AI Growing Up

The Changes and Opportunities

James F. Allen

■ Many people make many confusing claims about the aims and potential for success of work in AI. Much of this arises from a misunderstanding of the nature of work in the field. In this article, I examine the field and make some observations that I think would help alleviate some of the problems. I also argue that the field is at a major turning point, and that this will substantially change the work done and the way it is evaluated.

AI has always been a strange field. Where else could you find a field where people with no technical background feel completely comfortable making claims about viability and progress? We see articles in the popular press and books regularly appear telling us that AI is impossible, although it is not clear what the authors of these publications mean by that claim. Other sources tell us that AI is just around the corner or that it's already with us. Unlike fields such as biology or physics, apparently you don't need any technical expertise in order to evaluate what's going on in this field.

But such problems are not limited to the general public. Even within AI, the researchers themselves have sometimes misjudged the difficulty of problems and have oversold the prospects for short-term progress based on initial results. As a result, they have set themselves up for failure to meet those projections. Even more puzzling, they also downplay successes to the point where, if a project becomes successful, it almost defines itself out of the field. An excellent example is the recent success of the chess-playing program DEEP BLUE, which beat the world chess champion in 1997. Many AI researchers have spent some effort to distance themselves from this success, claiming that the chess program has no intelligence in it, and hence it is not AI. I think this is simply wrong and will spend some time trying to argue why.

So how can we explain this strange behavior? I can't give a complete answer but can point out some contributing causes. First, unlike most other sciences, everyone feels that they are familiar with the object of study in AI, namely, human intelligence. Because we as people are so complex, and so much of our behavior is subconscious and automatic, we have a hard time estimating the difficulty of tasks that we do. Consequently, we have a hard time accurately evaluating progress. For instance, to us, processes such as seeing, language understanding, and commonsense reasoning appear straightforward and obvious, while tasks such as doing mathematics or chess playing seem difficult. In contrast, it is the perceptual and commonsense reasoning tasks that are most difficult for machines to capture, whereas more formal problems like game playing are much more manageable. Thus, the generalizations that one might make by thinking of machines as people will tend to be highly inaccurate. For instance, one might think that because a machine can play excellent chess, it must be superintelligent, possibly more intelligent than any human. In order to defend against this "threat," people feel they have to put down work. In contrast, if machines become capable of some truly difficult task such as understanding natural language, people may very well take this in stride and focus on how "intelligent" the system appears to be given what it says.

This inability to fathom the true difficulties involved is not limited to the general public. As mentioned previously, researchers themselves can be equally wrong in their evaluations. In the researchers' defense, the field is very young. By analogy to a mature science such as physics, we are at a stage prior to the development of calculus and Newton's laws. Our youth should be evident from the fact that many in the first generation of AI researchers are still active. We still haven't produced a good, clear explanation of what the field is, what its goals are, and what methodologies are appropriate in research. I'm going to spend a fair amount of my time today on some of these thornier issues that define what the field is.

It might help to draw a parallel between the development of AI and the stages of human development. I should warn you, however, that this is a highly idealized version of how humans develop, and as you'll see, it's a little optimistic. In this naive model, there are three stages of development. First, we go through an infancy and a childhood, where we develop basic skills. There's a very strong emphasis on creative exploration, and we're governed more by emotion than by reason. Then we move into the second stage, adolescence, where we develop a sense of self as well as self-reliance and discipline. We acquire the skills and habits that are going to last us a lifetime. Finally, we attain adulthood, where we apply and further develop our skills with practice. We train future generations and ultimately attain a very deep understanding of life.

So, how does this apply to scientific disciplines? Let's consider how this model applies to the development of modern aviation based on fixed-wing aircraft. The infancy of aviation was a long period of time, several thousand years, where people experimented with balloons and artificial wings and hopping machines, and they jumped off cliffs and did many other things in an attempt to fly. Adolescence began with the development engines light enough to be able to lift themselves off the ground. These first prototypes also got the development of aircraft off the ground! Although the *Kitty Hawk* wasn't a viable airplane and stayed in the air only for a few seconds, it actually was the first time in fixed-wing aviation that there was an experiment that achieved the goal even partially. From then on, researchers could measure incremental progress, and a burst of work shortly led to the first viable airplanes. The field eventually moved on to adulthood with the development of the science of aeronautics and a deep understanding of flight that is used today for a wide range of applications.

So, at what stage is AI? Well, we've certainly been having a childhood, and in fact, we may still be in it. We can see the typical properties of childhood such as learning of basic capabilities. We've made tremendous progress in a wide range of areas over the last few decades, including theories of search, representation, pattern matching and classification, and learning and a host of applications, including robotics, natural language, and vision. In addition, we've certainly been governed more by emotion than rationality. We've been prone to making outrageous claims; we've tended to fall under the sway of the "popular kids on the block," with everybody rushing off and pursuing the latest popular ideas; and we've certainly succumbed to bullies. We've certainly experienced a childhood.

We've made tremendous progress in a wide range of areas over the last few decades, including theories of search, representation, pattern matching and classification, and learning and a host of applications, including robotics, natural language, and vision.

Equally obvious, we're far from adulthood. We don't yet have a good sense of self in that we have no consensus on a definition of the field, and we need to learn some responsibility and self-control in that we have no generally agreed-upon methodologies for AI research. So the critical question is whether we've moved into adolescence yet, and if so, how far have we developed? I would claim that we're right at the beginning of adolescence. We are at a similar transition point to the first flight in aviation. The field of aviation was changed dramatically by the development of working prototypes because for the first time, experimental work could be supported. Until then, there was no reasonable form of evaluation because no one could get off the ground. Consequently, it was hard for the field to progress. After the first flights, there was a tremendous surge of progress and development in aviation.

Given all this, I believe that we're at a similar transition point to the first flight because we are now able to construct simple working artifacts which then can be used to support experimental work. This is a critical event in the development of the field that I believe will revolutionize the way the field operates and the way it is perceived. Now, some people will claim that we have been building working systems for decades in AI, so I need to qualify what I mean by *working*. For much of AI's brief history, a system has been said to work if it could run on a set of predefined or "canned" scenarios. That's not what I mean by working. Rather, I say a system works if we can specify a domain that supports a range of tasks, and the system can handle randomly generated or selected tasks in that domain with a reasonable degree of success. The domain and tasks may be precisely specified, as in game playing where the rules completely define the possible behaviors and outcomes, or more abstractly defined, as in tasks to support problem solving, or only defined by naturally occurring behaviors, such as in visual interpretation or spoken language understanding. In all cases, the system works if it can perform its tasks successfully over a wide range of specific situations and problems within the domain.

Depending on the area, we have different degrees of sophistication in working systems. In well-circumscribed domains, such as game playing, we have high-performance working systems. Most newsworthy lately, of course, we now have a chess-playing program that can beat the world champion. Even in more general domains, we can see substantial progress. Expert systems are widely used in industry to do things such as controlling factories and diagnosing faults and problems in mechanical devices, although such applications are often not publicized as AI. We have robots crawling on Mars, and spacecraft that will be controlled using AI planning and agent technology. We have systems that can roughly classify the content of news articles and e-mail messages and conversational agents that can carry on simple conversations with people using speech. Some of these systems are quite primitive compared to what we would like, but they're all working by the definition I gave earlier. You can give them new scenarios and have them go through and actually do something. With the capability to evaluate, we can start the long process of incremental progress that is essential for the development of the science.

Given the progress mentioned previously, why does AI have such a bad name in certain circles, and why do many still think AI is impossible or that the accomplishments so far are not really AI work? A large part of the problem is that the goals of the field are misunderstood, not only by the layperson but, even more damaging, by the researchers themselves. As a result, it can seem that AI is pursuing a hopeless dream and that all these success stories are not really the product of AI research. It is critical that we reexamine the goals of the field and to attempt to define it in a way that is inclusive—a definition that embraces the diversity of work and does not exclude large sections of the field.

What Is AI?

I looked in the introductions of five introductory textbooks to AI and collected the definitions that were presented. I was surprised at the range of different definitions and the lack of an overall consensus. But one thing they all had in common was that all the definitions focused on what the *artificial* part was, but *intelligence* was left undefined. In fact, most definitions of AI used the word *intelligence* in its definition, leaving that term completely unexamined. I think this causes considerable social problems in the field and so will attempt to elaborate on the notion of intelligence and in what sense it can be artificial.

Let me start with definitions that should not be used to define the field of AI because they are not broad enough and encourage exclusion of work rather than inclusion. AI is not the science of building artificial people. It's not the science of understanding human intelligence. It's not even the science of trying to build artifacts that can imitate human behavior well enough to fool someone that the machine is human, as proposed in the famous Turing test. These are all fine motivations of individual researchers, but they don't define the field. Specifically, they are not inclusive enough to include the breadth of work in the field, and they encourage fragmentation.

These definitions also lead to behavior where success is downplayed because it falls short of these definitions of AI. To illustrate this, I want to give my favorite ways of putting down or excluding AI research. Most recently, I've heard variants of all of these in response to the success in chess playing. The first attack can be paraphrased as, "Well, the program only does one thing (say, play chess), and it's not aware of the larger context. Therefore it's not intelligent." The underlying assumption made here that intelligence means that a system has to duplicate a person, and since people are obviously aware of much larger context and do many things, anything that can't do that is not intelligent. Note that this is an absolute statement. Such people are not saying that the system is not very intelligent compared to human abilities; rather, they are claiming flat out that it is not intelligent. In fact, many go further and use this argument to claim that current chess-playing programs don't have anything to do with AI research. The second way is really a variant on this theme and can be paraphrased as "Well, people don't work that way, so this is not intelligent." A quite different reason, and my all-time favorite can be paraphrased as "Well, I understand how that system works, so it can't be intelligent." It seems that there has to be something mystical about intelligence that forces us to say that if we can understand the mechanism, then it lacks intelligence. If something is just a mechanism and deterministic, then it can't be intelligent. These three methods of excluding work from the field are all motivated by the definitions mentioned previously.

So, if that's what AI is not, what is AI? Let's start with this mystical word *intelligence* and try

In fact, most definitions of AI used the word intelligence *in its definition, leaving that term completely unexamined. I think this causes considerable social problems in the field and so will attempt to elaborate on the notion of intelligence and in what sense it can be artificial.*

to tease out a few properties of it. Then we'll deal with the *artificial* part and see if we can come up with a better definition of the field.

Readers familiar with work in knowledge representation or philosophy or linguistics will realize that we are unlikely to produce a precise definition of *intelligence*. They know that most words that describe properties or categories of things in English or any other natural language inherently lack precise definitions. For instance, consider trying to define the notion of a chair. We might try to define a chair by its physical characteristics; say, it must have some legs and a back and a place that you sit. But counterexamples are easy to find—there are chairs that don't have legs because they consist of solid blocks. In addition, you can find chairs that only have little backs or no back at all. And I'm sure someone has made a chair out of stone. There's a sliding scale between what a chair is and what is just a big stone. So it seems impossible to define the notion of a chair in terms of its physical characteristics. You might try to define a chair in terms of functional characteristics. We might want to define a chair as something that we sit on. Of course, this doesn't exclude large rocks that we sit on. But what about a stone that someone intends to use as a chair but has never yet been sat on? And if a chair seats more than one person, isn't it a sofa, not a chair? So does a chair stop being a chair if we can squeeze two people into it? Again, we have a continuum of functional definitions between chairs and sofas.

The only things I can think of that aren't able to have intelligence in some sense of the word are inert things like rocks. I have not been able to think of a possible scenario where someone could convince me that one rock's smarter than another one.

While there's no precise defining criteria, we're all perfectly comfortable with the notion of a chair. This is the kind of concept we deal with all the time, concepts that are given the term *natural kinds* by philosophers. Given that we can't define the notion of a chair precisely, we can't expect a precise definition of a more complex thing such as intelligence. But we can learn some properties of it and maybe develop some common intuitions about intelligence that will help us define AI.

One way to explore the notion of intelligence is to look at work by people who have studied it in depth, for instance, in the education literature. And what we see there is the suggestion that there really is no single form of intelligence. Rather there are many different forms of intelligence, or different intelligences, that people have. There's analytical intelligence, which consists of the skills that are measured by typical intelligence tests. While some people have taken that as the definition of intelligence, many feel that it captures only one particular skill that people may have. But there are many other aspects to intelligence or types of intelligence. There's creative intelligence, which reflects one's ability to be innovative and produce original ideas. There's communicative intelligence reflecting our communication skills, physical intelligence reflecting our agility and skill at sports as well as others. While theories differ on the number of types of intelligence, it is clear from this literature that there is no single notion of intelligence. Rather, the notion is being broken down into particular capabilities.

Next we can try a linguistic analysis of the term. The word *intelligence* is a nominalization of the adjective *intelligent*. It is an interesting fact that most properties described by adjectives are relative to the thing you're modifying. For example, even the quite concrete adjective *large* means something different when talking about a mouse and when talking about an elephant. The same thing happens with *intelligent*. We can argue about whether my dog's smarter than your dog by saying things like, "My dog knows five tricks; yours only knows three. My dog knows 20 words of English and can do all sorts of things, and yours doesn't know hardly anything!" While you may dispute the facts of the case, we generally agree on what properties we are talking about. But you see a very different argument if a cat owner is arguing with a dog owner about which pet is more intelligent. The dog person will say, "My dog does more tricks," and the cat owner replies, "Well, why would you want a pet to do tricks? My cat's much more independent." The argument rapidly changes form and becomes a debate about what intelligence is. And there's never a resolution to this because people are using different criteria when applying the notion to dogs than to cats. So, intelligence is a relative concept that may capture different properties depending on the thing it modifies.

So, now for some empirical work. Given intelligence is relative, what kind of things can be said to be intelligent? Clearly we apply the term to people, and earlier, I illustrated the case of applying the term to cats and dogs. In general, I think we are comfortable using the term for all mammals and probably other vertebrates. For instance, we could talk about one iguana being smarter than another iguana because of the way it interacts with its environment. Note that there's no magic property of intelligence that all these animals have; rather, the notion may change from species to species. But we are comfortable comparing the intelligence of animals of like species. Does the term apply beyond the animal kingdom. I am a little uncertain with bacteria. Initially, I thought not, but on second thought, I imagine some microbiologist could convince me that some bacteria are more intelligent than other bacteria. Again, though, note that the notion of intelligence as applied to bacteria may have nothing to do with intelligence as applied to humans.

What about machines? The advertising industry is absolutely positive that we have intelligent machines. We have intelligent cars, intelligent microwaves, all sorts of different machines have some notion of intelligence. And while I'd never depend on the advertising industry to define anything accurately, this does indicate that people generally feel comfortable applying the notion of intelligence to machines. Certain machines will do something better than other ones and might adapt better and, thus, are more intelligent at the functions they perform. While some AI researchers and philosophers might think it outrageous to claim that a thermostat can be intelligent, I doubt that most other people have difficulty with the concept. Clearly, one thermostat can be better at sensing and controlling the environment than another, and we are comfortable at characterizing this difference as a difference in intelligence.

The only things I can think of that aren't able to have intelligence in some sense of the word are inert things like rocks. I have not been able to think of a possible scenario where someone could convince me that one rock's smarter than another one.

So what can we learn from this exploration? We have developed a notion of species-specific, or more interestingly, task-specific intelligence, which is a measure of how well some entity can perform a set of tasks. Looking at the previous analysis, a prerequisite for something to be intelligent is that it has some way of sensing the environment and then selecting and performing actions. Note that everything discussed earlier except rocks can do this. To use a very popular term these days, these are all things that we might call *agents.*

So intelligence is a term applied to agents that characterizes the degree to which they are successful at performing certain tasks. The term is relative both to the agent and to the task you're talking about. Notice that this definition doesn't require, nor does it prohibit, a symbolic representation of the world, or acting like a human, or anything else. Note also that it means something can have intelligence even if we understand how it works.

Given some intuition about the notion of intelligence, we can now consider the artificial part of AI. In what sense is the intelligence we create in a machine artificial? Given the previous discussion, it seems that intelligence as applied to machines to compare different machines is real intelligence; namely, it compares how well they perceive the environment to perform certain tasks. I think that the artificial part of the phrase comes from the fact that we try to get machines to display intelligence that we would normally only apply to people. So it is this switching of senses that makes it artificial. To try an analogy, we might try to develop cat AI by training a dog to behave like a cat. We would then evaluate how good a model we would have for the dog using the measures we would usually use for cat intelligence.

Pulling this all together, here's my attempt at a one-sentence definition of AI: "AI is the science of making machines do tasks that humans can do or try to do." Note that I haven't used the term *intelligence* in the definition. Instead, I've used a slightly less mystical word *task*, which implies the requirement of acting in response to an environment to achieve goals. Note that this definition also is quite inclusive—you could argue in a sense that much of computer science and engineering could be included in this definition. For instance, adding numbers is a task that people do, and so you could argue that building a machine that can add numbers up is AI. And actually, I think that's probably right. I think the inventors of the first adding machines would have seen themselves as the precomputer-era AI pioneers. But in current times we understand very well how to build machines that add numbers, so it's not particularly an interesting issue with regard to intelligence, and the field focuses on the more complex things that people do.

Pulling this all together, here's my attempt at a one-sentence definition of AI: "AI is the science of making machines do tasks that humans can do or try to do."

Developing Good Work Habits

With a working definition of the field in hand, the next thing to worry about is developing good work habits, namely, our methodology and modes of research. To start this discussion, I want to describe some of the prime mistakes that we have made in the past. I'll try to do this without insulting specific people and would include myself as one of the offenders as well. After pointing out the mistakes, I want to turn it around and argue that we probably couldn't have done better previously, given what we knew at the time (a principle characteristic of the infancy of the field). But we can do better now.

So here are some mistakes. Each of these has a perfectly reasonable set of starting assumptions but a bad ending result.

The microworld mistake: In order to study a problem, you design a highly simplified microworld that captures part of the problem. You then develop and test systems that can do tasks in this microworld. This is all very fine as the start of the research project. There are two mistakes that follow. One involves generalizing from this initial work to make claims about the original problem without ever testing beyond the microworld. The other involves forgetting about the original problem completely and basing all future work on the simplified microworld.

The abstraction mistake: A related but different problem is the *abstraction mistake*. Here we identify a few properties of a real task and produce mathematical abstractions of these properties. Again, both of those steps are perfectly good as initial exploration. But then we work with the mathematical abstractions and never come back to the issues that came up in the original task. In some cases, new subfields of research have arisen based solely on this level of abstraction, and work becomes farther and farther removed from the original motivating problems.

The magic bullet mistake: The third class of mistakes covers a wide range of issues I call the *magic bullet mistakes*. These all use the same form of argument: “We just need to do this one thing *X*, and then intelligence is just going to happen. So we don’t have to worry about intelligence per se, we’re just going to work on *X*.” Again, typically *X* is a reasonable thing to work on in its own right; the problem is in the claims that are made that cannot be evaluated until the indefinite future. Here are some classic magic bullets:

> If we get the right learning algorithm, intelligence is just going to emerge. So all we need to do is study learning.
>
> If we work on basic mechanisms, such as motor control and sensing, intelligence will just emerge.
>
> If we encode enough commonsense facts about the world, intelligence is just going to emerge.

As I said before, all these are reasonable things to be doing, in context, but they’re not going to suddenly solve the problem of artificial intelligence. And they shouldn’t be sold as that.

From a distance, its easy to laugh at everyone who has made these mistakes, which includes most researchers in the field. But in fact, in our defense, remember we were in our childhood. We were allowed to make mistakes then; that’s part of the evolution of the science. We started off with no workable theories and learned a lot. And given the power of computers until recently, even if we had had great theories, we wouldn’t have been able to make them run.

But things have changed now, and we have the ability and opportunity to do better. The opportunity is that we now have a wide range of concepts, techniques, and tools to support intelligent behavior. And in addition to that, we have achieved some basic competence in supporting real-time perception and motor skills. Consider the power of some of the techniques that we have developed: game playing, where we have built the world chess champion; scheduling, where we have fast, heuristic scheduling algorithms that yield dramatic speedups over traditional methods; decision making, where we have expert systems as standard tools in many companies and products; and financial forecasting, where we don’t hear much about what people are doing, but Wall Street firms seem to hire AI researchers at a rapid rate.

On the perception side, robots with vision are revolutionizing manufacturing. In addition we have a robot crawling over Mars, and inexpensive robots are readily available that are on the retail market for a few hundred dollars. We have viable commercial speech-recognition systems now, supporting over 30,000-word continuous speech, with 95-percent accuracy after training on a single speaker. This is a very impressive performance compared to where we were a few years ago and good enough to support many applications.

Given that we have basic perceptual capabilities and we have a wide range of higher-level reasoning techniques, how are we to proceed? First, we must identify the particular intelligence we want to study. It doesn’t make sense to try to capture all forms of human intelligence simultaneously. Rather we must study more mundane and limited tasks (or intelligences). For example, we might want to build a house-cleaning robot or an assistant for logistics planning or a telephone-based automatic customer-service agent. Each one of these tasks defines a particular capability or task that could support long-range research efforts. For the selected task, we then define a set of telescoping restricted domains, all the way from the long-term dream down to some absolutely trivial.

The next step after defining the set of telescoping tasks is to define the rules of evaluation. Ideally, the evaluation measures should apply over the full range of tasks that are possible, from the trivial to the most complex. It is an added bonus if the evaluation criteria can also be applied to people doing the same task, for then we can compare system performance with human performance on the identical tasks.

Now we are ready to develop a model and system for the very simplest task domain. Since the initial task should be highly simplified, we should be able to accomplish this in a relatively short time period and establish some baseline of feasibility for this line of research. We then evaluate the model on the simple task domain. Once we have reasonable performance, we then evaluate the model on the task domain the next level of complexity up.

By focusing on a simple task first, we can ground the research and evaluate capabilities early on. Evaluating on the next level of complexity up would have a nice sobering effect on the claims that we would then make about how general our systems are. It would also identify the next set of critical problems for the research project. This approach still allows the use of microworlds, abstraction, or magic bullets for preliminary research, but it forces us to continually ground the research back to the task as we go along.

To reiterate the strategy, we study intelligence with respect to a task, and we simply turn real problems into manageable ones by restricting the task rather than the abilities of the machine.

An Example: Conversational Agents

Let me give a concrete example of this method for research based on our own work at the University of Rochester on conversational agents. We'll start with the dream and then refine that down into more and more simple domains. Then we'll establish the rules of evaluation that we're going to use throughout this, and consider what initial steps we've taken.

The dream is a fully conversational, multimedia system that converses with a human level of competence on arbitrary topics, essentially a machine you could just sit down and talk to. Examples in the popular media would include HAL in *2001: A Space Odyssey* and the computer in the *Star Trek* series. Obviously, we're not going to achieve this level of competence in the near future. But we can start refining it down, as shown in figure 1. We can, for instance, consider conversational agents that collaborate with people to solve concrete problem-solving tasks. This includes most forms of interaction that we would probably want a computer to engage in: diagnosis, design, planning, scheduling, information retrieval, command and control, and a wide range of other task-directed activities. This restricted task is still too general for immediate solution, so we refine it again to concentrate only on transportation planning and scheduling tasks.

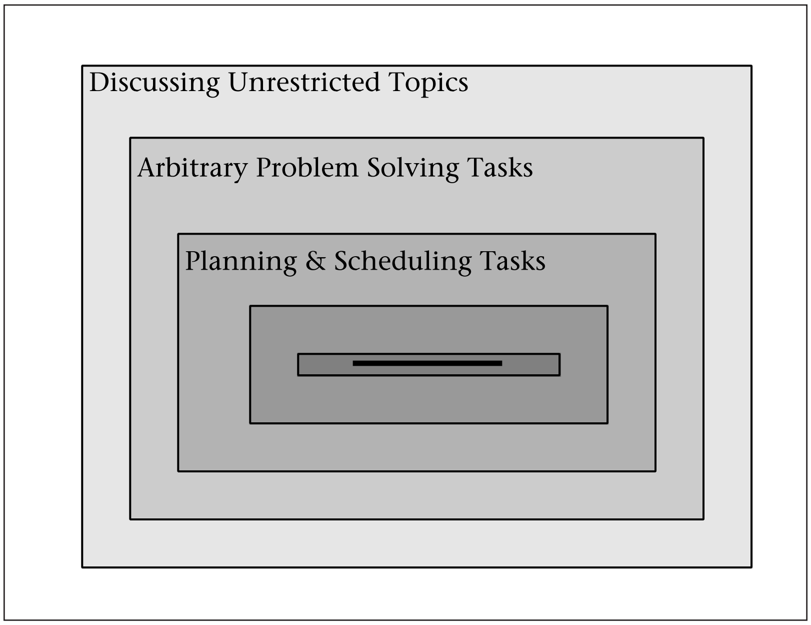

Discussing Unrestricted Topics
Arbitrary Problem Solving Tasks
Planning & Scheduling Tasks

Figure 1. Telescoping Tasks by Restricting Topics for Conversational Agents.

Once we get down to a fairly concrete level, we can start parameterizing domains along other dimensions. So, with planning and scheduling, for instance, we can parameterize the size of solutions required, the number of actions in a solution; the number of different action types that you need to reason about and their complexity, the complexity of temporal interactions in the domain, reasoning about quantities, static worlds versus dynamic worlds that change as you go along, and so on. There are many different parameters, and it is important to define the different dimensions of the domain so that we have a rich understanding of the problems that are there and those we're abstracting away as we simplify the tasks.

Note that this is a very different way of breaking up the problem than one might take in defining a conversational agent. We might, for instance, have tried to restrict tasks by limiting the vocabulary, the range of language that can be understood, the range of conversational acts handled, or the level of complexity we can handle in the discourse. The problem with simplifying in this way is that it does not simplify the problem for the person who must use the system. Rather it complicates things as they must learn a set of rules that restrict their natural behavior. So the computer's task becomes simpler, but the human's task becomes more complicated. Our approach is not to limit the person except by having them focus on solving the task. That way, the complexity of the task they have to solve naturally limits the complexity of the interactions that they typically attempt.

Since the goal is a conversational agent that collaborates with the user on planning and scheduling tasks, a natural evaluation criterion is how well the system enhances human performance on the task. This criterion can be applied in any of the different tasks, whatever level of complexity, and it provides a way to determine whether we're successful or not. In addition, we can use it to compare with human performance on the same tasks and, say, compare a person working with the system versus working alone, or compare it to two people working together.

Since the goal is a conversational agent that collaborates with the user on planning and scheduling tasks, a natural evaluation criterion is how well the system enhances human performance on the task.

In addition, because this project is studying conversational agents, an additional restriction is introduced that spoken language is a primary mode of communication. Note that this is an assumption not directly motivated by the evaluation criteria. If the goal were solely to improve human performance in planning and scheduling tasks, we might need to evaluate whether language is the best way to do it. By stipulating this restriction at the start, however, we focus the research. Additionally, we require that no restrictions be placed on what the user may say and that the system should function as well with novices or experts. As a consequence, we allow only minimal training, a few minutes at most, before a person is able to use the system.

Given this, in 1995 we explored the fundamental feasibility of this line of work. To do that, we defined the simplest domain we could come up with that still required extended interaction and constructed a system to handle it, TRAINS-95. The TRAINS-95 domain required a person to interact with the system to construct efficient routes for trains given a route map (see an example in figure 2). To complicate matters, in each session, problems such as bad weather were randomly generated in the scenario that only the system initially knew about. In addition, trains would be delayed if they attempted to move along the same track segments at the same time. Much of the effort in TRAINS-95 was aimed at developing techniques for the robust understanding of language in context in the face of speech-recognition errors, and we could get away with quite simple dialogue models. Readers interested in seeing a video of the session can obtain a copy from the web site at www.cs.rochester.edu/research/trains.

We had a few friends try TRAINS and found that they often could successfully solve problems using the system. In other words, it seemed to work, and we could generate good videotapes and example sessions for papers, and so on. In the past, this would have been enough to warrant claiming success. But given that we had achieved some basic competence, new questions arose that we couldn't ask before: How well does it work? What factors contribute to its success? We couldn't ask such questions before because we didn't have a conversational system that worked even badly (that is, drawing a metaphor back to flight, we'd never have gotten a conversational system off the ground).

To better answer how well it does work, we performed an experiment. We recruited some undergraduates, showed them the two- and one-half–minute videotape, and let them practice on the speech recognizer and then gave them cue cards with the problems to solve on it. We measured how long each session took, whether it ended in a working plan, and we measured the quality of the solution by considering how long the plan would take to execute. Seventy-three out of 80 sessions resulted in successful plans, which we considered to be an excellent success rate. However, it's also important to realize that TRAINS didn't really improve on a person doing the same task. In fact, what it really shows is that people can solve simple tasks under duress and eventually coach a reluctant system to get through and get to the solution. So we hadn't succeeded based on our prime evaluation criteria of improving human performance. But we did establish the basic feasibility of the approach.

Having a complete system running also allowed us to do other experiments that can lead to incremental improvements as well as explore other interesting issues. For example, we were interested in how well the robust language processing worked. To explore this, we compared using the system with speech input, which had about a 30-percent error rate in this experiment and with keyboard input, which had about a 3- to 4-percent error rate due to typos. Interestingly enough, even with the high error rate, the speech-driven sessions took only 66 percent of the time of the keyboard while producing the same quality of plans. And when given the choice, 12 out of 16 people chose speech to interact with the system rather than with keyboard. I'm not saying that speech is better than the keyboard in general for human-computer interaction. We would need to do much more elaborate experiments with alternative interfaces to establish something like that. But it is true that in the TRAINS task, robustness of the language processing allowed speech to be a viable, in fact, a preferred mode, of interaction.

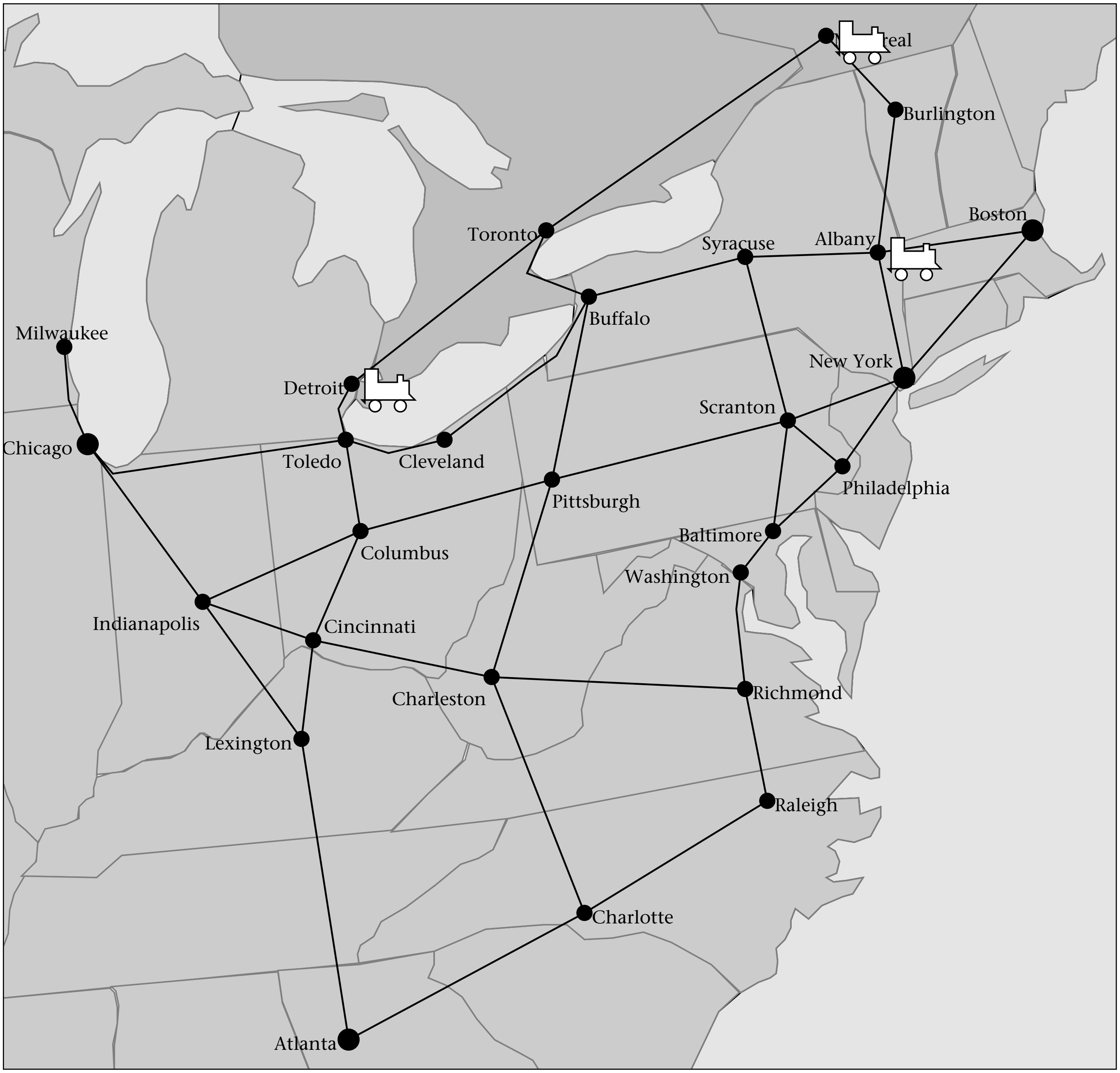
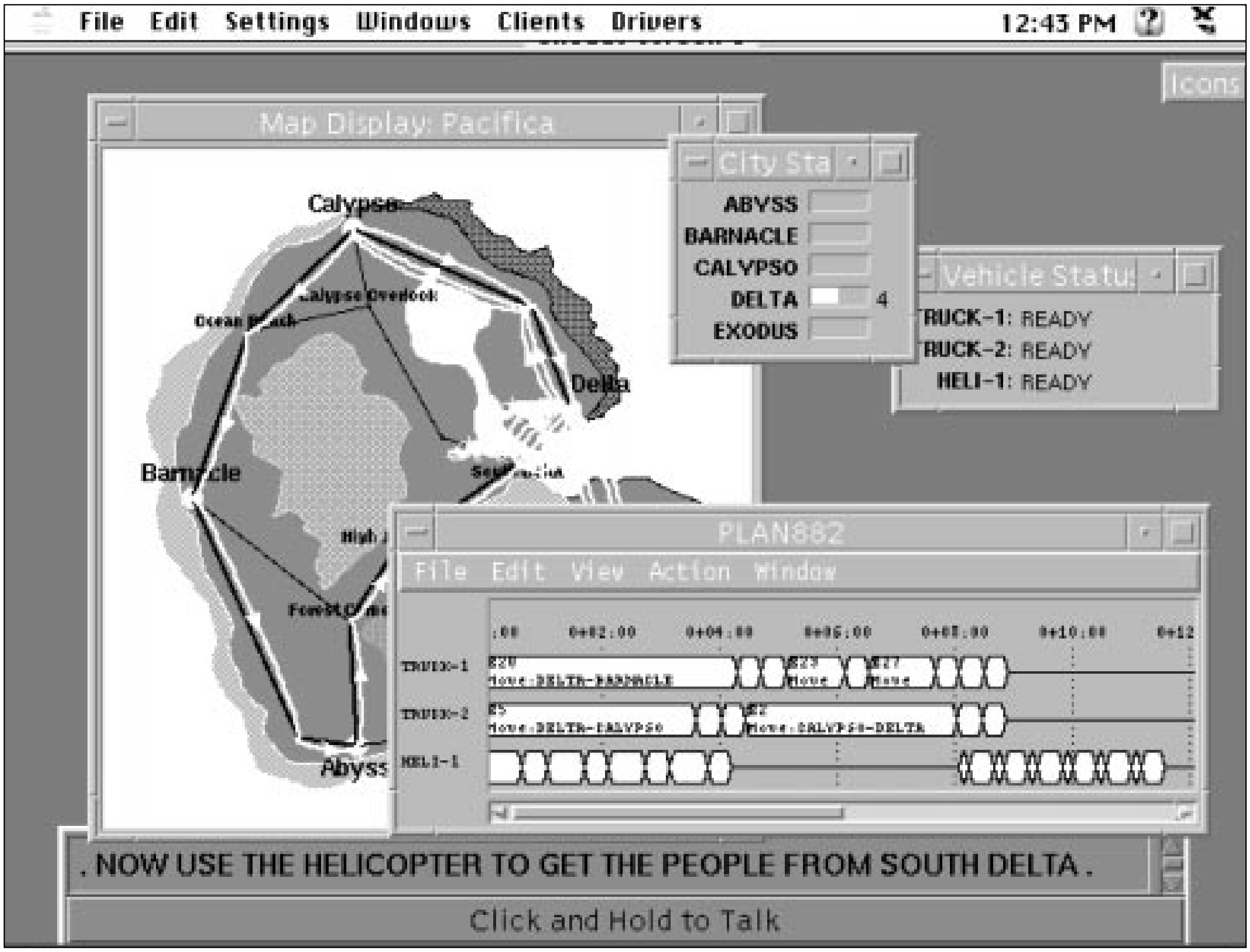


Figure 2. A Map Used in the TRAINS System.

With basic feasibility established on the base-level task, we were ready to move on to the next stage of the research. Our goal was to define a task that was still simple but one where we hoped that we could actually show that a person with the system can improve on human performance; that is, it has got just enough complexity. The result was a new task of moving cargo and personnel around from different locations under time and cost constraints, with multiple modes of transportation. There's increased opportunity for complex interactions because you can move things, drop them off somewhere, and have something else pick them up. Options can be compared based on cost and time. In addition, the world might change during planning, so we might have a plan almost constructed and then new information comes in that changes the situation. The need to handle a more complex domain forced us to handle much richer language phenomena that arose because of the more complex task.

We have just completed the first experimen-



ABYSS	
BARNACLE	
CALYPSO	
DELTA	4
EXODUS	

RUCK-1: READY
RUCK-2: READY
HELI-1: READY

. NOW USE THE HELICOPTER TO GET THE PEOPLE FROM SOUTH DELTA .

Figure 3. Developing a Plan with the TRIPS *System.*

tal version of TRIPS (the Rochester interactive planning system). A typical problem in TRIPS involves an island, and the task is to evacuate all the people off the island using some available vehicles before a storm hits. By varying the number of people, the types of transports available, and the time before the storm hits, we can generate a wide range of problems of varying difficulty. Figure 3 shows a screen shot from a session in which a plan is partially developed. Currently, TRIPS runs reliably only on scripted sessions where we have planned in advance the interactions that will occur. However, everything the system does is real. After evaluating TRIPS, we will then continue this process, improving the system, generalizing the tasks and domains, and so on. Along the way, new research problems arise that we would not have thought of otherwise, and the evolving system serves as an excellent test bed for supporting a wide range of research in collaborative planning, interactive dialogue, and spoken-language understanding.

Questions

There are a few questions that I believe that people will ask about what I've been saying.

Isn't Building Systems an Awful Lot of Work?

The answer is, yes, but so is building bridges. AI won't develop without systems, just as civil engineering would not exist unless someone built bridges. This doesn't mean that everybody has to build systems. In civil engineering, since bridge construction is well understood, the field can progress using controlled experimentation and theoretical analysis. But it's still ultimately grounded back to actual bridge construction. And that's where everything ultimately gets tested.

Are You Saying That the Last 30 Years of Effort Have Been Wasted in AI?

Absolutely not! In fact, if you look at the TRAINS and TRIPS systems, they build on a long history of AI research, ranging from the seventies, with ideas in chart parsing, STRIPS-style planning, hierarchical planning, and constraint satisfaction; the eighties, with unification-based grammars, incremental semantic interpretation, hierarchical discourse models, temporal reasoning, and plan-recognition algorithms; and the nineties, with statistical language models, robust parsing, and temporally based planning. And there are many others. If this work hadn't been done, we would not have the system today. So this really is an incremental development rather than a radical change from what we've done so far.

Well, System Building Is Applied Work. What I'm Interested in Is the Science, the Theory of AI.

System building is the experimental foundation on which the science of AI is based. If we don't have the experimentation, we don't have the science.

Well, How Can Work Progress, Then, Except for the Few Well-Funded Institutions?

Not everybody has to build full systems. But I would point out that if you work in a subarea of AI where nobody's building systems to test theories that are being developed, you might consider what your foundations are. But once you have the grounding in some experimental systems, then each subarea can define other relevant methods of evaluation.

And, for example, in the natural language field, there are many different methods now that allow people to do experimentation without building complete systems. There are large, shared databases of annotated corpora, annotated with features that have a communitywide agreement about what is crucial for the language-understanding task. There's agreement on some subproblems that are crucial to the entire process, such as extracting out sentential structure. And while it would be nice to see more, there is a beginning of sharing of system components, such as speech recognizers, parsers, and other components, where people can actually build systems using other people's work. For instance, in the TRAINS and TRIPS systems, the speech-recognition technology was built at Carnegie Mellon. If we'd had to build that ourselves, we would never have gotten it done.

Parting Thoughts

AI is a young field and faces many complexities because of the intimate relationship between the object of study (namely, intelligent behavior) and people's intuitions about their own intelligence. Despite a growing record of success in the field, many discount this progress by defining the successful work out of the field. Much of this is a result of the lack of an inclusive definition of the field. I have proposed a new definition and have argued that we are at a critical point in the development of the field because we can now construct simple prototype working systems. This new capability opens the door to new methodologies for evaluating work in the field, which I predict will lead to an accelerating rate of progress and development.

Acknowledgments

This article is an edited transcript of my keynote address given in Providence, Rhode Island, at the Fourteenth National Conference on Artificial Intelligence in 1997. I want to thank everyone in the Computer Science Department at the University of Rochester over the years for providing a supportive environment that has encouraged and nurtured my work. The work at Rochester reported here was done by the many dedicated students and researchers in the TRAINS project. References to their work can be found on the TRAINS web page, www.cs.rochester.edu/research/trains. Thanks to Henry Kautz and George Ferguson for helpful comments on a draft of this article. I also want to thank the Defense Advanced Research Projects Agency, the National Science Foundation, the Office of Naval Research, and Rome Laboratories for supporting my work fairly regularly over the years.

James Allen is the John Dessauer Professor of Computer Science at the University of Rochester. He received his Ph.D. in computer science at the University of Toronto and was a recipient of the Presidential Young Investigator Award from the National Science Foundation in 1984. A fellow of the American Association for Artificial Intelligence, he was editor in chief of *Computational Linguistics* from 1983 to 1993. He is also the author of several books, including *Natural Language Understanding,* Second Edition (Benjamin Cummings, 1995). His research interests concern the connection between natural language and commonsense reasoning, including work on spoken dialogue, language understanding, knowledge representation, planning, and reasoning. His e-mail address is james@cs.rochester.edu.