Abstract

The intelligent use of tools is a general and important human competence that AI research has not yet examined in depth. Other fields have studied the topic, however, with results we can compile into a broad characterization of habile (tool-using) agents. In this paper we give an overview of research on the use of physical tools, using this information to motivate the development of artificial habile agents. Specifically, we describe how research goals and methods in animal cognition overlap with those in artificial intelligence. We argue that analysis of activities of tool-using agents offers an informative way to evaluate intelligence.

Introduction

The use of tools is a hallmark of intelligent behavior. It would be hard to describe modern human life without mentioning tools of one sort or another. For example, before writing this paragraph, I brushed my teeth with a tool, cooked and ate breakfast with tools, unlocked my office door with a tool, and jotted down notes with a tool. At my desk I am surrounded by tools that support cognitive as well as physical activities. A calendar aids my memory, a calculator improves my arithmetic, a whiteboard allows me to apply my visualization abilities, and of course my computer facilitates an even wider range of cognitive activities. In philosophical circles, some hold that tool use is central to intelligent behavior and that it should rival language in the study of cognitive phenomena (Preston 1998).

Tool use can be thought of as a specialized case of problem solving, but it can also be viewed in more general terms. Nilsson (1995) writes, “[I]ntelligent programs need to be able to find out about what knowledge and tools are available to match the problems they face and to learn how to use them,” arguing that systems with such capabilities may gain general human competence. This challenge has been taken up, in part, by researchers in robotics (Bluethmann et al. 2003; Stoytchev 2003) and intelligent user interfaces (St. Amant and Zettlemoyer 2000). It also drives the narrative in this paper; we believe that physical tool use can act as a precursor to cognitive tool use and thus more intelligent effective behavior.

Research on tool use

Accounts of tool use have appeared in the AI literature (e.g., (Agre and Horswill 1997; Bluethmann et al. 2003; Brady et al. 1984)), but the most extensive analyses are found elsewhere, in particular in the field of animal cognition. The most widely accepted definition of tool use, in any account, is that a tool is a material entity that is used to transform the environment in order to achieve a goal. This definition has a long history in ethology (e.g., Tinbergen 1952), and it has been applied to a wide range of animals, from insects to primates to humans. The overall goal of our research is to characterize the class of habile (tool-using) agents in computational terms. Broadly speaking, a sophisticated habile agent can reason about how to exploit objects and features in its environment to reach its goals. A physical habile agent can perform spatial and physical reasoning (both qualitative and quantitative) about objects in relation to the agent’s own location and dimensions. Most importantly, a habile agent can reason about the relationships between its capabilities, the problem at hand, and the tools that it can bring to bear in order to transform the problem into one that is more easily solved. While specialized physical habile agents already exist (e.g., some factory robots and most prominently Robonaut (Bluethmann et al. 2003)), the goal of our research is to develop agents with much more general physical tool-using abilities—such agents may not even explicitly incorporate the concept of tools, but can opportunistically exploit objects with appropriate properties.1

This paper gives an overview of research on the use of physical tools, using this information to motivate the development of artificial habile agents. In the following section, we outline research on natural tool use in fields outside AI, focusing in particular on work in animal cognition. We use our brief survey to motivate a set of desirable properties in an artificial habile agent. We then describe an approach to defining general tests of agent intelligence based on tool-using ability. We conclude with a discussion of how consideration of tool use may provide useful challenges for AI research.

1For contrast, in response to a request for a wrench, Robonaut first decides whether or not it has a wrench in its hand (Bluethmann et al. 2003, p. 193). For biological agents, however, the use of a tool fundamentally alters the agent’s evaluation of information from its sensors and of the capabilities of its actuators. Informally, tool users are aware of a tool in hand; a decision is not needed.

2Work on the evolution of human cognition and tool use is also relevant (Deacon 1998; Gibson and Ingold 1993; Mithin 1996;
Thus tool use is the external employment of an unattached environmental object to alter more efficiently the form, position or condition of another object, another organism, or the user itself when the user holds or carries the tool during or just prior to use and is responsible for the proper and effective orientation of the tool.

Although this definition may appear unwieldy, all of its conditions turn out to be needed to distinguish tool use in the animal kingdom from other activities. Three examples illustrate the insights we can gain from animal tool use.

Some of the simplest animals to use tools are wasps (Anomophila armata and A. Yarrowi) that pound earth down into a nest with the help of a pebble (Oswalt 1973). Thus activities recognizable as tool use can be found in agents with limited manipulation, no capacity for learning, and only the simplest processing capabilities. This indicates that with careful programming it is possible to build a range of tool-using behaviors into an artificial agent (Bluthmann et al. 2003).

In laboratory experiments, the tool-using capabilities of several species of primates have been tested using a transparent tube in which an item of food is placed, as shown in Figure 1a. The end of the tube is too narrow for the subject’s hand to enter, but tools such as reeds are made available for the subject to use. The results of experiments with capuchin monkeys (Cebus apella) are especially suggestive. Visalberghi and Limongelli (1996) observe that capuchin monkeys often succeed in this task by rapid-fire selection of different strategies, though trying out many inappropriate strategies along the way. For example, given a bundle of reeds too bulky to fit through the tube in aggregate, one monkey hit upon the step of unwrapping tape from around the reeds, but then tried to push through the tube with the tape. For AI researchers, this example is interesting less because of the success or failure of the monkeys but instead because tool use can be “forgotten” in use, suggesting that an experienced user of a tool concentrates on the task at hand and not the tool itself, whereas novice users may instead direct more of their attention to proper use of the tool (Baber 2003).

As other studies show, for humans as well as other animals, an internal body schema is not constrained to modeling morphological boundaries, but can be extended to incorporate objects attached to bodies, such as clothes, ornaments, and tools (Berlucchi and Aglioti 1997). For instance, when one becomes skilled with a hammer, the hammer feels like a part of the body. It has long been observed that tools can be “forgotten” in use, suggesting that an experienced user of a tool concentrates on the task at hand and not the tool itself, whereas novice users may instead direct more of their attention on proper use of the tool (Baber 2003).

This is just a small sampling of research on animal cognition and tool use; Beck (1980) and Tomasello and Call (1997) give excellent introductions to animal and primate tool use respectively. Our purpose in reviewing such results is to highlight the common concerns of animal cognition and AI researchers, and to suggest the benefits of cross-fertilization for evaluating the intelligence of artificial agents. In addition to the AI-centric observations made about animal cognition work above, there are more general...
areas of overlap.

- Animal cognition researchers are acutely aware of the pitfalls of wishful thinking. Much of Povinelli’s research, for example, is motivated by the possibility that some general problem-solving skills attributed to chimpanzees can be explained by simpler mechanisms (by analogy, in the same way that stimulus response agents are simpler than planning agents.) His experiments are designed to tease out and eliminate various potential explanations for sophisticated surface behavior.

- Animal cognition research is also careful in distinguishing different capabilities in non-human intelligence. For example, while chimpanzees are generally smarter than dogs, dogs outperform chimpanzees in some tasks that for humans require the mental capacity to track invisible displacement of objects (Tomasello and Call 1997). That is, human development is a useful guide for understanding the scaffolding necessary for human intelligence, but some level of intelligence can be reached by taking different paths.

- Animal cognition research has produced a wide range of tasks that are common across the field. Food tube experiments are one example. Working with such common evaluation environments is now commonplace in subfields of AI such as planning, machine learning, and message understanding. One lesson that AI can take away is the depth of human results that is less informative involves context. Most food, though more intelligent animals such as chimpanzees and elephants also use tools for grooming and self-defense. For artificial tool-using agents, we are interested in a more important aspects of tools in use. For example, the U.S. Patent Office describes wrenches as “engaging a work part and exerting or transmitting a twisting strain thereto, or means for imparting or transmitting an actuating force to such a tool” (as quoted in (Hohn 2005, p. 60).) Taxonomies of human tools for subsistence (e.g., spears and fishing implements) due to Oswalt (1973) come closer, but are too abstract. Baber gives an extensive overview of tool use and cognition from a human factors perspective (Baber 2003), but no general taxonomies of action. Perhaps most promising is a small taxonomy that describes software tools as metaphors for physical tools (St. Amant and Horton 2002):

  - **Effective** tools produce a persistent effect on materials or the environment, e.g., hammers, screwdrivers, and saws.
  - **Instruments** provide information about materials or the environment, e.g., calipers and magnifying glasses.
  - **Constraining** tools constrain or stabilize materials for the application of effective tools, e.g., clamps and straight edges.
  - **Demarcating** tools impose perceivable structure the environment or materials, e.g., a carpenter’s pencil or push-pins.

To shift perspective to the actions of the agent, the agent uses effective tools to transform its actions on the environment and uses instruments to transform its perception of the environment. The agent uses constraining tools to stabilize the environment for its effective actions and uses demarcating tools to add information to the environment in aid of perception, by creating external representations. These categories of action can guide the selection of tasks by which we can evaluate the competence of a given agent and judge whether its design is functionally adequate.

**A “tooling test”**

We can take the taxonomy given above to describe the tool-using behavior of a sophisticated autonomous agent. Consider an urban search and rescue robot some years from now:

In the course of exploring a ruined building, the robot stops in the entrance to a room, blocked by a gap in the flooring too wide for its wheels to cross. It retreats to the hallway, where it has previously seen loose panels of wall board. It pushes a panel over the gap and crosses into the room. The robot’s first target is a pile of rubble near the entryway. The pieces are heavy, but a nearby length of pipe serves as a good lever to lift some of the debris aside. The robot continues its search when nothing of interest is found. Because lighting is variable and navigation uncertain, the robot marks each region of the room it has examined with a splash of fluorescent paint; this makes it easy for the robot to tell what remains to be done, and it informs other searchers (human or robotic) that the room has been explored. On completing its inspection, the robot sprays a fine mist of particles into the air, looking for are air currents that may indicate otherwise hidden passages. No movement. The robot continues to another area of the building.

This description illustrates use of each of the classes of tools given in the previous section. While perhaps useful as a challenge problem, the scenario given above is far too specific (and too difficult given the current state of the art in
robustics) to guide research toward intelligent tool-using ability. Instead, we propose a multi-stage test for intelligence, in which each stage that is passed gives evidence for greater intelligence than needed in the previous stage. The test will involve the subject manipulating physical tools in order to accomplish specific tasks; obstacles will be set in the way that can be overcome by the use of these tools. We call this test the tooling test.\(^3\)

Just as with Turing’s imitation game (Turing 1950), we need to establish a few basic assumptions about the subject of the tooling test, whom for convenience we will refer to as T. First, we need to have some understanding of T’s physical capabilities. If T were a small child, for example, it would not be reasonable to give him a forty pound sledge hammer and the task of driving railroad spikes. If T is non-anthropomorphic in its physical or sensory capacities, we need to be even more careful about making inferences based on what T can do (sometimes described as the effectivities of an agent (Turvey and Shaw 1979).) Tools reflect the capacities and limitations of their users as much as they do the functions they are designed to facilitate. Second, we need to know something about the goals that T might plausibly have. These might involve a desire to gain food or other sustenance, desire for a reward of some type, or perhaps a desire to leave the experimental setup. This assumption is almost universal in animal cognition experiments, and even if T is, say, a mechanical robot, it will eventually be necessary for T to recharge itself. In our discussion, for convenience, we will use acquisition of an object as a generic placeholder for the actual goals that T may have. Third, as with the imitation game, we assume that T is basically cooperative, willing to go along with our experiments. We can draw no conclusions from a test subject that sits like a stone in the middle of a room.

Given these assumptions, we can define a sequence of incremental tests of T’s tool-using capabilities.

**Simple tool use.** In this stage, T must apply simple tools to reach its goal. These simple tools are physical extensions of T’s unaided effectivities. For example, the object to be acquired might be placed out of reach, but it can be retrieved by poking it with a pole. We refer to this as an example of simple tool use only partly because the tool itself is simple; more importantly, the behavior of the tool is transparent to the tool user. Other scenarios that entail simple tool use might involve placing the target object in a transparent container that can only be opened by using a lever to lift a heavy lid, or by using a hammer to force a peg out of position, or by using forceps to retrieve the object through a narrow opening. In each case, T is prevented from carrying out a task because of physical constraints, but an appropriate tool is available to overcome these constraints, and the mechanics of the tool-based activity are observable as the tool is used.

These sorts of tool-use tests are a staple in animal cognitive research, as described above. The classic monkey and bananas problem is probably the best known of these. Results have helped researchers to better understand the extent to which animals can reason about spatial relationships, kinematics and dynamics, qualitative physics, and so forth, in comparison with humans. By increasing the complexity of the tasks to be performed (we can imagine arbitrarily complicated Rube Goldberg devices that must be worked through) T has the opportunity to display more and more sophisticated tool-using behavior.

**Tool construction.** Construction of novel tools and refinement of existing tools is only rarely seen among animals, and on an evolutionary scale it is a relatively recent capacity for human beings. We can test T’s capabilities in this regard by providing it with tools suitable for making or modifying other tools. For an example of tool modification, suppose that the object acquisition task involves turning a bolt. T tries to turn the bolt with a wrench, but the wrench has too short a handle to provide sufficient leverage. A length of pipe is also nearby, which T uses as a sleeve to extend the wrench handle, to successfully complete the task. Examples of tool construction are comparable, working from raw materials rather than tool components. (Note that one of the earliest examples of tool construction, rock knapping, is a demanding skill for modern humans to acquire, but more sophisticated tool-making tools can ease such tasks.)

Surprisingly, chimpanzees and humans are not the only creatures capable of tool construction and modification. For example, Sterelny (2003) describes crows able to bend a piece of wire into a shape appropriate to retrieve food from a complicated apparatus. In the end, it is not clear that there is a conceptual difference between tool use directly on materials and tool use for making other tools, but this turns out to be a soft boundary between human and animal tool users.

**Non-transparent tool use.** A more difficult set of problems arises in scenarios in which the mechanisms of a tool, in its operation or its effect, is partially or wholly hidden. For example, imagine picking a simple lock, or working the levers of a complex machine with an opaque housing, or even fishing for a wire inside a wall by feeding an electrician’s snake through a small hole.

What we gain from posing such problems is the opportunity to observe T’s exploratory strategies. Search is often held to be a core aspect of intelligent behavior. Comparable tasks are used in studies of human development and animal cognition. For example, one experiment design involves hiding an object in a container, and then moving this container between a series of other containers in different locations, with the object deposited in one of those containers. An unsophisticated search strategy for the will examine containers in locations that have not been visited; this is a strategy followed by most primates (but, surprisingly, not domesticated dogs, as mentioned above.) Piaget (1955) has described solving such problems as Stage 6 behavior, associated with mental representations. In these cases, T must reason about what is happening out of sight; behavior that relies purely on observations, without inference, will be less
successful, or at least less efficient, than behavior that relies on some understanding of what can possibly be happening behind the scenes.

There are two extensions possible here. First, let’s suppose that the problem is so difficult that blind search will take T quite a long time before reaching success. Enter a collaborator, C. C carries out the appropriate action or sequence of actions to accomplish the task. In our tool-using examples above, C might move the lock pick into a specific sequence of configurations, or move specific levers in order, or move the electrician’s snake along a particular path. If T can solve comparable problems after having observed C’s action, we count this as an example of learning, a key facet of intelligence. Again, we can vary the internals of a given apparatus to test the extent to which T has generalized its observations, and we can make the learned procedure as complex as desired.

Second, we can introduce uncertainty to the problems. Suppose that the tool or environment responds only probabilistically to the correct manipulation. Does T learn despite failure? Such learning again indicates some level of intelligence.

Collaborative tool use. A sophisticated ability to solve problems using tools, either alone or after instruction, captures many of the characteristics we associate with intelligence: the ability to reason about causality, physics, spatial relationships; planning and learning; even creativity and opportunism. One aspect of intelligence that is missing, however, is the ability to communicate. This is addressed in a stage that involves our collaborator as a part of all tasks. C attempts to solve a task but cannot succeed, lacking some crucial physical aid or tool. For example, acquiring a target object might involve prying up a lid that is too heavy or large for one to lift alone, turning a nut on a bolt that rotates freely unless held in place, and so forth, T passes this test by observing C’s actions, recognizing C’s implicit goal, and contributing appropriately to reach the goal.

There are two variations of interest in these scenarios. In one, C works toward one of T’s goals; in the other, C pursues a goal to which T can be assumed to be indifferent. Both of these variations require T to effectively assume C’s perspective for the purpose of goal recognition, but the latter case is more difficult in that C’s goal is one that T would not ordinarily have.

Discussion

We can make the relationship between the Turing Test and the Tooling Test explicit as follows: in each of the above scenarios, we can substitute for T a robot remotely controlled by a human operator. Given sufficient practice, an experienced operator will be able to reach some level of competence in solving the problems posed, even given the limitations of the robot’s hardware and the communication link. An independent observer, or even the collaborator in the experiments, could then decide whether the robot were acting autonomously.

Viewed in this way, the tooling test is clearly derivative of the Turing Test and its variants. It has some similarities to Harnad’s Total Turing Test (Harnad 1992), in that it involves physical interaction with the environment, but it is more restrictive than the Turing Test. The activities required of the agent are well circumscribed: solve this test, and then solve another test. As described, although the test requires cooperation, it involves no explicit communication in the form of language.

Despite the restrictions on generality and the use language, the tooling test offers some advantages over the Turing Test. Based on Paul Cohen’s criteria for a good test of intelligence, we can say the following:

- The tooling test is a reasonable proxy function. For a fanciful example, if aliens descended on Earth in a working space ship, we would have strong evidence for their intelligence even if we were unable to communicate with them or even to observe them communicating with each other. Their mastery over their physical environment would be sufficient to indicate much broader competence.
- The tooling test is specific. Different instances of the test require the exercise of different types of intelligent behavior, as discussed in the previous section. Further, the tests are independent of culture. A conversant may fail the Turing Test by being unaware of current events, the emotional connotations of words, the subtleties of human relationships, and so forth, but this is hardly convincing evidence for lack of intelligence. While Turing expected machines successful in the imitation game to be reasoning the way that humans do, in these tests we are willing to accept completely unexpected solution paths as evidence of intelligence.
- Perhaps most importantly, the tooling test is diagnostic, in several ways.
  - Each trial has an explicit success criterion: has a goal been achieved? This obviates some traditional concerns about the validity of the Turing Test (e.g., Is ten minutes enough for a conversation? How sophisticated must the human conversant be?)
  - The test is extensible. Like the Turing Test, the tooling test can be extended to as great a sophistication as desired. The equivalent of a deep discussion of King Lear might be the construction and operation of a complex piece of equipment, based on exploration, observation, and learning.
  - The result of a sequence of tests is not a binary intelligent/non-intelligent judgment. Rather, just as in research on animal tool use, we can say that a tool-using agent is more or less intelligent, based on the sophistication of their actions.

- A simple version of the tooling test can be solved today, with current technology, as demonstrated in existing systems (Bluthmann et al. 2003).

The tooling test is subject to some of the same philosophical objections that the Turing Test has raised, most obviously
that it takes a behaviorist approach to the attribution of intelligence to an agent. We believe that this concern is alleviated by the use of comparable tests in animal cognition. Researchers have long since abandoned the notion that animals are automata, and assess different aspects of the intelligence of non-humans without qualifying their observations as being of the mere imitation or simulation of intelligence. Chimpanzees, elephants, and so forth have some measurable degree of intelligence; we can apply the same metrics to artificial agents.

Acknowledgments
This effort was partially supported by the National Science Foundation under award IIS-0083281. The information in this paper does not necessarily reflect the position or policies of the U.S. government, and no official endorsement should be inferred.

References