# Will Robots Ever Have Literature?

Jerry R. Hobbs

Artificial Intelligence Center

SRI International

Let me begin with a statement that may be provocative: I believe "reductionist" is a good word.

Evolution has proceeded by "levels of organization", representing "levels of competence". Each level is characterized by the achievement of stable forms, out of which larger structures can eventually be constructed. Molecules are stable forms constructed out of atoms; stars, rocks and cells are stable forms constructed out of molecules; multicellular animals, including people, are stable forms constructed out of cells; and social organizations are stable forms constructed out of people. *The classic sciences—physics, chemistry, biology, psychology, sociology—each take on a level of organization to describe, producing what might be called a "level of description".*

I believe it is barren to argue about whether a particular field of inquiry is a "science" or not. But there is a stage that some sciences have gone through and others have not, that represents a qualitative advance. It happens when it is understood, at least (and generally no more than) in principle, how the entities and processes at one level emerge from entities and processes at lower levels.

Geology passed through this stage in the 1960s with the widespread acceptance of plate tectonics. Before then, ever since the eighteenth century, explanation in geology bottomed out in a mysterious process of "uplift". Plate tectonics explained uplift in terms of underlying physical processes. At that point, geology became, in the sense of reduction I intend here, a "reduced" science.

This kind of reductionism does not mean the terminology and laws of the higher science can be restated in terms of the terminology and laws of the lower science. The higher science generally concerns itself with emergent entities whose boundaries become very fuzzy when unpacked into the entities of the lower science. Moreover, it does not mean that prediction becomes possible in the higher science, resident on the laws of the lower science. The entities of the higher level are generally very complex dynamic systems of entities at the lower level, and although gross regularities may be established, the fine details of higher entities and processes cannot be derived. We understand the underlying physics of rivers, hurricanes, and volcanoes, but we can't predict their behavior, except within very coarse limits.

The central metaphor of cognitive science is "The brain is a computer". In the long history of inquiry into the nature of mind, the Computer Metaphor for Mind gives us, for the first time, the promise of linking the entities and processes of intentional psychology to the underlying biological processes of neurons, and hence to physical processes. We could say that the Computer

Metaphor for Mind is the first, best hope of materialism.

The jump between neurophysiology and intentional psychology is a huge one. We are more likely to succeed in linking the two if we can identify some intermediate levels. A view that is popular these days identifies two intermediate levels—the symbolic and the connectionist.

Intentional Level

|

Symbolic Level

|

Connectionist Level

|

Neurophysiological Level

The intentional level is implemented in the symbolic level, which is implemented in the connectionist level, which is implemented in the neurophysiological level. The aim of cognitive science is to show how entities and processes at each level emerge from the entities and processes of the level below. In my view, this picture looks very promising indeed. The elements in a connectionist network are modeled very closely on certain properties of neurons. The principal problems in linking the symbolic and connectionist levels are representing predicate-argument relations in connectionist networks, implementing variable-binding or universal instantiation in connectionist networks, and defining the right notion of "defeasibility" in logic to reflect the "soft corners" that make connectionist models so attractive.

Mainstream AI and cognitive science have taken as their task to show

how intentional phenomena can be implemented by symbolic processes, which may as well be viewed as logical processes, provided "logical" is interpreted loosely enough. The most common theories at the symbolic level use a terminology that is borrowed from intentional psychology but is given precise computational definition. The vocabulary includes terms like "belief", "goal", and "plan", and the thing whose behavior is being modeled is usually referred to as an "agent". The agent could be a human or a robot, and from this point of view the question in my title, "Will robots ever have literature?" is the same as the question, "Why do people have literature?"

Beliefs and goals can be thought of simply as data structures which interact with the agent's perceptions and actions in the right way. Specifically, the agent's beliefs are, for the most part, consistent with its perceptions, and its actions tend to bring about its goals, given its beliefs.

We of course differ vastly from the robots that have been constructed to date—in material, in complexity, and in how we came about. But we nevertheless recognize that we are in the same epistemic situation that robots are in, and this fact throws light on some recent debates in literary theory and in wider intellectual circles.

At SRI we have a robot named Flakey who rolls around the hallways on four wheels, sensing his environment with sonar and a TV camera. He interacts with the world through very narrow windows as he engages in very primitive behavior. He looks for doorways, and he avoids obstacles. In addition, there are fleeting images that give him pause. He is programmed not to run into them, but they do not have stable locations, and if when he senses again, they are gone, he proceeds. These fleeting images are people, his creators. Flakey is completely unaware of most of what is in his environment, just as people, without artificial aids developed only within the last century, are completely unaware of the vast bulk of the electromagnetic spectrum.

Flakey perceives whatever observable data is made available to him by his sensors; he, in a sense, constructs a theory that explains that data; and he acts on the basis of this theory. His actions modify the world in some way, and the cycle begins again, as he observes how the world has changed.

We humans find ourselves in the same epistemic situation. We perceive certain features of our environment, we form the best theory we can to explain them, and we act on the basis of this theory. We observe a richer set of features than Flakey, we are far more adept at forming theories of the data and the theories are far more complex, and we are capable of many more actions. But the situation is qualitatively the same.

The data is generally relatively sparse, and consequently underdetermines the theory. Many theories will explain the same data, although of course the more data there is to be explained, the fewer the available theories. One often suspects that some literary theorists jump from this perfectly true observation, that the data underdetermines the theory, to the unwarranted conclusion that therefore theories are completely **un**determined by the data, and that they are therefore arbitrary.

We occupy an Olympian point of view with respect to Flakey. We know the truth about the world he occupies, and we can tell when he has incorrect beliefs, when he has constructed the wrong theory. We know there are some things he will never get right because he lacks the language for it and lacks the means for extending his language to encompass them.

By analogy with Flakey, we can understand how **we** could have the wrong theory of the world, and how our very language permits only a narrow range of theories, how our language embeds a theory, how even the very data that a theory is supposed to explain, above the most primitive level made directly available to our senses, is itself the product of a theory at a lower level. *In a way, even the sense data is theory-laden. Just as we have equipped Flakey with the sensors that seem to be the most functional for a creature of that sort in his environment, so evolution, that massively parallel problem-solving process, has done for us.*

But the analogy also enables us to imagine that there is a truth of the matter, even though we can never know it. It allows us to understand how one theory could be better than another, in the sense that it explains more of the data more simply. It allows us to understand how one language could be better than another in the sense that it constrains the set of possible theories less, or leads us to better theories sooner. We can imagine what an Olympian point of view would be like, even though we don't occupy an Olympian point of view with respect to ourselves.

A common device of deconstructionist critics is to show that a literary work whose content is $P$ in fact by its style or form conveys $\neg P$, thereby undercutting the whole $P - \neg P$ distinction, and calling into question the language of which $P$ is a part. This is a move that an AI researcher should be very sympathetic with. He or she is in the business of making up languages for capturing some phenomenon, and has experienced failures in the process. He or she is very aware that the choice of a language sharply delimits what can be thought. This is true for our language about language, just as for our language about other things.

*The analogy also throws light on the controversies in the past few decades concerning the nature of the interpretations of texts. Texts are observable data—at some level, whether of words, letters, or patterns of ink on the page. To interpret a text we need to come up with a theory that explains that data. An interpretation is a theory of the text. But theories do not arise de novo out of the observable data. They are a product of the interaction of the data with what we already know. We can represent this schematically by the formula,*

$$F(K, T) = I$$

*$T$ is the text to be interpreted, $K$ is the knowledge base used or assumed, and $F$ is the process that accesses $K$ and $T$ to produce an interpretation $I$. Just how each of these is realized in data structures and computational procedures is a matter of healthy debate in artificial intelligence.*

*From this point of view, many of the controversies in literary theory about the nature of interpretation simply evaporate. The debates have centered on the question of whether interpretations can be fixed, and if so, how. From our point of view, E. D. Hirsch's contention that the only meaning of a text is the author's intended meaning reduces to the statement that the only $K$ against which the text should be interpreted is the $K$ that the author assumed* *to be shared with the audience. New Criticism can be viewed as attempting to standardize $K$ in another way to include only those beliefs that an informed, but not too informed, reader would possess. Stanley Fish can be viewed as having observed that $K$ is a necessary argument of $F$ and then neglecting $T$. For early Fish, $K$ is the beliefs of the individual reader; for late Fish, it is the shared beliefs of an interpretive community.*

*The simple truth is that interpretations are a product of texts and a set of beliefs. Fix the set of beliefs and you fix the interpretation. Just how $K$ is to be fixed is not a question of the nature of interpretation, but rather a question of the function of literature. If we view literature as a way of holding conversations with the great minds of the past, then we should probably use the $K$ that the writer assumed to be shared with his or her readership. If we are interested in the psychology of the writer, we are justified in using facts about the writer of a more private nature, "to what lady, while sitting on what lawn". Or we can try to read in a purely exploitative manner, ignoring the author's intentions, interpreting against our own $K$, and getting what we find most useful out of the words on the page.*

We can also understand by analogy with robots a very significant feature of our epistemic situation. We, both humans and robots, are very finite creatures. We have very limited computational resources. We cannot reason about the whole world at once. The way we proceed is to fix the vast majority of our language and our knowledge base, declaring it in a sense to be unproblematic. They are the resources we will use. We then focus on a few facts or concepts which we take, for the nonce, to be problematic, and we solve the problem the best we can, given the resources. This process may yield a refined notion of those facts and concepts, which then take their place among the “unproblematic” background resources, as our focus shifts to a new problem. There is no particular difficulty in the same concept being used as one of the background resources, and at the same time being treated as problematic. We are using an old version of a concept to bootstrap into a new one. You might think of this as proceeding from one approximation to the truth to a slightly better one. There is thus no more paradox in using language to reason about language, than there is in using ink to write about ink.

The Computer Metaphor for Mind gives us a very concrete model of cognition that throws light on many epistemological problems that have puzzled philosphers over the centuries and confused many literary theorists in recent decades. On the other hand, the study of literature poses significant challenges for cognitive science, and I want to discuss some of those next.

Whatever else it is, literature is a kind of discourse, which is a kind of intentional behavior. It is therefore something that cognitive science ought to be able, eventually, to explain. In fact, it is a particularly challenging something to be explained.

*There are a number of phenomena that are sometimes described as challenges—indeed, that are sometimes cited as refutations of the possibility of formal or computational approaches to literature, that are simply not refutations, or even challenges, at all. Among these are ambiguity, metaphor, and the open quality of literary texts. I want to discuss, or rather dismiss, each of these* *in turn.*

*One sometimes hears that computers cannot handle ambiguity. It is certainly true that a computer cannot follow an ambiguous instruction without first settling on one reading or the other. But the same is true of people. Computers can certainly* **represent** *ambiguity—it is simply a disjunction about the meaning of something. And they can certainly reason by cases, that is, reason about what would happen if each of the readings were true. And they can refrain from action when there is insufficient information to make a good decision. There are no special problems here, no special differences between people and computers.*

*Related to this is the assertion that computers are only capable of binary distinctions. This is true in the sense that it is constructed out of binary elements. Literary theorists are justifiably very suspicious of systems, such as those that characterized Structuralism, that posed a small number of binary distinctions, because they necessarily result in gross oversimplifications of the phenomena. But the problem here is not the binary distinctions, but the small number of them. In a system capable of making 100,000 binary distinctions, such as language or such as a formal logic with 100,000 predicates, you can make more distinctions than you could ever use, or even imagine. Probabilistic and fuzzy logics are not different in kind from ordinary logics. They are merely a way of avoiding deeper analysis of a situation, in terms of binary distinctions on a large number of very fine-grained predicates.*

*Metaphor is also sometimes contrasted with formal or computational approaches; it is sometimes taken as a proof that literature is beyond the scope* *of standard cognitive science. But in fact quite a number of computational treatments of metaphor have been developed, of more or less sophistication. Their structure is all basically the same. A mapping is established between the theory of a source domain, generally very rich in relations and inferences, and the theory of a target domain, generally less rich. That mapping is then used to import new relations and inferences from the source theory to the target theory.*

*Another frequent argument against a cognitive treatment of literature is that literary works are "open", in Eco's sense. They never stop yielding interpretations. But far from being an argument against a cognitive approach, cognitive science can help explain what makes a text open. A literary work presents data about a character or situation, and the reader's task is to arrive at a theory that explains the data. But as I said before, the data usually underdetermines the theory. A multitude of theories are possible. This is especially true when the data we are presented with has, on the surface, a somewhat contradictory nature, as happens in the richest literary works that try to be true to the world's complexity, as for example what we learn in Shakespeare of Shylock or of Malvolio. In this case easy theories are not available, there is no clearly best theory, and new theories are always invited.*

There are two phenomena central to literature that cognitive science has said relatively little about so far: fiction and narrative. What is their nature and function? But I think we can at least sketch a plausible account now. Before doing this, however, I first have to sketch plausible accounts of emotion, imagining, and social interaction.

I think a plausible account of emotion goes something like this. Our higher reasoning processes are at least mammalian and very probably higher primate. Those functions that we share with lower animals, including those most essential to survival, are not handled exclusively or even primarily by the higher reasoning processes. Emotions are the means our brain uses to impel us to generally appropriate actions without extensive reflection, often in situations where there is no time to reflect or where reflection is likely to lead us to the wrong conclusions. Emotions are a kind of evolutionary vestige, a leftover reptilian cognition. If you want to know what it is like to be a lizard, imagine yourself in a moment of stark terror, then imagine yourself lying on the beach completely blissed out, then imagine yourself continually alternating between these two states. Those who glorify our emotions are in fact glorifying the rather complex cognitive elaborations we construct around our emotions. These elaborations are, of course, tremendously important for an understanding of both literary works and our response to literature.

Cognitive science has almost nothing to say about the subjective experience of emotions. But we can begin to say something about the cognitive elaborations of emotion, by saying something about the combinations of beliefs and goals that are associated with various emotional states. Thus, pleasure is associated with, among other things, a focused belief that one's goals will be satisfied. Fear with a belief that they won't.

Imagining is very much like reasoning about the things we believe except that the causal connections with perception and action do not hold. The ideas we entertain do not have to be consistent with what we have perceived and what we otherwise believe, and, except for rare and generally unfortunate instances, they do not influence action. Imagining has two roles, one very rational and one quite curious. The rational explanation of imagining is that it is a way of working out the solutions to problems before they occur. It is a kind of counterfactual reasoning: "If this were to occur, then ..." The curious role involves its connection with emotions. Why should it give us pleasure to imagine winning the lottery, when we know perfectly well our chances are slim? Pleasure is associated with *any* proposition we entertain whose content is that our goals will be satisfied, whether or not we believe the proposition. It is as though emotional responses were not hooked up with goal- and belief-states quite right.

It is possible that this function of imagination can be reduced to the first function, however. Insofar as the function of emotion is to impel us to generally appropriate actions without extensive reflection, often in situations in which there is no time to reflect, the emotional response to imagining can be seen as a part of the problem-solving process, a quick heuristic—reptilian cognition again. We imagine a situation and perhaps practice a response, and the emotional reaction mediates between the imagining and the response, simply because that's the way it works in real situations.

A paraphrase of Horace's view of the function of literature thus provides a summary of the two roles of imagination: We imagine things to instruct and delight ourselves.

Fiction is discourse, which is social interaction, so before addressing fiction, we have to address social interaction. But this is an area where again I think there has been a lot of confusion.

Every entity in the universe is embedded in an environment. The environment impinges on the entity and changes its structure in various ways. The entity in turn influences the environment in some small way. Indeed, the environment is constituted of a large number of such entities. There is no particular mystery here. This is the way things are at every level of organization, from the quantum on up. Viewed from a synchronic perspective, it may seem paradoxical for an entity to be deeply influenced by an environment that is constituted precisely of entities like itself. But at every level, we can see how the situation evolved through slightly less complicated entities interacting with slightly less complex environments. There is a paradox only where we cannot imagine incremental progress toward the current state, and I know of no such cases.

There is nothing special in this regard about human beings embedded in social organizations. The social organization impinges on the individual, for example, by being the source most of his or her beliefs. In turn, the individuals, through their interactions with each other, constitute the social organization. But the evolution of this situation is clear. We have numerous examples of simpler arrangements, in the human and animal worlds. Similarly, there is no particular mystery about thought taking place against a background of already existent thoughts. Our adult mental states are the product of decades of development and experience and three and a half billion years of evolution.

There has been a great deal of work in AI on how a society could be composed of computational agents. A society of agents, robots or humans, is consitituted by conventions, or mutual beliefs, that arise from communication, agreements, and copresence, among other things. A mutual belief that $P$ among a set of agents $S$ occurs when each of the agents in $S$ has a belief that $S$ mutually believes $P$. For robots, this could simply be a representation of the expression $mutually\text{-}believe(S, P)$. The agent would also have to have the proper associated axioms for the predicate *mutually-believe*, allowing it, for example, to conclude individual belief from mutual belief. (If a society of agents discovered by communicating their experiences to each other that there were large areas of coincidence in their beliefs, thereby creating large areas of mutual belief, one can see that "truth" would be a useful concept for them to have.)

Mutual imagining would be like mutual belief except that it bottoms out in imagining rather than belief. That is, a set $S$ of agents mutually imagines that $P$ when each of the agents in $S$ imagines $P$, and they each believe that they all imagine $P$, and they each believe that they all believe that they all imagine $P$, and so on. The origin of any instance of mutual imagining will be either an explicit agreement or an implicit agreement by virtue of conventions in the society of agents. The functions of mutual imagining parallel the functions of imagining for the individual agent—cooperative problem-solving and enjoying the pleasure of one another's company.

Fictional discourse is an invitation to mutual imagining, in which the author provides explicit propositions to be imagined and the audience makes what they take to be the necessary minimal changes to the set of mutual beliefs the fiction is created with respect to. The functions of fiction are the same as the functions of mutual imaginings. Novels can be likened to experiments. Situations that are more or less possible, but not actual, are set up in a carefully controlled framework, and the author and the readers can explore the consequences of these situations.

Two central questions concerning narrative are "What is narrative?" and "Why, among the various forms of discourse, does narrative have its peculiar power over us?"

The traditional view in AI of agents—robots or people—is that they are planning mechanisms. They have goals, perhaps including a single top-level goal of "I thrive", and they use their beliefs about what kinds of things tend to cause what other kinds of things to decompose these goals into subgoals, and the subgoals into further subgoals, until the process bottoms out in executable actions. As the agent works through the actions of its plan, it monitors the environment to check on the success of the plan. When the plan fails, the agent modifies the subsequent steps in the plan to achieve its goals in another way, and perhaps to repair the damage it has done.

*This view of people should not be foreign to postmodern intellectual fashion, which views people's actions as primarily driven by their interests.*

A narrative is a species of discourse in which an entity, usually a person, is viewed as just such a planning mechanism, attempting to achieve some goal, generally in the face of some obstacle and working out and working through the steps of a changing plan to achieve the goal. Since plans are constructed out of our beliefs of what causes and enables what, narrative presents a character, like us a planning mechanism, maneuvering among these causal connections, attempting with or without success to create a satisfactory outcome.

The peculiar power of narrative derives precisely from this. A narrative describes a planning mechanism planning its way toward a goal. We are planning mechanisms, continually planning our way toward goals. Thus, narrative presents us with situations and events precisely as we would experience them when we are most engaged with the world.

Much of what is most powerful in literature is a conjunction of the two categories—the fictional narrative. It is an author's invitation to the readers to a mutual imagining, to delight and instruct, by the creation of a possible world and possible characters striving toward goals, told in a way that corresponds directly to our own experience as we plan our way toward our goals in a world that denies us of so much of what we desire.

For me, the most interesting issue, and the most mysterious one—the most challenging—where cognitive science and literary theory can interact profitably is in what ought to be the fundamental definitional question in literary theory—what makes a literary work good? What makes a text literature?

The category of literature, and especially of the canon, is in bad repute these days in literary studies, and for good reason. The serious study of X always casts doubt on the commonsense notion of X. The boundaries are fuzzy, the motives for classifying something as X are suspect, and definitions are elusive. The field of linguistics has had precisely this problem with "language", and anthropology with "culture". And indeed biology with "life" and physics with "matter".

Nevertheless, there are two important questions that we need to answer: "What should I read?" and "What should we all read?" The first question is a subquestion of the more general question, "How should I spend my time?" The second question is a subquestion of the more general question, "What common experiences should we all have if we are going to engage in collaborative action together?"

In part, the answers to these questions, especially the second, have to be instrumental, or political. For example, if we are going to participate in a multicultural society, then we should have some awareness of the fine details of each other's experience. We need to see the common humanity that lies behind differing practices.

But part of the answer to these questions has to involve a notion of literary quality, a notion of one work being better, in some sense, than another. It is simply a fact that Shakespeare is better than Harold Robbins. This is part of the data to be explained. But what is the explanation? What sense can we make out out this in cognitive terms? What are the cognitive components of literary quality?

One aspect of the aesthetic response is something that the cognitive psychologist Tom Bever has proposed a characterization of. In his view, an aesthetically satisfying experience is one that "stimulates a conflict in perceptual representations, which is resolved by accessing another representation that allows the two conflicting ones to coexist." I would add that very often the conflict is resolved by tapping into a large, highly structured conceptual schema that is heavily charged emotionally. Bever has examined such successful cultural artifacts as the song "Happy Birthday" and the pattern "One shave and haircut, two bits." In his account of "Happy Birthday", for example, we cannot decide which of two possible keys the song is in until the ambiguity is resolved precisely on the mention of the person's name.

A fairly clear example of this phenomenon in poetry is found in the Middle English poem,

> Western wind, when wilt thou blow?
> The small rain down can rain.
> Christ, that my love were in my arms,
> And I in my bed again.

*The first two lines create an atmosphere of longing by expressing a desire that is beyond the control of the poet to satisfy, thereby indicating that the desires expressed in the second two lines are also beyond control. The first line expresses the instrumentality, the second the result.*

*The third and fourth lines each express other things that are commonly longed for.* The third line by itself creates an image of the couple embracing, standing up. The fourth line by itself creates an image of the poet lying in bed alone. It is when we try to put these two images together to form a coherent picture of the whole that we are forced to reinterpret them as the couple lying in bed together, making love.

There is certainly more to the concept of literary quality or good-ness than this particular aesthetic response, as powerful and pervasive as it may be. I don't know of any deeper or more extensive analysis of literary quality in cognitive science. But it is part of cognition and it is one of the defining characteristics of literature, so it ought to be an ideal problem for collaborative work between cognitive scientists and literary theorists.