journal of Experimental and Theoretical and Artificial Intelligence,* 4(3), 185-211.

Defays, D. (1995). A Study in Cognition and Recognition. In D. Hofstadter & FARG *Fluid Concepts and Creative Analogies*: *Computer models of the fundamental mechanisms of thought*. New York, NY: Basic Books. 131-155.

Dennett, D. (1984). Cognitive wheels: the frame problem of AI. In *Minds, Machines and Evolution*. C. Hookway (ed.). Cambridge, UK: Cambridge University Press.

Dennett, D. (1991). *Consciousness Explained*. NY: Little, Brown.

Dreyfus, H. & Dreyfus, S. (1987). How to stop worrying about the frame problem even though it's computationally insoluble. In Z. Pylyshyn (ed.) *The Robot's Dilemma: The frame problem in artificial intelligence*. NY: Ablex.

Frege, G. (1952). *Translations from the philosophical writings of Gottlob Frege*. P. Geach and M. Black (trans. and eds.) Oxford: Basil Blackwell.

French, R. (1995). *The Subtlety of Sameness: A Theory and Computer Model of Analogy-Making*. Cambridge, MA: The MIT Press.

French, R. M. (1997). When coffee cups are like old elephants or Why representation modules don't make sense, *Proceedings of the International Conference New Trends in Cognitive Science*, A. Riegler & M. Peschl (eds.), Austrian Society for Cognitive Science, p. 158-163.

Harvey, I. (1992). Untimed and misrepresented : Connectionism and the computer metaphor. CSRP 245, University of Sussex.

Haugeland, J. (1987). An overview of the frame problem. In Z. Pylyshyn (ed.) *The Robot's Dilemma: The frame problem in artificial intelligence*. NY: Ablex.

Hofstadter, D. & Mitchell, M. (1991). The Copycat Project: A model of mental fluidity and analogy-making. In K. Holyoak and J. Barden (eds.) *Advances in Connectionist and Neural Computation Theory, Vol. 2: Analogical Connections*. NJ: Ablex.

Hofstadter, D. (1979). *Gödel, Escher, Bach: an Eternal Golden Braid.* NY: Basic Books.

Hofstadter, D. (1984). The Copycat Project: An experiment in nondeterminism and creative analogies. MIT AI Memo No. 755.

Janlert, Lars-Erik (1987). Modeling Change — The Frame Problem. In Z. Pylyshyn (ed.) *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*. NJ: Ablex.

Janlert, Lars-Erik (1996). The Frame Problem: Freedom or stability? With pictures we can have both. In K. Ford and Z. Pylyshyn (eds.) *The Robot's Dilemma Revisited: The Frame Problem in Artificial Intelligence*. NJ: Ablex.

Kokinov, B. (1994). The context-sensitive cognitive architecture DUAL. In *Proceedings of the 16th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things*. Chicago: University of Chicago Press.

Maturana, H. & Varela, F.J. (1980). *Autopoïesis and Cognition : The Realization of the Living.* Boston Studies in the Philosophy of Science, t. XLII, Boston, D. Reidel.

McCarthy & Hayes, P. (1969). Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer & D. Michie (eds.) *Machine Intelligence 4*. Edinburgh, Scotland: Edinburgh University Press.

McGraw, G. (1996). Letter Spirit (part one): Emergent High-level Perception of Letters Using Fluid Concepts. Unpublished Doctoral Dissertation. Indiana University.

Miller, G. (1956). The Magic Number 7, ±2: Some limits on our capacity for processing information. *Psychological Review.* 63:81-93.

Mitchell, M. (1993). *Analogy-making as Perception*. Cambridge, MA: The MIT Press.

Newell, A. & Simon, H. (1976) . Computer science as empirical inquiry: Symbols and search. *Communications of the Association for Computing Machinery*. 19: 113-126.

Parisi, D. (1997). Artificial life and higher level cognition. *Brain and Cognition, 34*, 160-184.

Rumelhart, D. & McClelland, J. eds. (1986) *Parallel distributed processing.* Cambridge, MA: MIT Press, Bradford.

Russell, B. (1924). Logical Atomism. In *Logic and Knowledge* (1956), R. Marsh (ed.) London: Allen and Unwin.

Shastri, L. & Ajjanagadde, V. (1993). From Simple Associations to Systematic Reasoning: A connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences, 16,* 417-494.

Sougné, J. (1996). A Connectionist Model of Reflective Reasoning Using Temporal Properties of Node Firing. *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*. Mahwah, NJ:LEA.

Sougné, J. & French, R. (1997). A Neurobiologically Inspired Model of Working Memory Based on Neuronal Synchrony and Rythmicity. In J. Bullinaria, G. Houghton, D. Glasspool (eds.). *Connectionist Representations: Proceedings of the Fourth Neural Computation and Psychology Workshop.* Springer-Verlag. 155-167.

Thompson, E. (1997). Symbol grounding : A bridge from artificial life to artificial intelligence. *Brain and Cognition, 34 :* 48-71.

Toth, J.A. (1995). Book review. Kenneth M. and Patrick J. Hayes, eds., *Reasoning Agents in a Dynamic World : The Frame Problem. Artificial Intelligence, 73 :* 323-369.

van Gelder, T. (1995). What might cognition be, if not computation ? *The Journal of Philosophy, XCI, 7 :* 345-381.

van Gelder, T. (1997). The dynamical hypothesis in cognitive science. *Behavioral and Brain Science.* (In press).

Waugh, N. & Norman, D. (1965). Primary Memory. *Psychological Review* 72: 89-104.