Considerations Regarding Human-Level Artificial Intelligence

Nils J. Nilsson
(nilsson@cs.stanford.edu)
Robotics Laboratory
Department of Computer Science
Stanford University

January 2, 2002

AI researchers have several overlapping objectives. Among these are: to build systems that aid humans in intellectual tasks; to build agents that can function autonomously in circumscribed domains; to build a general science of intelligence as manifested in animals, humans, and machines; and to build versatile agents with human-level intelligence or beyond. In these notes, I list what I think are some important considerations for those working toward building human-level AI agents.

1. First, we might ask whether there is any particular methodology, among the many that exist, that stands out as being particularly relevant to human-level AI. Among the alternative methodologies, in no particular order, are those that stress the importance of:

    a. logical representation and reasoning
    b. neural networks
    c. probabilistic representations and inference
    d. object-oriented representations
    e. evolutionary computation
    f. modeling human and other animal cognition and behavior
    g. exhaustive or heuristic search methods
    h. phenomena that "emerge" from the interaction of relatively simple agents with each other and with complex environments
    i. "societies" of simple agents
    j. case-based representation and reasoning
    k. building a series of increasingly complex agents---following, more-or-less, either biological evolutionary history or human developmental maturation

    Some argue that human-level AI can best be achieved by modeling, at various levels, the processes thought to take place in animal and human brains. These arguments are countered by the observation that biological processes, arising as they did from evolutionary bricolage, may not be the only or even the optimal ones for artificial devices having human capabilities. The brain modelers reply that any so-called "sub-optimality" may stem from useful or even necessary computational compromises. I think it's too early to take sides in this argument, and that we should be open to good ideas that might arise from any of these methods. Humility about method befits an endeavor having such a long way to go.

2. The profound differences between "wetware" and "hardware" imply that there can be great differences between machine and human intelligence. Some AI researchers draw inappropriate conclusions from this fact. The simple version of their argument is something like, "computers should do what they are good at (fast computation, exhaustive search, detailed simulations), and humans should do what they are good at (intuition, judgment, creativity, …)". While the potential for differences exists, that does not necessarily imply that machine intelligence cannot, if we want it to, mimic and exceed human abilities in all the kinds of things humans are good at.

3. Hierarchical (and/or heterarchical) architectures will be important. One example is the so-called three-level robot architecture in which the lowest level deals with servo control of effectors, the intermediate level combines low-level actions into more complex routines, and the top level plans goal-achieving sequences of intermediate level actions. Each of these broad "levels" may themselves be more finely divided.

   At least at the lower levels, much of the action computation and control will be handled by a large number of (possibly competing) built-in mechanisms. Their responses may depend on state or otherwise be influenced by data read from various memory structures---in addition to depending on immediate sensory stimuli. Examples of some possible mechanisms include neural networks, finite-state machines, "subsumption" architectures, and "teleo-reactive" programs. Some of these action-computing mechanisms may be self-adjusting through experience and practice or otherwise teachable.

4. A key unsolved problem is developing the sensory and perceptual apparatus required for agents to function in complex environments. Understanding "what is out there" (or "what is being signaled from out there") will require a lot of world knowledge from within the agent about what *might* be out there (or what *might* have been signaled)---in addition to information obtained directly from the sensory input stream. We still don't know enough about what world knowledge might be needed, how it should be represented, whether or not it can be learned, and how it should be used. I believe that probabilistic reasoning and filtering will play important roles here because the sensory stream may be noisy and there will be uncertainties about the environment.

5. Some researchers argue that intelligent machines must be "embodied." They would claim that a human-level "computer in a box," like HAL 9000, is impossible. The extreme form of the argument has it that a human-level intelligent machine must be built in the form of a human (head, arms, torso, legs, etc.) While I think that human-level intelligence will require "sensory grounding," I don't believe that embodiment requires anything more than providing sensors and effectors appropriate to the tasks envisioned. Embodied or not, a computer-based agent could, of course, have the very programs that make it intelligent transferred to a disembodied computer.

6. There is a need for agents to have and to be able to use declarative knowledge in addition to the "procedural knowledge" which is encoded in special-purpose routines. Only those programs and circuits in which it is embedded can employ procedurally represented knowledge. Declaratively represented knowledge, on the other hand, can be used for a wide variety of more general purposes---some of which might be unforeseen when the knowledge is installed or otherwise obtained. Also, in order to interact with humans at a high level of understanding, agents will need to be able to respond appropriately to declarative statements and commands---such as "Rooms on the second floor can accept deliveries only on Tuesdays." Additionally, we want agents to be able to learn from books and other declaratively expressed material. I include many representational forms under the heading "declarative knowledge." Some examples are logical sentences, Bayes networks, geometric maps, graphical image models, episodic memory, and case studies. Recent work has emphasized object-oriented forms of both logical and probabilistic knowledge representation schemes. In fact, there has been progress toward unifying probabilistic and first-order representations.

   For versatile use of declarative knowledge, agents must be able to draw inferences from it. Since humans don't always reason in a logical manner but nevertheless often reason usefully for their purposes, non-logical inference methods might be useful for agents also. In any case, computational tractability probably will require compromising logical soundness and/or completeness. It is possible that the decision by an agent to begin to *reason* about how to act, rather than to allow the automatic evocation of a low-level action, should itself be governed by (adaptable) low-level mechanisms.

7. Some of the body of programs and knowledge required for an agent with human-level intelligence will need to be produced by three different methods. First, much of this body will need to be written by human programmers. Second, the agent will need to learn some of it on its own (and/or be taught it by non-programming humans). Third, the agent will need to synthesize some of its programs via some kind of planning method. Most likely no one of these three methods alone will be adequate.

   Furthermore, the three methods must be capable of operating in sequence. For example, human programmers might build a "first-cut" agent. Next, through self-learning or by being taught, this proto-agent might modify and/or add to some of its human-provided programs and representations. Subsequently, human programmers might correct, add to, and/or modify some of what has been learned or taught. And, at any stage, the agent itself might add to and/or modify some of the control programs through its own automatic planning methods. The planned programs might then be further modified either by humans or by learning methods, and so on. This three-pronged style of agent development places constraints on at least some of the kinds of representations and languages that can be used. Namely, they must be understandable, synthesizable, and modifiable by any of the three sources. The requirement to be able to modify a human-generated program by machine learning methods probably rules out the use of a programming language such as C. The requirement that humans be able to understand and modify a learned program probably rules out neural

networks. At the intermediate level, teleo-reactive programs and the models they access are possible candidates that appear to satisfy the constraints.

8. Research on human-level AI should avoid concentrating on the development of specialized "tools," such as programs for job-shop scheduling, network flow, and logistics planning. Important as they are for many applications, these tools can and will be developed by those parts of computer science most concerned with them. Taking as a clue the fact that unaided intelligent humans aren't particularly good at tasks such as complex scheduling but can use and learn to use scheduling tools, AI researchers should work toward building what I call *habile agents*---ones that are capable of using and/or learning to use tools. The internet makes a wide variety of such tools readily accessible to humans and to habile agents.

9. There is speculation about whether or not a human-level intelligent agent necessarily would have or should have "emotions." As we understand better the physical basis of our own emotions, it may be that some of them will be recognized as essential aspects of our problem-solving, communicating, and coping mechanisms. Some of what we will want to build in to human-level agents to enable them better to problem solve, communicate, and cope may be analogs of human emotional mechanisms.

10. Will human-level intelligent agents have "free will" or be "conscious"? My opinion is that if they have mechanisms that allow them to consider alternative courses of action and choose from among them based on anticipated consequences, they will have the same kind of free will that we have. Additionally, if they can introspect, name, and reason about these selection processes, they may even *claim* that they have free will---just as we do. Agents that can discuss these topics with us and make such claims might also declare they are "conscious," and I guess I would have to believe them---just as I believe similar claims from people.