WHAT IS THE PURPOSE AND STRUCTURE OF INTELLIGENCE?

Ronald L. Rivest
August 16, 1983
MIT Laboratory for Computer Science, Cambridge, MA 02139

The brain is not an appendix. It is not vestigial or ornamental. It has survival value. The "purpose" of the brain is to improve the chances thaat its owner will survive and flourish in a complex environment. It does so by providing memory, intelligence, and adaptability. The purpose of this paper is to explore some of the ramifications of this point of view for artificial intelligence.

We begin by observing that intelligence and intelligent behavior does not arise from the void, but from the interaction of a being capable of intelligence with a complex environment. All three are essential: the being, the environment, and their interaction. Much previous work in AI has implicitly held the view that intelligence was somehow intrinsic to the being. Considerations of the environment or the interaction of the being with the environment were often neglected. But it is clear that the utility of intelligence to the being is a function of the nature of the environment it lives in, and the character of the interactions it can have with the environment.

The interface between the being and the environment can be described as a set of input channels (senses) and output channels (effectors). The interaction between the being and the environment can be described by describing what information flows through these channels. Any information the being has about the environment is obtained from its senses. Any effect the being has on the environment is made via its effectors. This is all straightforward.

We need to understand better the nature of an environment that can favor the evolution of intelligence. Clearly, the environment should be complex. What does this mean? Perhaps a complex environment should be defined as one with much "hidden structure"; structure that is not immediately evedent to the senses, but which can only be inferred on the basis of extensive experience.

Second, the environment must offer intelligent beings an increased chance of success (survival). Intelligence must help the being to avoid fatal mistakes and to adapt to changing circumstances. A being that is fed or not each day solely on the basis of a cosmic coin-flip would have no need for intelligence.

Our framework presupposes that the being "lives" in the sense that it participates in a continuous interaction with the environment from its creation until its death. It "has experiences" and "learns from them" (presumably). Its life has a "past", a "present", and a "future".

The fact that the being lives in time, and not somehow outside of time is a central observation on our analysis. the survival of the being does not depend on isolated instances of problem solving, but on the ability of the being to continuously interact with the environment to its own advantage and security.

How, then, does the being's brain improve the survival chances of the being? I propose that the function of intelligence is to provide knowledge of the future. This proposition represents one of the major points of this paper, so it requires some careful explanation and discussion.

For the purpose of survival, the future is all that matters. All events that have occurred up to and including the present instant are fixed and unchangeable. The survival of the being is a question about the future of the being. The (future) actions of the being may determine whether the being dies in the next few seconds or decades from now. Knowledge of the past is only useful to the extent that it provides a basis for future action.

It is amusing to consider a being with a built-in "crystal ball" (perhaps it is born with one in place of a head). The crystal ball allows the being to predict the future perfectly. One can easily see what a tremendous survival advantage that would provide! The being would have no need for senses since the sensory inputs would have been perfectly predicted. The being would have no need to laboriously reason out the consequences of various possible actions, since the crystal ball can predict these perfectly. Thinking as such would become superfluous to survival. Avoiding danger (nearby tigers) would be trivial.

I would suggest that a head (brain & senses) is a good approximation to a crystal ball. Its purpose is the same (providing knowledge of the future), and its construction is the best that Nature can provide.

What is "the future"? It is not a single, fixed, linear sequence of events like "the past". Rather, it is a collection of possibilities and potentialities, predictions and expectations, branching out to infinity. It is indeterminate in the sense that which events actually occur will depend on free choices by the being and on random events in the environment. This is why the future can not be known in the same manner as the past. Knowing the future means knowing the entire branching tree of "possible futures" (perhaps with associated probabilities). "Knowing the future" does not preclude choice or even uncertainty. (Uncertainty may be an instrinsic feature in a nondeterministic environment, where a "perfect" crystal ball could at best provide imperfect predictions.)

The distinction between "present sensory information" and "future expectations" is a very important one in the theory being presented here. I began slowly to realize how deeply our acts of "perception" are entangled with acts of "expectation". The depth of this entanglement was very surprising to me. I would like to try to convey this appreciation to the reader.

One has the feeling most of the time that the immediate environment is hwat might be called "directly perceivable". One can walk about, manipulate objects, interact with other people with the comfortable feeling at each moment one is "in contact with solid reality". Nearby objects are perceived as recognized simply by glancing at them; one knows where they are relative to oneself. Things are "under control"; there are no surprises. One is "in touch". One has the impression that the state of the local environment is being "directly perceived" by the senses. I would like to argue that this feeling of "direct perception" is wrong in that most of our perception of the immediate invironment can be categorized as "expectations" rather than "sensory perceptions".

This argument is a little difficult to make, for it requires arguing against one's natural feelings and intuition. However, I would argue that the feeling of "direct perception" is a convenient fiction maintained by unconscious micro-processes in our brains. A useful fiction, but a fiction nonetheless.

Of what does our "perception of the immediate environment" consist? One might tyr some representation of objects in three-dimensional Euclidean space with color and texture added as a candidate. I would argue that this is a poor approximation to what we really maintain as a "local world model" in our heads.

To begin with, why should we maintain a "local world model" at all? I would argue that it forms a basis for choosing actions in the immediate future. It is a guide for navigating and predicting in the space of possible micro-futures. Knowing the door is closed, I can wisely decide to open it before trying to go through it. Clearly, the local world model is useful in working with these micro-futures. However, I would argue that these concepts are in fact nearly merged in our brains, and that our local world-model is made up of expectations and micro-futures. (Can you think of a door without thinking of it opening and closing?)

It has widely been observed that sensory perception is perhaps more an active than passive process. For example, we note that the fovea of the eye is capable of perceiving only over a small angular region. The process of perceiving a simple object (i.e. a cup or a pen) may involve dozens of eye fixations as the different features and edges are rapidly scanned. (The model of the visual system as camera which takes a single "snapshot" to be processed by the visual cortex is highly inaccurate.) As each new feature is scanned, expectations are raised about the locations and characteristics of other features. These expectations guide the scanning and eye-movement process. Eventually the location and identity of the object is decided upon.

If I close my eyes for a moment, my local world-model does not dissolve or change significantly. I expect that, once I open my eyes again, my desk will still be thhere as before. It is this expectation, rather than the direct perception of my desk, that is an integral part of my local world-model.

Consider the task of crossing a road. I look left: the road is empty. I look right: also empty. While I am looking right I may maintain a "mental image" of the road to the left as previously perceived. However, it is obviously no longer directly perceivable. It does, however, represent an important part of my local world model. Important not because it represents what I saw, but because it allows me to predict (expect) that stepping into the road is a safe operation. Once I look right, the empty road perceived to the left changes status from that of an more-or-less immediate perception, to that of a prediction (if I look left again, the road will still be empty). This prediction is typical of the hundreds of such predictions that a typical local world-model would contain.

I would maintain that the local world-model consists of a set of predictions which can or are likely to be verified or disproved in the very near future. Each such prediction may be in the form of an if-then rule, e.g. "if I lean backwards, I will feel the back of the chair on my back" or "if I look up, I will see a round ceiling light" or "if I press down with my pencil and move it, I will see a black line". These all represent futures that can be realized by selecting the action specified. They are, however, only predictions that may fail to be ture. The back of my chair, unnoticed by me, might just have been devoured by a horde of hungry termites, so that I will fall on the floor when I lean back. (Whew! I just tried it. It's still there ...) In general, these predictions are so unerringly relaible that they can be substituted for direct perceptions. What is the need to actually look up at my ceiling light if I am sure that it is really there? A good crystal ball would, we recall, obviate any need foor the senses. Our local world model is thus made up of a combination of direct perceptions and more-or-less infallible predictions. The solidity and reliability of our contact with the world may be a function of the exterme accuracy of these predictions, rather than evidence for the "direct perception" view.

Some of the elements of my local world-model may represent expectations such as: expecting to hear thunder after seeing a lightning flash, or expecting to be squirted when my four-year son points a squirt gun at me. They begin to involve more directly the temporal aspects of the future, compared to more static expectations such as "if I look up, I will see my ceiling light." These expectations may arise quickly and be quickly verified or contradicted. Yet they represent a part of my local world-model as real and reliable as my ceiling light.

I would like to give two reasons why we should accept the proposal that a local world-model is a set of predictions which can be quickly verified or disproved.

The first reason is that "predictions" are sufficiently general as an information-representation structure that no generality is lost. For example, the sense inputs can be encoded as predictions, valid for very shoort periods of time. Having seen a cup, we can predict (for a short while) that we will continue to see a cup. The world tends to be static, on a small enough time scale.

The second reason is that "predictions" capture all that is immediately relevant to the survival of the being. Sense data that has no predictive power is treated as noise and ignored or forgotten. Any part of the local world-model that is not useful in helping the being predict the set of possible futures and courses of action is not terribly useful to the being. For example, background noise is generally useless in this sense.

The construction of a world-model from "predictions" has two intriguing consequences:

1) The world model becomes more difficult to update when actions are taken, relative to a "direct perception" approach where all consequences are directly perceptible,

2) Treating a (accurate, reliable) prediction on an equal status with a direct sensory perception allows for the creation of "generalized sensory inputs" and the building of concept hierarchies.

We discuss these issues in turn.

Suppose the world-model of the being contains the prediction "if I look straight ahead, I will see an ice-cream truck." Suppose now the being turns right (say by 90 degrees). The prediction in the world-model is no longer correct; it can be easily disproved. (By looking.) However, a new prediction should be inserted in its place: "if I turn left by 90 degrees, I will see an ice-cream truck." Somehow the being needs to know that taking a left-turn should be sufficient reason to replace the first prediction by the second in his world-model.

One might argue that people have models of space "wired in", and that such transformations are automatic and unlearned. Perhaps, I would suggest that they are learned. As partial evidence, I would argue that our world-models may contain dozens or hundreds of predictions that are unrelated to spatial operations, and which need to be transformed quickly as actions are taken. A good chess player maintains a complicated world-view with a variety of attack and protecting relationships that need to be updated with each move. A driver (especially in Boston) has learned how to maintain an accurate world-model as he interprets brake and turn signals, traffic lights, and horn signals. Riding a bicycle is another skill which requires the more-or-less automatic updating of the world-model of the bicycle's behavior as actions are taken. Using a typewriter is another example of a situation where a number of simple predictive rules must be learned. (When you hear the bell, the type-ball can be observed on the right-hand side of the page. If you press Return, it will move to the left-hand side.) It would be difficult to argue that we have models "wired-in" for all of these tasks. Yet we seem to be able to learn the rules of these micro-environments and to work within them as automatically and unconsciously as we work with the rules of spatial transformations. I would suggest that a powerful, general mechanism is employed which is capable of learning the appropriate world-model transformations. This will be discussed at more length later.

The second consequence of using predictions as the atomic constituents of a local world-model was that they allow for "generalized sensory inputs" and the building of concept hierarchies. We begin by noting that the sensory channels have only a very limited expressive capability. They may provide a rich set of cues about the state of the environment, but they do not give a complete description of the current state (as the direct perception model would suggest). They certainly have a very limited expressive capability intrinsically, they are only powerful because of what they can suggest (allow one to predict). On the other hand, the vocabulary of predictions can be arbitrarily complex. The complexity of the local world-model is not thereby limited by the language descriibing sensory inputs, but can be made complex as new concepts and generalized actions sare introduced to permit description of higher-level (but still reliable) predictions. As the language describing predictions is enriched, the atomic components of the world-model (the predictions) can achieve ever-higher levels of complexity (hierarchical depth). Since we have the illusion that each prediction in our local world-model is "directly perceptible," an enriched prediction vocabulary allows us to "see" deeper levels of meaning in the current situation. An expert skier "perceives" the slopes and moguls in terms of their possibilities and opportunities in a way that a novice is incapable of imagining.

This brings us to the question of "learning." Fortunately, the frameqork developed so far provides some guidance as to what should be learned, which is often a difficult or confusing question. In our case, the being should be able to learn

1) how to update a world-model as actions are taken and sense inputs received,

2) new concepts that facilitate the description of new and more useful predictions in the world-model.

We shall see that these activities are nicely synergistic.

I would suggest that one might define an intelligent being as one that is capable of being surprised. A being that is capable of being surprised must have expectations or predictions about the world, since a surprise is simply an event thatt is contrary to expectation. However, this definition would be incomplete without indicating that the being outht to "learn something" when it is surprised. (You can trick a fool many times, but a wise man only once.) I would suggest that the important learning of type (1) (how to update a world model) happens when "surprises" occur.

Suppose the current world-model contains predictions P may suppose that each such prediction is off the form, "if X then Y" where X and Y are statements in a suitable language (which may mean that they include other predictions as components). Suppose then that X occurs. Here, perhaps, X might be a specific action or specific sense input. Then Y should be included in the current world-model. If Y specifies a directly observable attribute of the environment, it can be tested directly. In this case a sense input would be input which might contradict the presence of Y in the world-model. A new transformation rule might then be generated, which would predict the failure of Y. The new rule might hypothesize tthat some of conditions P interesting to note that I am proposing that only elements of the current world-model (i.e. predictions) are suitable bases for such rules. The only interesting statements about the current state of the environment may be statements about its future (predictions).

We note that the default situation is that a prediction remains in the current world-model until it is contradicted by a sense datum or by a contrary prediction made by a newly invoked rule. This default is the right one to have.

The details of how the local model updating might occur are left pretty vague here, but are fortunately in a format similar to much of the "learning a rule form instances" literature. A set of predictions is the precondition, an action is taken, and some new predictions are represented in the (correct) new world-model. The predictions that ought to be in the new world-model may have to be inferred backwards from the interaction following the action.

Concept learning fits nicely into the same framework. One of the most important ways new concepts can be induced is by generalizing the preconditions for a prediction. I favor the definition of a chair as something one can sit on. As another example, turning right 180 degrees and turning left 180 degrees are actions that should have the same predictive consequences; we can generalize them to "turning around." Similarly, our notion of a "ball" is a generalization of the set of predictions associated with a ball indicating what one can do with it. When one thinks of a ball, one thiks of holding it in one's hand, or looking at it, or throwing it, catching it, bouncing it, etc. The entire notion of a ball is just an interrelated set of expectations and predictions. In general, our concepts of objects may be just sets of mutually predicting expectations, that tend to enter and exit our would-model as a group.

Again, the above discussion is rather vague. However, the underlying thrust is that new concepts are developed as a means of unifying and generalizing previous predictions. While it may seem strange to think of objects as complexes of expectations, I feel that this is a fruitful point of view.

SUMMARY

We present the thesis that the purpose of intelligence is providing knowledge about the future, and the corollary proposition that a local world-model should consist of a set of predictions and expectations. These predictions can be taken as atomic units in the development of higher-order concepts and rules; having a reliable prediction about a future course of possible action should be considered as an event as reliable and useful as a direct sensory perception; it is "seeing" the future. This framework provides a natural test-bed for developing a theory of learning: when surprises occur new prediction rules can be hypothesized, and concepts can be induced by generalizing over several rules.

ACKNOWLEDGEMENTS

This work was supported in part by NSF grant MCS-8006938.

I would like to thank Pat Winston, Peter Andreae, Bob Giansiracusa, Gary Drescher, and Phil Agre for many helpful discussions and pointers. (Of course, they bear no responsibility for any of the inadequacies of this paper.) Also, some of the points made here have been made previously by others (e.g. Schank's observations about the connections between memory and learning and

failed expectations). A later version of this paper will include some of the relevant references.

To the reader: COMMENTS, CRITICISM, and REFERENCES TO RELATED LITERATURE would be MUCH APPRECIATED!