Weyhrauch, "Building Cs. Artifacts"

CONSCIOUSNESS:
Distinction and Reflection

Edited by
Giuseppe Trautteur

BIBLIOPOLIS

ISBN 88-7088-341-8

Printed in Italy by Arte Tipografica s.a.s., Napoli

Table of Contents

Building Conscious Artifacts

Richard Weyhrauch

1. Introduction

I was looking forward to this Workshop because of the wide range of people attending it. I am generally thought of as being on the farthest theoretical edge of representation theory, but here I would like to address, as seriously as we can, the idea that we should try to actually build an artifact that has human level – at least human level – survival capabilities and performance capabilities. In fact, I am not much interested just in trying to understand what is, for example, consciousness or cognition, but I am really interested in discovering what kind of object, in the engineering sense, I need to make if I am going to succeed in having artifacts that are capable of substantive enough interactions with us that I would be interested in inviting them along as a dinner companion, who could be as interesting as the biological fellow travelers, like the people in this audience, that I spend time with.

For me the issue is what kind of artifact can come to know his world in the way that we know it. That is, has enough common ground with us so that we have a mutual understanding of the world, that we can enjoy meaningful conversations with him and that we have realistic hope that he will contribute to the richness of our lives in the way that our colleagues and friends do.

If our level of expectations is less than this we are unlikely to be satisfied with the results, and probably cannot even hope to solve real problems. Building artificial minds is not an activity that will mature out of partial solutions – wanting everything is a reasonable level of expectation as a starting point for this Workshop. And so, I guess that the core of what I hope we talk about and the direction I hope the discussion takes is not just abstract theorizing, but rather, that we judge ideas on the basis of whether they produce practical blueprints for building artifacts with a real mental life, who would be

both knowledgeable and conscious. Anything short of this begs the issue.

2. Where the Workshop stood

I did not know how my talk would come out, because I have two different points of view and they lie at such radical ends of the spectrum that I was afraid I would end up wobbling back and forth. I had decided that I would not give a detailed technical talk, although the FOL work (Weyhrauch 1980) does provide what I think is an interesting cognitive architecture. What I proposed was something else.

One of the reasons that I suggested that we all get up and give an introductory talk on the first morning was that I was very much interested in how we could fit together and make use of the fact that the groups of people at this Workshop was incredibly diverse, with very different ideas and backgrounds. One of the things that I, for myself, wanted to get out of this Workshop was some understanding of what other people are doing and how their different subject areas could contribute. I did not think that the extreme technical details of the FOL system that I usually present would add much at that point. Instead, in my talk I encouraged everybody to participate actively as a way of prodding us into confronting what we are doing here, and to look at things from as many different directions as possible.

Since I took our task to be the project that I described rather crudely in the previous section – to ask if we want to have artifacts that are "just like us", things that have a good understanding of the world and could act as colleagues, companions, or aides and make important contributions to our society in the same way we hope that we do and because we were such a diverse group of people from so many different disciplines, I started my talk with a survey of the attendees so that everyone could get a sense of the intellectual landscape covered here. I had people raise their hands to a set of questions and since the answers were sufficiently interesting, I thought I would summarize them here. After all, we were here to discuss consciousness, and I was interested in the simple question of who thought what existing things were in fact conscious.

After we surveyed the research areas of the attendees

(psychology, computer science, physics, cognitive science, neuro-physiology, biology, ...) I continued by asking some questions about the subject of the Workshop – namely consciousness. I started by asking about spoons. No one thought that spoons were conscious. Okay, What about wrist watches? They move and they tell us the time. What about plants, you know, corn or roses? No one was willing to take the position that roses were conscious. What about animals? Here a few people thought that animals were conscious, but people suddenly wanted to be more selective, they asked what kind of animals and what kind of consciousness! So we started at the bottom; what about flat worms? No one was willing to vote for consciousness in lower life animals like these. What about chickens? What about chimpanzees and gorillas? And a favorite topic of Americans: what about whales or dolphins? What about people? Although the responses were on the rise, I was surprised to see that not everyone responded, "yes", to the question about people. Then I asked the question in the negative: is there anyone who does not think that people are conscious? and surprisingly there were some. When asked to explain, it seemed that people did not think that the answer was so straightforward. Some thought that people were only conscious some of the time, and others asked whether consciousness in people is under their conscious control.

Well, this was not a scientific survey, but it revealed much about just how differently people thought about consciousness. "Is there anyone here who thinks that there are no conscious animals and that the only known artifacts that are conscious are humans?" One, two... At this point someone mentioned self-consciousness! Who or what was self-conscious? We repeated the survey about self-consciousness. This revived the discussion about being unconscious and I think that it is almost unanimous that sometimes people are unconscious. Next we asked whether organizations and some form of societies were conscious. What about ant colonies? Does anyone think that they are conscious? What about self-consciousness for ant colonies?

Although fairly sophomoric, given the make up of the Workshop, this questioning was actually more interesting than I thought it was going to be. The diversity of opinion is quite surprising and it was quite wide. By the time we had gotten

to chimpanzees more than half of the group thought that there was consciousness involved. With the caveats about humans sometimes being unconscious – virtually everyone thought that humans were both conscious and self-conscious –, there was a slightly smaller number of people who thought that gorillas were self-conscious, and a fair number of people thought that societies were conscious.

Then Trautteur asked if there were degrees of consciousness or whether you are just conscious or not. After much discussion we finally asked: do you think that consciousness is primarily a mechanism of gradual differentiation of some form, where there is some form of full consciousness at one end and non-consciousness at the other? About one third voted for all or nothing.

Then someone asked if "consciousness" is culture-relative. Do we imagine that, at any particular time in history, consciousness is culturally determined, or is it, so to speak, developed through time? Is it geographical? Do we think that Italian consciousness is different from American consciousness?

What about the development of self-consciousness, is it catastrophic or developmental? Is consciousness, to some degree, related to language?

After some discussion, Phil Johnson-Laird, brought up to the surface the idea of hierarchies of consciousness by remarking, "That is to say that if you really practice hard you become aware that you are aware that you are self-aware". Someone then suggested we vote on whether people thought that this talk about awareness and self-awareness and awareness of self-awareness, and however many iterations of that kind of notion you want, requires that a cognitive architecture must be a system that is based on some kind of recursive or hierarchical structures. Unfortunately neither the discussion nor the results of the vote were recorded.

3. Where I stand

Having set the context with the above survey, let me now try to explain a little bit more about what my own efforts are about. Most of my academic life has been spent developing and describing and theorizing about what in Artificial Intelligence was called representation theory. Now

it is more common to talk about such things like cognitive architectures or some other such phrase. I have built one major computer system, FOL, a representation system based on first order logic. This puts me clearly at the logical based system end in the spectrum, so one thing I want to discuss is why I think logic is relevant to the activity of building minds. A lot of people think, for example, that formal logic has nothing to do with cognition, nothing to do with problem-solving or psychology or philosophy, but that it is simply a formal trick for entertaining the theoretician.

But in my entire experience of working with logic, my motivations have always been driven by the desire that I expressed at the very beginning which I take very seriously: the desire to build an artifact that is aware of the world – let us leave the boundaries of awareness and consciousness aside for a moment – who has the ability to have conversations with us, and with whom, in Maturana's words, we share a domain of consensus that is sufficiently rich so that it supports real communication.

All of my research is directed in the service of that aim. And of course, it is also a way of understanding ourselves because we are, among other things, the primary example we have of such a cognitive artifact. But it is not just an exercise to duplicate ourselves that I think is interesting. I wonder what it would be like to know such robots and to have this kind of associate. Without noticing it we have become used to this idea in our everyday life. We have already entered a world where children grow up in a world where the idea of highly cognitive robots or maybe extra-terrestrial intelligence are not surprising or shocking or even strange; we have all seen Star Wars and identify with R2D2 and C3PO (and more recently Commander Data of Star Trek, the next Generation). We no longer think about those things as strange, in fact we are more likely to imagine them as the colleagues and friends of the biological people in those shows. In our own mental life we do not think about them as being very different from us.

Of course, in Star Wars one of them, 3PO, gets mangled and is, in a certain sense, brought back to life, which is not a property that we easily attribute to human beings, but the question that interests me is, cognitively, what kind of stuff is in such an object. What is exciting is that we can imagine

what it would be like to have a performance-competent robot, not of the industrial kind used for building cars, but one for traveling around with us through life, like friends do.

I want to be clear that I have confidence that it is possible for us to build such things and that it should be a scientific goal to do it. My confidence does not necessarily depend on duplicating brains. Built minds will most likely be different. On the other hand, it is important to study brains and to study ourselves to understand the distinctions we have worried about above. Why do we think that some things are conscious and some things are unconscious? Can a non-biological thing ever be conscious? Why do we think that some things do learn, but that others cannot? We could ask the same range of questions we asked above about learning and about problem-solving and many of the cognitive activities that we engage in.

I find it unlikely that the electromechanical solutions to these problems will be very like, in their atomic details, the biological solutions that we are. And so, one of the tasks when asking the questions of this Workshop, and one of the nice things about having so many disciplines here, is to try to understand what level of detail is appropriate for this discussion, and what a theory of conscious artifacts might be like. What level of description should we try? On one end, if we consider pure computer science and we want to put a cognitive architecture into the memory of a computer, then we are talking about data structure design; on the other end of the spectrum we could have a completely abstract conversation about ontology in its purest form, about whether consciousness exists. In the middle we could look at how the different descriptions, of all of the disciplines represented here, can be merged to the service of my quest to actually construct conscious artifacts.

3.1 The seeds of an approach

Let us start to give a little more detailed view of the task. What kinds of things do we have to account for? One entire area that I find missing from representation theory (in addition to problem-solving competence, which is bad enough itself) is that, in the computer science world at least, hardly anyone notices that dealing with perception is a vital

component of any theory of representation that could be used to build a robot – let us call it an individual. One of the questions we are going to have to answer is how does this individual come to know what world is in and how can he come to know it?

Now we have already stepped into a problematic hole. Once we use the word "individual", there are classical philosophical questions about what we mean by being an individual at all. A simpler question is: how do we know where our boundaries (as individuals) are? Trying to give an explanation about what an individual is raises serious questions about boundaries. This reflects itself in our entire philosophical history. How can we understand what an individual is.

The computer program I built in the late seventies is called FOL. It was based on First Order Logic, and let me start off in general and then let me say something about what logic has to do with any of this.

Let us start with the presumption that we are presented with an individual. Now I ask (before we address anything of a level as high as consciousness) how do we know that we have an individual? How do we know that we have a coherent ongoing entity and what do we mean by that? How do we know that he has a reasonable understanding of the world, and the old philosophical question of how do we know that, as time proceeds on, we are dealing with the same individual or not.

How does he preserve a certain amount of consistency? Given that we had a reasonable individual before, in an effort to understand how its environment affects it, how do we know that we end up with a reasonable individual later? What does this passage through time consist of?

Passage through time consists of sensory input, of internal mental activity, of a complicated collection of activities: it walks, it talks at the same time, some of it may be conscious, some of it may be unconscious. Is there some way of giving a description of what it means to be an individual, where the boundaries of this individual are and what does it mean for him to preserve a reasonable understanding of his world through time?

If this individual that we are going to build is to have

survival capabilities, we are immediately presented with the problems of philosophy, cognitive science and artificial intelligence in an attempt to understand what coherence-preserving transformations on an individual might be.

You can see here that I am trying to outline a strategy for how I believe that a theory of the coherent persistence of individuals through time must be formulated. What I want to do is to explore the idea that a reasonable way to look at our task is to explain how physical interactions through time change an individual.

Let us consider how this strategy might affect linguistics. In my view, the meaning of a sentence is not something that can be determined independently of the hearer. That is, for me, the meaning of a sentence is the effect that its being heard has on the hearer. This interpretation precludes the idea that meanings can be represented as structures that are independent of and outside the individual who is doing the hearing.

Fernando Flores once commented to me that we should not use the word "data" with respect to the acquisition of knowledge, and that the appropriate word is "capta", that is, what is important is what we captured (heard), not what was given (spoken) – (cf. the Latin roots). I fully agree with that; a consequence is that we need to understand what kind of structure our hearer is, i.e. what is the architecture of that individual, and when he hears something, how does he change and what does he become.

This is the kind of theory that I want to propose. For me, the question of our consciousness, the question of our cognition, the question of our ability to learn, the question of our understanding are all questions about what kinds of changes can happen in individuals through time as the result of being and as the result of being in the place that they are.

But part of being "in the place that you are" is how does this individual, given its boundary, connect to the place that it is. That is, if we build this individual, how do we arrange it that he is not simply being manipulated by external forces, but that he is in the world? (has being?) How can it come to have a life of its own?

We have not asked questions about whether being conscious or self-conscious has anything to do with a notion

of being alive. How is it that we distinguish ourselves as not being simply a part of the mechanical linkage of universe, but actually as "the thing we are" in "the place we are"? And how is it that we come to know the place we are at? If our theories of what cognitive architectures do not explain how we connect to our exterior and do not explain how it is that we come to know what is outside us, I do not think that we have a very successful theory.

My point of view is that for everything we know the world must be contained inside our own structure, and any differentiation that we can make about the world is an architectural/structural property of us. That is, if one day I believe something and another day I believe something else, then this change must be only detectable by examining me. It cannot be embodied in some part of my exterior.

From this perspective, what we need to be looking for in a theory of individuals, is an explanation for the changes in my physical state, my emotional state, my internal state or whatever state that is me. This is the kind of theory that I want. I want a theory that specifies an architecture so I can look inside and tell what actual structure is there and, if it hears something, for example, how does that structure change. (To be clear this does not imply a homunculus, or that a look inside will find some sentence that I believe.)

This is why we need to think "capta". If you do not believe that it is capta that's important, examine your own behavior – frequently we lie to our friends because we know how to interpret what they hear and not what we literally say. Very frequently we will swear that we heard some remark, but be told: no, that's not what I said. And so, in some real sense, there is no information going back and forth. There is only what you catch out of the flow, on the basis of what you hear. That is, how you change as a result of having heard. I am trying here to give a vision of the kind of theory, that I think one needs in order to explain the autonomous self (including the topic of this Workshop, consciousness).

Question: It seems that you are pushing the capta point of view too far. On the whole, people can succeed in communicating by using language. Therefore, when I say something to you, it must be something that my knowledge of the language gives me, that must remain relatively invariant.

I agree that most utterances are dependent upon their context for their interpretation, and always the context includes the hearer's state of mind, but I think, I would argue that there is a bit of data and a bit of capta. If you put everything in what effect the language has on you, you overlook the possibility that a person can genuinely misunderstand something.

Answer: I certainly believe that people communicate, and I certainly believe that with some relatively large probability a substantial part of your understanding of the world is like my understanding of the world. The question we need to look at is whether or not this common understanding is an integral part of our structural architecture or whether it is embedded in the content (software) of what we have come to know, either evolutionarily – i.e. in a frozen, pre-fabricated way –, or through our personal experience. I would argue that, in fact, you cannot assume the importance of any "data" and successfully build cognitive artifacts. The reason is that we would then have to actually agree, among all of us, that something was irreversibly correct.

Even if we could all agree that something is correct (true?), it would be the idea that this "truth" was part of the architecture of the universe, probably would be beyond what we would be willing to agree with. Let me present an example: consider the most notorious motorcycle gang you can imagine. They are tough and in some ways their notion of fairness and ethics are unknown to you, and because of your cultural deprivation you have not enough experience of the world to predict their behavior. Furthermore, suppose you have heard that they enjoy taking professors and people who hang out around universities and, just for fun, terrorizing them and maybe even beating them up. They are sufficiently different from you that you do not have any idea of what they might do. Now, picture yourself in a bar, minding your own business, when the door of the bar opens and twenty members of this gang, all in black leather jackets, looking really mean, burst into the bar. The leader walks up to you and says: "O.K., prof, what is 2 plus 2?" Your first tendency would be to say "four", but you hesitate, wondering what he really meant. But why? Yesterday this was an analytic truth! So here's the problem. I claim that if we can imagine that in the above circumstances you hesitated for even a microsecond,

because you are wondering what he meant by that, then you have confirmed my point. If data were the issue, since we are talking about a totally analytic notion, it would be safe to respond without thinking. But no, you know in your heart that you had better "catch" what he wanted you to answer before you actually commit yourself to something that you, in your intellectual life, might believe is an analytic truth.

And so, for me, the architectural question about natural language is how do we replace a data-based theory of language meaning with a capta based one. This, even though there is a substantial amount of general or consensual knowledge about our world. A corollary of this view is that indirect speech acts are the default and direct speech acts simply artifacts of convenience. The idea that meaning is, in some sense, contained in the words, rather than carried by words, is unlikely.

Let me give you another example. I believe if you go up to the vaporetto* ticket sales-agent and you hand him some money and say: "Please, give me a cup of tea", he will still give you a ticket and the correct change. Now, why is that? The answer is not that you were incoherent, but rather that he knew exactly what "Please, give me a cup of tea" meant; it meant I want to take a ride on the boat.

It is unrealistic to believe that you can extract that "meaning" by looking at the words. At best, if you ever expect "understand" that sentence, you have to look at the domain of consensus in which you are behaving, in order to catch the meaning. I believe that direct speech acts are in fact technical inventions that are only necessary when the domain of consensus is weak and we use them to clarify situations where we could not catch any meaning without more formal agreement.

A comment and a question: The comment is about this contextual dependence you mentioned and the fact that the interpretation of language may be a function of the particularly located specifics. It seems true to me and that is not something that logics deal with. Part of what you are mentioning, is that systematic regularities exist, and

* The "vaporetto" is the small boat which substitutes buses in Venice, where the Workshop was being held. [N. of the Ed.]

that even though meaning seems to be context-dependent, it is systematically context-dependent. You say this without challenging something else which you may be suggesting, that is, that the specificity of human actions transcends regularity, and in particular cases what is actually happening in the micro-details of a given human interaction. There is more to it than any systematic account can give. Specificity outstrips systematicity. That, I think, is the real challenge of context-dependence.

The question. When one considers meaning as the effect of communication on the hearer, one issue is whether we want to treat the phenomena we are exploring – consciousness, meaning, language and so on – as physically contained inside heads (or computers) or whether we should treat them, in some sense, as relational to the wider world.

The challenge that I would pose to you is that, even though I agree that there is something very important in what happens to people and so on and so forth, the challenge is that by reducing meaning to what goes inside the hearer, you are implicitly advocating a kind of physical reductionism.

A.: The fear of physical reductionism is grounded in people who believe that reasoning, and even truth, depends upon externals from the individual that holds them. Even if there is a unique actuality, which we all knew about and correctly believed in, it does not commit us to an architectural position that commits other beliefs to non-existence. So physical reductionism is independent of "truth" and of our connection to it.

I believe there are many traditionally accepted theories of external truth and belief that are trivially unusable as problem-solving tools for our computer individual, but have advocates in the literature. The position I prefer is not that anyone has privileged knowledge about what is, but rather that we share an infinite historical and evolutionary tradition that causes us to have similar ideas about things. And it is this tradition that holds together our ability to communicate and any attempt to imagine that there are, in a sense, objective externals will keep us from building successful computer individuals.

Q.: I think your ideas are clearly false. I would like to stress that I think it is ridiculous to say that meaning is in what

the hearer gets from communication. I think that you should take into account that the one who is speaking the sentence is trying to inform somebody else and by this act of informing he is trying to bring the hearer a particular idea, an idea in which both can find some sort of consensus. So there is an action on the part of the speaker trying to inform and there is an action on the part of the hearer that is trying to get message. And you cannot simply disregard that.

A.: I do not imagine that the intentions of the speaker are irrelevant to a large-scale understanding of a conversation, but I propose that the hearer's idea of the intentions of the speaker are in fact relevant to the hearer in terms of how the capta gets processed; that is, only through the hearer's understanding of the intention of the speaker. What the speaker's actual intentions are, what your intentions are when speaking to me, when speaking to this group, that is something I have no direct perception of.

I might believe that you want to inform me, so I may try to understand your point of view, but I still have no way, other than my internal mental model of you, to "know" what you intend. There is no connection. I think any other point of view implies that there is a physical connection between the speaker and the hearer. The idea that the hearer cannot make sense out of the words of the speaker unless he takes the speaker's intentions into account, seems right to me, but that activity of making sense is happening totally inside of the hearer's head, inside of boundaries where the speaker has no access.

Clearly there are some intentions that the speaker may wish the hearer to capture but, of course, it is a very important aspect of communication that one can tell lies. It would be impossible to tell lies if understanding meant reliably recovering the all the speaker's intentions.

Let me repeat my remark about meaning again. I believe that if we are going to understand how to build robots, we need to accept the fact that the only things that we have to work with are its internal state and the effect on this state when something comes in contact with it. In the case of speech this means what he hears. So we, standing outside, may imagine that there is a speaker that's causing this hearing to happen, but if you want to describe the robot's architecture,

if you want to describe what's actually happening, the only information we have is what actually impinges on him, and it cannot matter whether or not it was someone with, so to speak, good intentions or bad intentions who said it. Even the existence of a speaker is irrelevant. Unless you want to postulate some form of "action at a distance", the change in the individual is going to be the same, and that's all we get to measure. I believe that any cognitive architecture that ignores this will fail.

Aside: Simple physical reductionism is not the worst of it. Here's a claim I made a long time ago, I think well before 1980. It is my opinion that we will not be able to build robots with intellectual capabilities comparable to ours, that do not have the possibility of having the same level of psychosis that current human beings have. The capability of having a rich mental life is the capability of having a screwed up mental life. We will have the choice of either having an artifact that has a very, very simple (even possibly autistic) understanding of the world, or an artifact that has pathologies, perhaps even pathologies that are comparable and similar to the worse psychological problems that we have as humans. This is a consequence of the same disconnection I spoke of above in the case of language.

If we have highly cognitive and self-conscious and self-aware artifacts we will have psychotic ones. I do not believe that psychoses are an emergent phenomenon. I believe that the same machinery that allows you to have good ideas about the world would allow you to have badly misguided ideas about the world, and the disconnection that you have from your exterior that I described above, makes it problematic for anyone, from the inside, to distinguish between the two. Once you have the facility of genuine judgement about the world, just as in the case of people, you are going to have robots that have really mistaken understandings of their world. And so psychosis is not an emergent phenomenon. It's not something that you must add to a system, it is not something that you can choose to introduce into the system. It just comes with smartness or with self-awareness.

3.2 How to start

I think that we can now begin to see what a theory of

individuals might be like. We imagine again that we are presented with an individual.

We might represent him in a picture as a closed curve with an outline like a potato. He is self contained. The line represents his boundary. Perception is represented by the fact that things in his exterior may push on the boundary, and as a result he may change shape or be rearranged internally. Pushing on him may also cause him to push back. Pushing on him with speech may cause him to hear and he may push back by moving his lips and speaking, i.e. his boundary may be changed by internal forces (the desire to speak) and well as external forces (random noise from well intentioned speakers). This picture accounts for perception.

We cannot formulate our theory by asking when I push on the boundary here, what kind of changes are going to happen. We need to be careful not to imagine that this is like a function being applied to these things. In particular, if our robot exists in free space, then you can imagine, just as for ourselves, that sense signals and all kinds of other signals affect us not only asynchronously, but are also geographically distributed around our exterior. This means that we have more than just a functional representation between the state of our robot before and his state after. We can walk and see and hear "at the same time".

An adequate theory of individuals must account for this distributed behavior, and in fact must be able to do so locally in space as well as in time. The theory of FOL systems provides a framework for constructing the kind of system that can move in this way and in which it is possible to describe the kinds of changes that such structures can make over time.

3.3 Conversations

But in practice, what kinds of conversations do we have to support in order to imagine that we have a reasonably cognitive object at all? A robot must not only be able to discuss simple problems and their solutions, but other things, some of which impinge directly on consciousness. In some ways it is the quality of the conversations that we can have with him that will determine just how "intelligent" he is. We want to understand what kinds of questions we can ask him about what he is doing, and what we can expect about

the kinds of answers he can give us. If we want a fully capable individual then these questions include, for example, something I call the gossip principle. The gossip principle is the ability, at any moment, for a system to answer the question: what have you been doing lately? In other words if you are thinking about self-consciousness, the ability to ask yourself the question: what have I been doing lately? This is a very general principle that allows you to step back from your current existing activity and possibly try to make a judgement about it, a conscious judgement, about whether or not the activity that you are currently doing is satisfying your goals. This principle is problematic because it is hard to imagine what mechanism can stop you, at any time, and leave you in a coherent enough state to think about what you were up to. The problem of self-reference is, of course, very important, and I believe that it is important (and possible) to solve it without infinite regress and without the (possibly infinite or arbitrarily extendable) towers of theories approaches in the literature. None of these theories encompasses the gossip principle, which allows "self" application during its use in order to answer "Oh, I'm just using the gossip principle".

Another question of conversation relates to persistence. We have to account for why an individual, as it persists through time, can still call itself I. We have to account for the fact that "now" always tells us the correct time. The attempt to account for this using indexicals has always struck me as strange. By indexing the interpretations of words over time, it seems that we basically end up with infinite regress, e.g. if 'now-now' represents the meaning of 'now' now and the 'then-now' represents the meaning of 'now' before, is the now-now now the then-now by the time you finish reading this sentence? .... I believe that the treatment of time as a modality needs to be examined.

If we really want to build individuals that function at or near the levels that we do, many forms of modalities are really important. They cannot be avoided. The simplest kinds of problem-solving include interactions between understanding people's wants, understanding motives, having desires, revenge, time, etc.

The list of modalities is almost endless and one can supply very simple problem-solving situations in which they

are all necessary at the same time. One of the things that is discouraging to me about the course of research in this area is the problem of compactness, in the mathematical sense, that is, although we have people that are working on theories of time, theories of belief, theories of wants, desires, hatreds, etc., I see no evidence whatsoever that these things can be used in problem-solving contexts together, and virtually any interesting problem-solving activity that you can imagine is likely to require one to understand more of these at a time. It worries me that no one seems to worry about this. We have a long way to go to understand the nature of an architecture that can move back and forth between the many modalities.

3.4 Logic

Q.: Is formal logic relevant to any of this?

A.: I think of logic very structurally. One of the reasons for having a logical architecture is that we can explain why we believe that a robot can continue to have a coherent view of his world over time. I use the word 'coherent' here because I do not want to use the logical word 'consistency'. I do not believe that people are consistent in the logical sense, but I do believe that the vast majority of people have a quite reasonable picture of the world and that luckily that understanding is in fact close to what the world is like. As we think about individuals going through time we even imagine that for the majority the understanding of the world gets better. We mature. We have more experience. Again I ask how is it that we come to know. More particularly, we probably would not be very happy with individuals if their world models showed vast discontinuities. Religious conversions and strokes, leave people uncomfortable.

In FOL systems I use logic as a way of saying things about what it means for structures to be coherent and what it means to be coherency persistent through time. I believe that with respect to the kinds of architectures we are talking about, it is not enough to say: I have this collection of things. I do not believe that we have a blueprint if we do not have some reason for understanding why we have before us a coherent robot and why the actions we are imagining that he may do through time preserve that coherency. I am not just an experimentalist, who will build an artifact and let it run until it dies. I want

to have some confidence that, within some bounds, what I am building will work. In the case of FOL, rather than thinking about logic as trying to determine what's analytic, we use logic to tell us how to design data structures for theories. A physical data structure for theories. The notion of language, model, consistency, etc. in logic are used to understand how good those data structures are. Consistency of an individual theory is not a very important property, but we can use consistency as a measure of coherence and thus we can begin to say what it means to be coherent over time. And so, what the theory of FOL systems does, in fact, is to propose that we build individuals out of a collection of theories, collections of constructions, for which we have defined a certain notion of consistency. Relative to this notion and without prejudice about what kind of operations you use to represent changes through time, or what external properties you might want to preserve of our robot, some form of coherency preservation makes sense.

Notice here that we differ from normal discussion of logic, because we are talking about consistency-preserving changes in systems, not validity-preserving changes. I do not talk about abstract "truth", in FOL we talk about preserving the consistency of the individual. It is consistency preservation that is at the core of ongoing individuality itself. This individual may have certain beliefs, and later he may have completely other beliefs. What is important is not that he has preserved any particular "truth", because things stop being true, even if some philosopher would object to that, but the world changes (I think that statement is probably unproblematic) and what we really want in an individual is that whatever understanding he has now is at least as coherent as the understanding he had then.

Think of passage through time as an operator that acts when, for example, some vision happens. Then try to imagine the way our robot's structure is updated as a consequence of seeing. To succeed in individual building we need to ask how the effect of what is seen or heard is incorporated into our robot's existing state, into his existing structure, and we would like to, at least in the normal cases, say that he has gone from coherent state to coherent state. Notice that our description does not imply that the state he was in

corresponds, in any fundamental way, to what his exterior is like, any more than the consequent does. But, at least, it is a way of describing what we think the effect of something happening to him through time is. We have some reason to believe that if he were coherent before, he is coherent after. Therefore, if his model of the world was reasonable before, and if this operation corresponds to an actual event of the world, then his model of the world later is going to be reasonably coherent with respect to the way it actually is.

I propose this kind of paradigm for thinking about how to describe the kinds of internal events inside individuals, and I have used logic to build these structures. I have used logic not as a problem-solving tool, but as a representation tool using theories as data structures, to give me a reasonable sense of what coherency means and what it means to preserve coherency. By using logic to specify these data structures, I claim that using FOL *contexts* I can actually provide operators that move these structures through time and representations of mental states. These FOL contexts have representations of objects in them and you can have many of these theories co-resident at the same time.

Q.: Why have you persisted over the years with first order logic when other "richer" systems abound?

A.: First order logic is incredibly important. It is certain that something of the declarative strength of first order sentences is a minimum capability that representational structures must have. I am not saying that the syntax of first order sentences is necessary, but rather it is the declarative content of first order sentences that is necessary. The essence of being first order is that whatever things I am thinking about I can name them, I can talk about the relationships between them and I can mention their parts.

Technically, this means you need individual symbols that can name things, you have function symbols which construct new things and tear things apart and you need relation symbols to name relationships. If, in addition, you want to say simple things like: everyone in this room wants to go to lunch, or there is an apple in the refrigerator, then having quantification is a reasonable feature. Now, this is an argument for first order ideas as minimal. But what about modal logics, do not you need temporal logic, what about non-

monotonic reasoning, etc? Well, part of the answer lies in how promiscuous you are willing to be with respect to ontology. In some direct way higher order things and modalities all become first order things if you are willing to admit that abstract objects are real. Disconnection with your exterior really is powerful. If you do not have any philosophical problem with that, if you do not mind that round squares are legitimate objects, and are clear that the existence of round squares does not impact on your coherency, then there is a deep sense in which first order has to be enough.

Now, I do not consider that to be an argument, if anybody wants to see this filled out in detail, they would have to understand a little bit of how I put together FOL, how I do metatheoretic reasoning and how one does what I call multi-context reasoning, that is, I replace higher order notions and modalities with contextual first order notions. And I would claim that even if that is not enough to do everything, which is probable, it can be used to represent the existing theories of modality, of non-monotonic reasoning, problems of reference, *de re* and *de dicto*, problems of cross-word individuation, etc. Furthermore, it does this in one architectural framework, and for that reason alone it is an interesting framework to look at. Even if you were to convince me that second order entities are interesting, it is not my task to simply increase the expressiveness of the consistent language of a system. I know how to increase the expressiveness of first order languages by introducing various higher order extensions. The question is: does that introduction do anything except expanding my problem-solving capability slightly? If you are interested in the simple case of meta-theory or the difficult case of the gossip problem, then you must have something else besides simple logic as a truth-determining tool in any event.

FOL uses logic in the local sense of having many internal theories, in order to be able to ask if something is true, but, at the same time, it is used to produce coherency preserving travels through time, adjusting one's theories based on what perceptions you have had in the meantime.

Q.: What you are trying to get away from, by the use of logic for problem-solving, is the well known problem of inconsistency which you talked about in terms of coherency, and then you said that you use logic to build representations

rather than to solve problems. I always think of logic as a process, not as a standing thing, and a representation is sitting there not doing anything, so what is logic doing in a representation if it is not transforming it into some other representation which, in fact, is doing problem-solving and where the problem of consistency and errors come in?

A.: Let me make another little step here. In FOL I give a complete description of what it means to be a theory represented as a finite data structure. It does not correspond exactly to what logicians describe as a theory because among other things it has the property that each of these theories is constructed as a finite data structure. So, we have taken some infinitary extensional object (a theory) which might have a nice extensional description in set theory, and replaced it, in all cases, with a finite data structure. Then, for these data structures, which we call FOL contexts, we define an appropriate notion of consistency, consequence and satisfaction. FOL contexts consist of a linguistic part, which represents the language that is to be used in this theory, a semantic part, which is an explicit representation of a partial model of the theory, and a collection of facts, each of which contains a first order sentence, written in the language of that context, about the objects of the theory. The important idea, however, is that we have reduced an infinite set theoretic object to a finitistically presentable data structure.

It is important that it be finitistically presentable, because it contains "functions" and "relations". It is also compatible with classical reasoning. So what we have done is to replace a theory with this data structure. Because we now have data structures for theories and each FOL context contains, as part of its structure, an explicit partial model, one of the things that we can use FOL contexts to talk about are FOL contexts themselves. Contexts reasoning about contexts rather than theories of theories. The finitized version of normal traditional metatheory.

By building hierarchies of these theories we can talk about the relations among theories and about groups of theories. In fact we can say in a "meta" FOL context, what it means to be a consistent context. In this setup we can now have a context that talks about another context which is itself not consistent. This pair of contexts, however, is consistent! It is in this way

that we can use coherency to describe inconsistent problem solving.

So here comes the punch line. By connecting together enough of these contexts (delicately of course) we build up a system that can be thought about as an individual. Although I will not do so in this paper we can describe what it means to be the boundary of one of these systems, and then we can ask about what properties of these systems should be preserved when various operations are performed. We call these external properties the physics of a system and they describe a set of constraints on the possible futures that a system may become. Of course, we probably do not want to consider operations that are not consistency preserving.

The point is that we are not using logic for problem-solving, although some forms of problem solving may go on, but as a data structure in which the mental life of an individual can be built. From the outside, we do not see theories. We simply talk to this thing and, on the basis of its responses, decide if it is cogent enough to be paid attention to.

Q.: It seems that your theories presuppose that there is a categorization of the world and so, one of my questions is, where does this come from? This is related to the problem of what is the state of one of these individuals. It seems that you have a theory about phenomena that you observe and that, as a theorist, determine to be a coherent state; and then some information comes in and by looking at the next state you can determine if it is still coherent. If this is the case, then you have a sort of external-view-point theory, but you stressed in the beginning that you also wanted the theory of the mechanism, and while I am willing to agree that this is, perhaps, an external theory, I doubt that it is going to be a satisfactory theory of the internals.

I am thinking particularly about a robot, as you said in the beginning, that is moving around in the world, that has to pick up information from the world through sensors and has to relate it with internal representations. I have questions about where the knowledge comes from. It seems to me that it is going to be very difficult to make a predictive theory of how these states will be making transitions from one to the other, because it seems to me that the complexities of the world

require that we understand much more of the phenomena of emergence of meaning and of interactions between the world and the growing individual.

A.: In fact, what I have presented up here is content free in the sense that there is no intimation about what is in any of the theories that I mentioned. You are quite correct, it does not address the issue of what is in a theory of the world, or how it might arise or how enough information might be incorporated into a system to make it functional. I have tried first to give a framework in which to discuss these problems. As anyone who has tried to build a robot knows, the question of how perceptual data become cognitive data is an immediate problem. On the other hand, the workers in representation theory do not even imagine that this problem exists. If we are to make progress we need to get both groups working on the same problem. One reason why problem-solving has not interested me much, is that I have never seen a representation theory that explained how the system itself could ever come to know he was solving a problem. The solutions obtained by AI problem solvers have never been solutions in the head of a robot: even the fact that they represent problems has always been only in the head of the researcher. We need to have representation theories that are powerful enough to contain the information that they are solving a problem.

Q.: Have you actually built a robot?

A.: I have not built a physical robot, but we have used FOL to build a purely electronic robot that had ears (a keyboard) a mouth (a CRT) and eyes (a simple set of sensors) but no mobility.

Q.: It seems to me that in human beings consciousness and self-consciousness and ego, among many other things we have, might have a strong survival value. Perhaps these things develop through evolution in a chance way, but in a way that maximizes their survival values of the species. So it seems to me that perhaps consciousness and self-consciousness might have co-evolved without an external design. Whereas here it seems to appear that you want to put something into the robot by an explicit design. Is there any survival value for consciousness in a robot? Is this a relevant question?

A.: I think that there is huge survival value in self

consciousness. What I would like to say about the former remark, is that I do not believe for a minute that we understand enough about the theories of our world to "put the stuff in there". But I believe that in order to build a successful robot we are going to have to know enough about what principles are used to construct theories out of experience so that our robot will be able to make some unpredetermined use of the sensors we give him. It is quite possible under those circumstances that a robot's understanding of the world could change quite radically, given some experience, and not only will we not have built such ideas into the robot directly, but, we, lacking this experience, may not be able to find them out.

4. Conclusion

In this paper I have tried to preserve some of the interactive flavor of the discussion. I apologize for mangling the questions of others into the form that was heard by me, but by leaving them anonymous I hope that any blame for foolishness can be absorbed by me. I hope that the Worshop and the ensuing papers will stimulate hard thinking about the creation of artificial consciousnesses and fully cogent robots. Perhaps even genuine artificial life. I do not think that these are impossible dreams or far off science fiction. Real sentient non-biological thinkers will come sooner than we think. Let us not be afraid and instead all work to make it happen.