Advances in Cognitive Systems 1 (2012) 15–21 Submitted 2/2012; published 7/2012

Beyond Idiot-Savant AI

Scott E. Fahlman SEF@CS.CMU.EDU
Language Technologies Institute, Carnegie Mellon University, Pittsburgh PA 15213 USA

Abstract

Why has progress toward broad, flexible, human-like AI been so slow? I suggest the key reason is that, in recent years, few researchers have actually focused on this goal. Instead they have focused on achieving super-human (but brittle) performance in a few narrow problem domains. That is a valuable enterprise, but some attention should also be paid to the original – and still unachieved – goal stated decades ago by the founders of the AI field.

1. Slow Progress Toward the Original Goal of AI

If measured by the number of useful applications, tools, and vibrant spin-off fields it has produced, artificial intelligence has been a spectacular success. However, many people (including me) believe that AI has been a disappointment in terms of achieving its original goal: to understand and, ultimately, to replicate the computational mechanisms responsible for human-like intelligence, in all its generality, flexibility, and resilience.[1]

Back in the early days of the field, we seemed to be making good progress toward this goal. There were a number of key discoveries along the way: first, that computers could manipulate symbols as well as numbers; second, that search through a space of possibilities, with occasional backtracking, was a powerful and resilient way to solve many problems; third, that human-like performance requires a lot of knowledge, not just search power; fourth, that it is too tedious to assemble and organize by hand enough knowledge for broad, general intelligence, so we also must find ways to increase the store of knowledge by learning. But somehow, since the mid-1980's, progress toward this central goal of AI seems to have run out of steam.

Why has this happened? One explanation is that the funding climate changed. In the early days, there was steady, long-term support for research on the central problems of AI – not an enormous amount of funding, but enough to let a small community of AI researchers focus on the most challenging fundamental problems. This effort attracted some of the most brilliant minds in the field of computing. But times changed. Sponsors lost patience with basic, long-term AI research; they began demanding work on specific applications, with frequent benchmarks, competitions, "go/no-go" decisions, and short-term deliverables. The patient, curiosity-driven funding that characterized the early days of AI has now become very rare.

[1] Elsewhere (Fahlman, 2008) I have listed some of the major elements of intelligence that we still do not understand after more than 50 years of work on AI.

But that is only a part of the story. I think that we face a more fundamental problem: in an odd way, AI has been a victim of its own success. AI researchers have been successful in producing super-human performance in a number of narrow, disconnected problem domains, and the resulting excitement has almost completely crowded out work on the original goal of creating flexible, integrated, human-like AI. We have seen one gold rush after another to exploit new, highly specialized technologies with their roots in AI. In the short run, this may be good for the field, since it attracts both people and money; in the long run, I think it poses a serious problem.

2. "Idiot-Savant" AI

There are many examples of narrow super-human AI, but the story is similar for each. First, researchers grappling with some important AI problem try a variety of approaches, inspired to some degree by the questions "How do humans perform this task?" or "What is really required to achieve human-like performance?" Then someone comes up with an elegant mathematical approach that, *under certain conditions* and *with sufficient computing power*, can produce results much better than an unaided human. In many cases, this leads to a commercially valuable technology. In some cases, it gives rise to an active field of investigation that takes on a life of its own, attracting many researchers, substantial funding, and spawning its own specialized conferences and journals.

There are many examples of these super-human AI technologies: computer algebra systems that can solve integrals that no unassisted human can handle; search-intensive chess programs that can consistently beat (almost) every human player; search engines that can browse and index the entire Internet, but without any understanding of the content; statistical machine translation systems that can produce useful (if imperfect) translations without ever considering the meaning of the text; statistical data-mining programs that can extract subtle regularities from a mountain of noisy data; poker-playing programs that can achieve expert-level results using powerful techniques from statistics and game theory; planning systems that, for some types of problems, can produce optimal or provably near-optimal results; theorem-proving inference systems that (for the problems that they can solve at all) guarantee soundness, logical completeness, and provable consistency; and statistical inference systems that (if their models and input probabilities are correct) can very precisely infer the probabilities of various outcomes in a way that no unaided human can match.

These developments are impressive, but in every case (so far, at least) they have contributed little or nothing toward achieving our original goal. The super-human techniques apply only to a narrow set of problems, or the assumptions underlying the mathematical model are unrealistic in practice, or the method is too computationally demanding for large problems – often problems that we humans can solve easily using our less formal approaches. Or all of the above. Each of these systems is impressive, and many are commercially valuable, but none would be called *intelligent*, in the normal sense of that word. None of these systems can begin to match the common sense, flexible problem-solving ability, or language skills of a five-year-old child.

The result of this focus on super-human performance is what we might call "idiot-savant AI".[2] Present-day AI systems can perform some amazing (and sometimes useful) feats of

[2] The reference here is really to the popular stereotype of the "idiot savant", for example as seen in the movie "Rain Man": that is, an individual with some very startling capabilities in areas such as counting or memorizing details, but one who is unable, without help, to meet the challenges of everyday life. This greatly over-simplifies the reality of what medical professionals now call "Savant Syndrome", an umbrella

intellect that go far beyond the capabilities of typical humans, but each such system does only one or a few things. They are not capable of solving any problems beyond their own narrow domain, or of blending these isolated capabilities into a system with even a normal level of general intelligence, let alone super-human intelligence. They depend on their human companions to identify and formulate the problems and then to apply the results in the real world.

Real human intelligence is a bundle of many capabilities that can work together in a flexible way to solve the problems of daily life – and of many professions such as medicine, architecture, or disaster-relief planning. Some specific super-human skills would be valuable in these areas, but only if they are embedded in a resilient matrix of general, common-sense capabilities.

3. An Example: Optimal vs. "Good Enough" Planning

To understand what is going on here, let us examine the evolution of one area – AI planning and problem-solving systems – in more detail. A lot of the early work in this area took an intuitive approach, informed to some degree by introspection about how humans approach complex planning tasks.

The first problem was to represent the universe in which the planning takes place, the allowable set of operations, and the preconditions and effects of each operation. (We still have not completely solved these representation problems, but that is a topic for another essay.) Given an adequate representation, the next problem was how to find a legal path from the current state to the goal. Sometimes a legal path is easily found; sometimes it requires a great deal of search and non-obvious application of the available operators, or even the invention of new operators.

Ideally, we would like both a reasonably efficient plan and a reasonably efficient planning process. One powerful idea is hierarchical planning: first, use high-level, abstract operators to sketch the outlines of a plan; then use more specific operators to fill in the details. Another powerful idea is to save a sequence of operations that is useful in one context, generalize it a bit, and to turn the sequence into a "macro-operator" that can be used in other problems.

These ideas were explored extensively in the early days of AI in the context of systems such as GPS (Newell, Shaw, & Simon 1959), STRIPS (Fikes & Nilsson, 1971), ABSTRIPS (Sacerdoti, 1974), SOAR (Laird, Rosenbloom, & Newell, 1987), and many others. My own BUILD program (Fahlman, 1974) was typical of early work in this area. This system attempted to generate a plan by which a (simulated) one-handed robot could build a specified structure on a table, given a collection of blocks. BUILD could be quite resourceful: it would first try a straightforward approach, placing the blocks one by one, starting from the bottom of the desired structure and working upward. But if the desired structure was unstable during the construction, it would consider more complex plans. The system would try to use other blocks as scaffolding or as temporary counterweights. If that did not work, it would try to build a substructure on the table and then lift the whole subassembly into place. BUILD would do some extra work to produce good plans – for example, it would eliminate redundant steps – but its plans were by no means optimal, and were never intended to be. The system just returned the first reasonably good plan that worked. In that respect, it seemed very human-like in its planning.

term that covers many different combinations of talents and cognitive deficits. Most are not "idiots" in the sense of being severely developmentally delayed. Some are autistic, some have other impairments, and a few have savant-like gifts without any obvious deficit at all.

Not long after BUILD was published, the AI planning field changed radically. Methods were developed that, for a certain limited class of problems, guaranteed either optimal results or results that were provably close to optimal. Other things being equal, that was good: who would not prefer an optimal solution over one that is merely “good enough”? Of course, other things were not equal. The optimal planning programs were computationally demanding because the programs either had to consider *every* possible solution or formally exclude some parts of the search space where no optimal solution could possibly be hiding. For many problems of interest, these techniques were computationally intractable, or at least impossibly inefficient, so that they were limited to small problems in very clean, easy-to-model domains. In the real world, it makes little sense to waste a supercomputer’s time seeking an optimal solution to some real-world problem when a single obstacle – one not represented in the model – could force the entire planning process to be re-run. (If you really care about optimality, then a quick local patch to the plan is not good enough.)

Given these limitations, some researchers felt that the obvious move would be to continue work on flexible, resourceful, trainable, “good enough” planning systems. After all, humans do not worry about optimal planning in our daily lives. “Good enough” planning is good enough for us, and we can show great cleverness and resiliency when things go wrong at execution time – as they so often do – forcing us to replan on the fly. We can even pass partially instantiated plans from one person to another in the form of informal high-level recipes: “To get from CMU to the airport by car, take Fifth Avenue to the Parkway East (heading west), cross the Fort Pitt Bridge, and just follow the ‘Airport’ signs from there.”

But the idea of optimal or near-optimal solutions, built on a sound and elegant mathematical foundation of theorems and lemmas, was too alluring to pass up. Since the mid-1980’s, the planning field has been dominated by this approach. Many of the papers at planning conferences in recent years have focused on how to deal with the resulting intractability, so that at least some problems of practical interest can be addressed. If an optimal solution is infeasible, one must at least prove something about how close a technique comes to the optimum – impossible in most messy-real-world planning domains. It is now difficult to publish planning results that do not address optimality concerns, and several generations of students have learned to take this optimal-planning approach for granted. Not only has a super-human sub-field of AI been spawned, but work on more human-like approaches to planning has shriveled, unable to thrive in the shade of this mighty oak.

4. Another Example: Theorem-Proving vs. “Common Sense” Inference

Another example can be seen in the area of knowledge representation and reasoning – the area in which I currently work. In the early days (the 1960s through the 1980s), this field was chaotic but seemed to be making exciting progress. It was clear to everyone that the ability to store and effectively use a large collection of knowledge was the key to common-sense reasoning, natural language understanding, and much else. People attacked the problem with a collection of ideas, including semantic networks, production rules, frame systems, and societies of specialized agents. But in recent years, one mathematically clean approach has become dominant. For many people, the obvious way to represent knowledge is to use first-order logic or some close relative. This approach has appealed to great thinkers since Aristotle and, after 2300 years of development, the mathematical foundations of this approach are very well understood.

A knowledge base requires some sort of inference capability, and to many the obvious candidate is some logically complete proof method such as resolution theorem proving (Robinson, 1965), which offers a guarantee (and a requirement) that the knowledge base is internally consistent. Unfortunately, even the full first-order logic is undecidable. One can devise more restrictive (i.e. less expressive) logics that are decidable, and restricting them even farther can yield a system that is computationally tractable. But in such systems – OWL Lite is an example – even disjunctions are not allowed. One cannot say that a cow must be either brown, black, or white. That is a serious restriction.

These systems have their uses. There are times when you want to be absolutely sure that your conclusions are correct – that is, if the assertions or axioms are all correct at the outset. Theorem proving was invented for a good reason, and it stands as one of the crowning achievements of human philosophy and mathematics. And, in practical terms, these systems are being used: they are quite popular in the "semantic web" community, so they must be meeting the needs and aspirations of that community, at least for now.

But if we want to do simple common-sense reasoning, or if we want to represent the content of stories in natural language, we need a representation that is *more* expressive than first-order logic, not *less* expressive. We must be able to represent and reason about statements such as "John believes that Fred loves Mary, but that is not really true." We need default reasoning with exceptions. And, I would argue (see Fahlman, 2011), we need some way to represent and reason about overlapping world models that share most of their information but that differ in some crucial details. All of these capabilities go beyond the standard first-order logic model. In addition, we must be able to reason over these structures efficiently, even as the knowledge base grows to tens of millions or hundreds of millions of entities and statements.

Unfortunately, we just cannot do those things in a system that insists on logically complete, provably consistent inference methods. So here we have another example of an "idiot-savant" system (some relative of first-order logic and a provably complete inference method) that rests on impeccable mathematical foundations and that delivers super-human levels of *certainty*, but that is unable to represent and reason about the plot of a children's story or a situation comedy on TV.

What is the alternative? In the 1980's there were many AI systems that avoided this problem simply by using limited, human-like reasoning instead of theorem proving: they would reason locally, follow chains of inference to a certain depth, find *most* of the inconsistencies and useful inferences, and then they would stop. The Scone system (Fahlman, 2012), which has been developed over the past few years at Carnegie Mellon, uses this general strategy.

This approach seems very much like human reasoning. We all harbor a few inconsistencies that have not yet been exposed and resolved, and we occasionally make an error in inference because we have not thought deeply enough. But these errors are small in number compared to the errors we make due in incomplete or incorrect knowledge, while the speed and representational power that we gain from this strategy is an important part of our so-called common sense.

5. The Long-Term Problem for AI

This, I think, is the problem: AI is one field with two very different sets of goals. It would be healthy for the field if these two approaches could coexist. One set of researchers could work on narrow, disconnected, super-human areas of AI and another set could work on the original core problem of broad, human-like intelligence. These efforts could reinforce one another, and some people would move back and forth between them in the course of a career. Unfortunately, it seldom works out that way, for two reasons.

First, when one of these super-human technologies takes off, it creates a "gold rush" that attracts talent and resources away from the broader core problems of AI. In recent years, it seems that 80 to 90 percent of the people at the big AI conferences are working on super-human AI problems and approaches, not on human-like AI. So it is little wonder that progress on the core problems has slowed down.

Second, researchers in some of these super-human areas often develop a contempt for the less elegant, less formal, human-like approaches in the same or neighboring areas: their own work is based on elegant mathematics and clean abstractions. Many of these researchers argue that their approach is scientific and *principled*, while those working on less formal approaches to human-like AI are just messing around. "That's the sort of thing we did in the old days, before we understood how to properly frame the problem. Anyone still messing with those *ad hoc* approaches must be doing so out of ignorance, unaware of all the amazing progress that has taken place in AI."

There has indeed been some amazing progress, and we should build on that whenever we can. But in most cases, *we are not really talking about the same area of research.* There is a place for optimal planning, but we also need to understand human-like "good-enough" planning, which is faster and more resilient and flexible. There is an important place for theorem proving, but we need something more quick and dirty if we want our systems to read and understand the daily newspaper or even to follow the plot of a simple children's story. The same argument holds for many other areas in which super-human AI has thrived.

Unless and until these super-human approaches can be extended to cover the kinds of large, messy, hard-to-formalize tasks that humans handle with such aplomb, we must keep working on these tasks by whatever scruffy means are necessary. Maybe some of these problems can be handled by techniques that will ultimately be formalized and wrapped in elegant theory, or maybe they are inherently messy, but any reasonable person must admit that AI still contains many challenging and important problems that do not fit into the elegant theoretical frameworks developed so far.

One might argue that these super-human techniques, however narrow and specialized, are *more* valuable than understanding and emulating human-like intelligence. After all, we already have plenty of humans, and we are producing more every day. Why not just focus on the areas where machines can *extend* human capabilities? There is some merit in that argument, but it would be a shame to let the scramble for narrow, super-human capabilities, rooted in elegant mathematical theory, completely crowd out the quest for human-like AI.

The goal of understanding and replicating human-like intelligence remains as one of the great intellectual challenges of mankind – one of the last great mysteries. This quest has proven to be more difficult than we thought it would be, and the solution is unlikely to rest on a foundation of clean, beautiful mathematics – at least, not completely. But that should not discourage us. If, along the way to understanding intelligence, we create some valuable technologies that provide super-human performance in specific narrow domains, that is a bonus. AI as a field may pause occasionally to exploit these new technologies, but we should not let them divert us from the ultimate goal. Better yet, we may be able to combine the results of these two approaches to get the best of both worlds: flexible, resourceful human-like AI systems with a "telepathic" link to an array of super-human tools, for the situations in which those tools are applicable.

Acknowledgments

I would like to thank Pat Langley for making a number of suggestions that improved this presentation.

References

Fahlman, S. E. (1974). A planning system for robot construction tasks. *Artificial Intelligence, 5*, 1–49.

Fahlman, S. E. (2008). AI: What's missing? *Knowledge Nuggets* blog. http://www.cs.cmu.edu/~nuggets.

Fahlman, S. E. (2011). Using Scone's multiple-context mechanism to emulate human-like reasoning. *Proceedings of the AAAI Fall Symposium on Advances in Cognitive Systems*. Arlington, VA: AAAI Press.

Fahlman, S. E. (2012). The Scone knowledge-base project. http://www.cs.cmu.edu/~sef/scone.

Fikes, R. & Nilsson, N. J. (1971). STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence, 2*, 189–208.

Laird, J. E., Rosenbloom, P. S., &Newell, A. (1987). Soar: An architecture for general intelligence. *Artificial Intelligence, 33*, 1–64.

Newell, A., Shaw, J. C., & Simon, H. A. (1959). Report on a general problem-solving program. *Proceedings of the International Conference on Information Processing* (pp. 256–264). Paris: UNESCO House.

Robinson, J.A. (1965). A machine-oriented logic based on the resolution principle. *Communications of the ACM, 5*, 23–41.

Sacerdoti, E. D. (1974). Planning in a hierarchy of abstraction spaces. *Artificial Intelligence, 5*, 115–135.