Is the Turing Test Still Relevant? A Plan for Developing the Cognitive Decathlon to Test Intelligent Embodied Behavior

Shane T. Mueller, Ph.D.
Klein Associates Division
ARA Inc.
Fairborn, OH 45324 *

Abstract

The field of artificial intelligence has long surpassed the notion of verbal intelligence envisioned by Turing (1950). Consequently, the Turing Test is primarily viewed as a philosopher's debate or a publicity stunt, and has little relevance to AI researchers. This paper describes the motivation and design of a set of behavioral tests called the Cognitive Decathlon, which were developed to be a useable version of an embodied Turing Test that is relevant and achievable by state-of-the-art AI algorithms in the next five years. I describe some of the background motivation for developing this test, and then provide a detailed account of the tasks that make up the Decathlon, and the types of results that should be expected.

Can the Turing Test be Useful and Relevant?

Alan Turing (1950) famously suggested that a reasonable test for artificial machine intelligence is to compare the machine to humans (who we agree are intelligent), and if their verbal behavior and interactions are indistinguishible, then the machine should be considered intelligent. Turing proposed that the test should be limited to verbal interactions alone, and this is how the test is typically interpreted in common usage. For example, the $100,000 Loebner prize is essentially a competition for designing the best chatbot. Although computational linguistics remains an important branch of modern AI, the field has expanded into many non-verbal domains related to embodied intelligent behavior, including specialized fields of robotics, image understanding, motor control, and active vision.

Consequently, it is reasonable to ask whether the Turing Test, and especially the traditional Verbal Turing Test (VTT) is still relevant today. Indeed, it is fair to say that almost no cutting-edge research in cognitive science or AI has a goal of passing the VTT. Indeed, current thinking about the VTT is that it is almost a joke (Sundman, 2004), or impossible goal that is not useful for current research (Shieber, 1994;). Yet there are cogent arguments that the test is relavant and useful. For example, Harnad (1989, 1990) has argued that an embodied Turing test is indeed useful, that it is actually the most useful (cf Harnad, 2000), and that it is even consistent with thought experiment originally described by Turing (see Harnad, 2004). Consequently, a version of the Turing test may be relevant, but how do we design one that is useful?

*Some of the research reported here was conducted as part of the U.S. DARPA program *Biologically Inspired Cognitive Architectures*. Paper submitted to the 19th Midwest Artificial Intelligence and Cognitive Science Conference, to be held at the University of Cincinnati, April 12-13, 2008.

Adapting the Turing Test for Modern Artificial Intelligence

To frame this argument, I first note that a general statement of the Turing test has three important aspects, each of which are somewhat ambiguous:

> Measurement of artificial behavior in (1) a specified domain that is (2) indistinguishible from (3) human behavior.

The Domain of the Turing Test. Harnad (2004) has argued that Turing's writings are consistent with the the first aspect (the domain) being a "sliding scale", and he described 5 levels of Turing Tests: 1. For a limited task; 2. For verbal context; 3. For sensorimotor context; 4. For internal structure; 5. For physical structure. Harnad argued that although Turing did not mean the first level (Turing-1), the Turing-2 is susceptible to gaming, the most useful version of the test is (Turing-3). This argument is useful because it means it is possible to develop versions of the Turing Test that are relevant to today's researchers. However, because Turing-3 is a superset of Turing-2, it means that it would be a greater challenge and perhaps even less useful that Turing-2. Yet, the other two aspects of the test may suggest ways to design and implement a useful version of the test.

The Meaning of Indistinguishible. A second aspect of the Turing Test is that it looks for "indistinguishible" behavior. On any task, the range of human behavior across the spectrum of abilities can span orders of magnitude, and there are artificial systems that today outperform humans on quite complex tasks. So, we might also specify a number of levels of "indistinguishible": at the minimum, consider the criterion of competence: the artificial system produces behavior that it at least as good as (and possibly better than) a typical human. A more stringent criteria might be called *resemblance*, requiring that typical inadequacies exhibited

by humans also be made, such as appropriate time profiles and error rates. Here, the reproduction of robust qualitative trends may be sufficient to pass the test. A test with a fidelity higher than resemblance might be called *verisimilitude*. For example, suppose a test required the agent produce behavior such that, if its responses were given along with corresponding responses from a set of humans on the same tasks, its data would not be able to be picked out as anomalous.

The criterion of verisimilitude is somewhat controversial. After all, if an artificial agent is smarter/stronger/better than its human counterpart, isn't that a sign of embodied intelligence? There are a number of contexts in which we would prefer verisimilitude over competence. For example, if the goal of developing an artificial agent is to replace a human, either as a teammate or adversary (e.g., for training), it can be useful for the agent to fail in the same ways a human fails. In other cases, if the agent is being used to make predictive assessments of how a human would behave in a specific situation, verisimilitude would be a benefit as well. Finally, as the agents were to be designed to have a computational organization akin to the human brain, behavioral performance profiles can be diagnostic measures of whether the artificial computation reflects the biological organization.

A criterion more stringent than verisimilitude might be called distributional: predicting distributions of human behavior. Given multiple repeated tests, the agent's behavior would be reproduce the same distribution of results as a sample of humans produces.

The Target of Intelligent Behavior. A third important aspect of the general Turing Test stated above is that an intelligent target which produces behavior must be specified. There is a wide range of abilities possessed by humans, and if we observe behavior that we consider intelligent in a non-human animal or system, it could equally-well serve as a target for the Turing Test. So, at one end of the spectrum, there are behaviors of top experts in narrow domains (e.g., chess grandmasters or baseball power hitters); on the other end of the spectrum, there are physically disabled individuals, toddlers, and perhaps even other animals that exhibit intelligent behavior. So, one way to frame a useable Turing-3 test is to choose a target that might be easier to mimic than an adult able-bodied human expert. The different version oft these three concepts are shown in Table 1.

This framework suggests that the Turing Test is indeed a reasonable criterion for assessing artificial intelligence, and is even relevant for embodied AI. By considering a generalized form, there are a number of ways the test can be implemented with present technology that allow for an embodied Turing-3 test to be constructed, tested, and possibly passed, even though the state of AI research is nowhere close to passing the traditional Turing-2, which is a subset of Turing-3.

In the remainder of this report, I describe just such a plan for testing embodied intelligence of artificial agents. It was an attempts to go beyond the VTT by incorporating a wide range of embodied cognitive tasks. In order to meet this goal, we chose a target that was at the lower end of the capability spectrum: performance that might be expected of a typical to 2-year-old human toddler. In addition, we relaxed the fidelity requirement to initially require competence, and later to require the reproduction of robust qualitative trends.

The Cognitive Decathlon

This research effort was funded as part of the first phase of DARPA's BICA program (Biologically-Inspired Cognitive Architectures).[1] The primary goals of the BICA program were to develop comprehensive biological embodied cognitive agents that could learn and be taught like a human. This limits the scope and difficulty of the tasks that could by accomplished in a five year program.

Goals

The test specification was designed to promote the goals of the BICA program, while encouraging the construction of models that were systematic, coherent and consistent. One hallmark of human cognition is its flexibility, and so performance should be produced by a single flexible system, rather than a set of special-purpose models cobbled together into a single meta-model. Thus, we designed the test specification to: (1) Encourage the development of coherent, consistent, systematic, cognitive system that can achieve complex tasks; (2) Promote procedural and knowledge acquisition through learning, rather than programming or endowment by modelers; (3) Involve tasks that go beyond the capabilities of traditional cognitive architectures toward a level of embodiment inspired by human biology; and (4) Promote and assess the use of processing and control algorithms inspired by neurobiological processes.

To achieve these goals, we designed three types of tests: Challenge Scenarios, the Cognitive Decathlon, and a set of Biovalidity Assessments. The Challenge Scenarios are designed to require integrated end-to-end systems, covering a wide range of capabilities over the set of test problems. The Cognitive Decathlon is intended to provide stepping stones along the way to the complex scenario tasks, testing specific systems and core competencies against human behavior. The biovalidity assessment is designed to determine how well the systems resemble the neural computation systems.

We designed a three-thrust test suite for pragmatic and conceptual reasons in order to best promote the goals of the program. Challenge scenarios were meant to be complex tests that couldn't be accomplished by small special systems; this encouraged coherent systematic architectures. Decathlon tasks were meant to be small targeted tasks could test the special systems in greater detail and provide useful comparisons to human behavioral data. The biovalidity assessments were designed to ensure that the large-scale and small-scale architectures were indeed inspired by the biology, and not just standard AI approaches mapped onto a set of brain regions.

[1] Phase I of the BICA program was the design phase. Later phases of the program were not funded, and so the Cognitive Decathlon has not yet been used to test embodied intelligence.

Target	Fidelity	Domain (Harnad, 2000)
1. Lower animals	1. Competence: can accomplish task target achieves	1. Local indistinguishibility for specific task
2. Mammals	2. Domination: Behavior better than target	2. Global Verbal performance
3. Children	3. Resemblance: reproduces robust qualitative trends	3. Global Sensorimotor performance
4. Typical Adult	4. Verisimilitude: Cannot distinguish measured behavior from target behavior	4. External & Internal structure/function
5. Human expert	5. Distributional: Produces range of behavior for target population.	5. Physical structure/function

Design of the Cognitive Decathlon

Rather than taking its inspiration from the types of tasks AI researchers have typically studied, we developed the tasks based on analysis of the core cognitive competencies of humans as they develop. Thus, we have specified a set of fine-grained behavioral tests that map onto core human skills, which was called The Cognitive Decathlon. Like the Olympic Decathlon, which attempts to measure the core capabilities of an athlete or warrior, the Cognitive Decathlon attempts to measure the core capabilities of an embodied cognitive human or agent. These tasks cover the basic range of human behavior, they are reasonable well-studied so that we understand how humans perform, and they typically have a number of computational and mathematical models available that implement theories of how humans perform the tasks.

Research on human development has shown that by 24-months, children are capable of a large number of cognitive, linguistic and motor skills. For example, according to the Hawaii Early Learning Profile development assessment, the linguistic skills of a typical 24-month-old child include the ability to name pictures, use jargon, use 2-3 word sentences, produce 50 or more words, answer questions, and coordinate language and gestures. Their motor skills include walking, throwing, kicking, and catching balls, building towers, carrying objects, folding paper, simple drawing, climbing, walking down stairs, and imitating manual and bilateral movements. Their cognitive skills include matching (names to pictures, sounds to animals, identical objects, etc.), finding and retrieving hidden objects, understanding most nouns, pointing to distant objects, and solving simple problems using tools (Parks, 2006). These component tasks of the Cognitive Decathlon were designed to exercise these core skills.

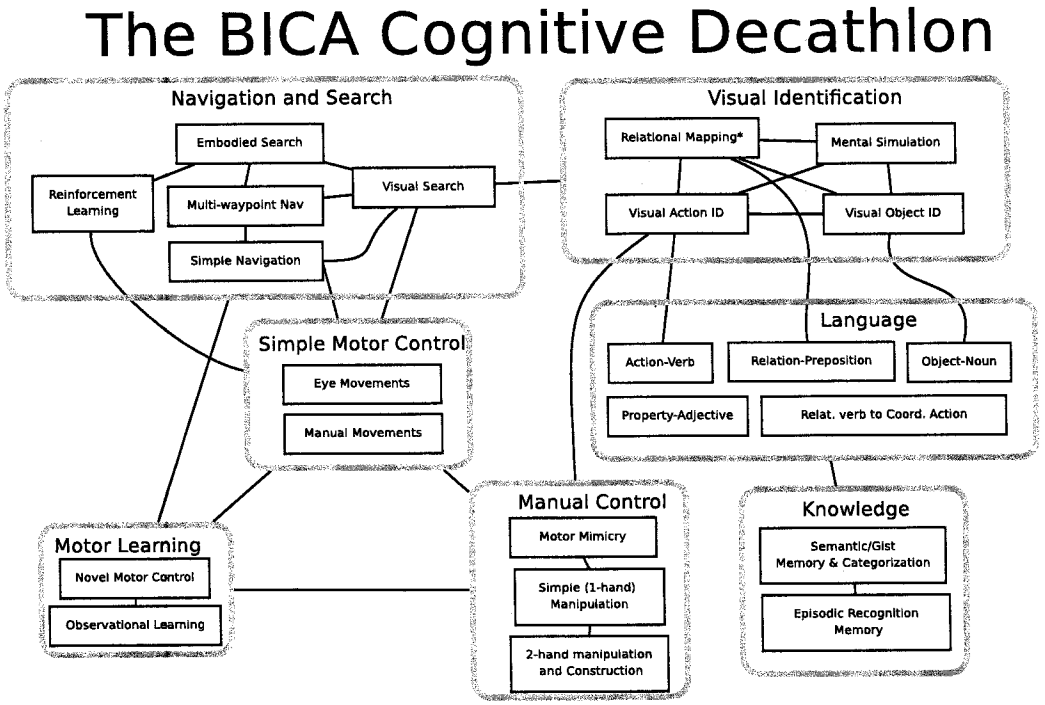
We anticipated that the agent would be embodied in a photorealistic virtual environment or robotic platform with controllable graspers, locomotion, and orientation effectors with on the order of 20-40 degrees of freedom. The EU RobotCub project (Sandini, Metta, & Vernon, 2004) is perhaps the most similar effort, although that effort is focused on building child-like robots rather than designing end-to-end cognitive-biological architectures.

Like the Olympic Decathlon, the BICA "Cognitive Decathlon" was designed to test a range of core skills used to accomplish more complex tasks. Despite its name, the decathlon involves roughly 20 sub-tasks or tests organized into six task categories. The primary motivation for these tasks is to test the component skills that are involved in solving the challenge problems against behavioral and biological standards. This design was chosen to guide the independent modeling teams in building coherent systems that solve complex problems in ways similar to human performers, while encouraging a reusable modular approaches rather than special-purpose engineered solutions. Additionally, the tasks limited scope provides a better comparison to empirical and neurobiological data. Prior research using these tasks has produced a wealth of empirical data on adults and children performance characteristics. We anticipated comparing agent performance to robust trends identified in these prior experiments, as well as conducting new experiments where necessary. We provide basic descriptions of these tasks below, along with some information on the prior research.

Table 1: Component tasks of the cognitive decathlon.

Task	Level
1. Vision	Invariant Object Identification Object ID: Size discrimination Object ID with rotation Visual Action/Event Recognition
2. Search	Navigation Visual Search Simple Navigation Traveling Salesman Problem Embodied Search Reinforcement Learning
3. Manual Control and Learning	Motor Mimicry Simple (1-hand) Manipulation Two-hand manipulation Device Mimicry Intention Mimicry
4. Knowledge Learning	Episodic Recognition Memory Semantic Memory/Categorization
5. Language and Concept Learning	Object-Noun Mapping Property-Adjective Relation-Preposition Action-Verb Relational Verb-Coordinated Action
6. Simple Motor Control	Eye Movements Aimed manual Movements

Figure 1: Graphical depiction of the Cognitive decathlon. Grey rounded boxes indicate individual tasks that require the same basic procedural skills. Black rectangles indicate individual trial types or task variations. Lines indicate areas where there are strong relationships between tasks.



Visual Identification

The ability to identify visual aspects of the environment is a critical skill used for many tasks faced by humans. This skill is captured in a graded series tests that determine whether an agent can tell whether two 'objects' or "Events" are identical; and what parts of two complex events or objects play corresponding roles.

The notion of sameness (cf. French, 1995) is an ill-defined and perhaps socially constructed concept, and this ambiguity helps structure a series of graded tests. Typically, objects used for identification will be comprised of two or more connected components, have one or more axes of symmetry, and have color and weight properties. Objects can differ in color, weight, size, component structure, relations between components, time of perception, movement trajectory, location, or orientation. In these tasks, color, mass, size, component relations are defined as integral features to an object, and differences along these dimensions are sufficient to consider two objects different. Neuropsychological findings (e.g., Wallis & Rolls, 1997) show that sameness detection is invariant to differences in translation, visual size, and view, and differences along these dimensions should not be considered sufficient to be indicate difference.

In the basic task, the agent will be shown two objects., and be required to determine whether the objects are the same or different. The different types of trials include:

Invariant Object Recognition. On same trials, the objects will be oriented in the same direction. On different trials, objects will differ along color, visual texture, or shape. Even poor visual systems should be able to perform well in this task,

Size Differences. Objects are perceived as maintaining a constant size even when the observer distance changes, creating large differences in the stimulus size. Some neural mechanisms involved in object identification have been shown to be invariant to differences in size, detecting whether two objects that are identical in shape. Thus, discriminating between two objects with identical shape but different size can be challenging. This type of trial tests the ability to discriminate size differences in two identically-shaped objects. Success in the task is likely to require incorporating at least one other type of information, such as body position, binocular vision, or other depth cues.

Identification requiring rotation. Complex objects often need to be aligned and oriented in order to detect sameness. On these trials, identical objects will be rotated along two orthogonal axes, so that physical or mental rotation is required to correctly identify whether they are the same or different.

Event Recognition. Perceptual identification is not just static in time; it includes events that occur as a sequence

of path movements and interactions in time. This test examines the agent's ability to represent and discriminate such events. The two objects will repeat through a short equally-timed event loop (e.g., rotating, moving, bouncing, etc.) and the agent is required to determine whether the two depicted events are the same.

Search and Navigation. A critical skill for embodied agents is the ability to navigate through an environment, which forms the basis for numerous search skills and aspects of spatial cognition. A graded series of decathlon events, described in the following sections, tests these abilities.

Visual Search

A core skill required for many navigation tasks is the spatial localization of a goal target. In the visual search task, the agent will view a visual field containing a number of objects, including (on target-present trials) the well-learned target light. The agent is expected to determine whether the target is or is not present, responding verbally ("YES" or "NO"). Behavior similar to human performance will be expected for simple task manipulations (e.g., both color-based pop-out and deliberate search strategies should be observed).

Simple Navigation. In this task, the agent will be given the verbal task cue "Find the target", and will be expected to identify and move to the red target light in a room containing obstacles. The target light will be visible to the agent from its starting point, but may be occluded at intermediate points, depending upon the navigation path. Obstacles of different shapes and sizes will be present in the room, and will change from trial to trial. On some trials, the path to the object may be obstructed by movable and manipulable objects, and success would require clearing these obstacles. Agents will be assessed on their competency in the task as well as performance profiles in comparison to human solution paths.

Traveling Salesman Problem. A skill required for many of the Challenge Scenarios is the ability to investigate multiple locations in a room, forming an efficient search path through to different points of interest. This requires prioritizing navigation to multiple points. This skill has been studied in humans in the context of the Traveling Salesman Problem.

The Euclidean TSP (E-TSP) belongs to a class of problems that are "NP-Complete", which means that algorithmic solutions can require exhaustive search through all possible paths to find the best solution. This is computationally intractable for large problems, and so presents an interesting challenge for classic AI approaches to intelligence, which typically rely on search through the problem space. Such approaches would produce solution times that scale as a power of the number of cities, and would never succeed at finding solutions to large enough problems. Yet human solutions to the problem are typically close to optimal (5% longer than the minimum path) and efficient (solution times that are linear with the number of cities) indicating that humans solve the problem in ways fundamentally different from traditional approaches. Recent research (e.g., Pizlo, et al., 2006) has suggested that humans rely on their visual systems to solve the problem, and such skill may form the basis of many human navigation abilities. Thus, this task is ideally suited for evaluating the biologically-inspired cognitive agents, as it tests skills (prioritized navigation) that are important for embodied agents and are solved by humans in ways that rely closely on the architecture of their visual system.

The agent will be tested by being given a verbal task cue ("Find the targets"), after which it will be expected to visit all the target locations. Once visited, each target light will disappear, to enable task performance without remembering all past visited locations. The agents' performance will primarily be based on competence (ability to visit all objects), and secondarily on comparison to robust behavioral findings regarding this task (solution paths are close to optimal with solution times that are roughly linear with the number of targets.)

Embodied Search. True search ability requires some amount of metaknowledge, to remember the places that have already been searched. In this task, the agent must find a single target light, which is located inside one of a number of occluders scattered around the test room. The target can be detected only when an occluder is approached. The target will be presented randomly, so that all locations have equal probability of hiding the target light. Performance will be expected to be efficient, with search time profiles and perseveration errors (repeated examination of individual boxes) resembling human data.

Reinforcement Learning. The earlier search tasks have fairly simple goals, yet human's ability to search and navigate often supports higher-order goals such as hunting, foraging, path discovery. Reinforcement learning plays an important role in these more complex search tasks, guiding exploration to produce procedural skill, and tying learning to motivational and emotional systems. To better test the ways reinforcement learning contributes to search and navigation, the agents will perform a modified search task that closely resembles the so-called Iowa Gambling Task (e.g., Bechara et al., 1994).

The task is similar to the Embodied Search Task, but the target light will be hidden probabilistically in different locations on each trial. Different locations will be more or less likely to contain the hidden object, which the agent is expected to learn and exploit accordingly. The probabilistic structure of the environment may change mid-task, as happens in the Wisconsin Card Sort (Berg, 1954), and behavior should be sensitive to such changes, moving away from exploitation toward exploration in response to repeated search failures.

Manual Control & Learning

Along with visual and navigational skills, the agents will have ability to control its arms and graspers in order to ma-

nipulate the environment. Initial simple control of these effectors will be tested in the Simple Motor Control test (see below). This event incorporates for levels that go beyond simple control.

Motor Mimicry. One pathway to procedural skill is the mimicry of the actions of others. This task tests this skill by evaluating the agents ability to copy manual actions. For this task, the agent will mimic hand movements of the instructor, including moving fingers, rotating hand, moving arms, touching a location, etc., but will not include the manipulation of artifacts or the requirement to move two hands/arms in a coordinated manner. Mimicry is expected to be egocentric and not driven by shared attention to absolute locations in space. Agents will be assessed on their ability to mimic these novel actions, and the complexity of the actions that can be mimicked.

Simple (One-hand) Manipulation. A more complex type of mimicry involves interacting with objects in a dexterous way. Based on simple verbal instructions, the agent is expected to grasp, pick up, rotate, move, put down, push, or otherwise manipulate objects, copying the actions of an instructor. Given the substantial skill required for coordinating two hands, all manipulations in this version of the task will involve a single arm/grasper. The agent will be expected to copy the instructor's action with its own facsimile of the object. Mimicry is expected to be egocentric and not based on shared attention, although produced actions can be mirror-image of the instructors. Agents will be assessed on their ability to mimic these novel manipulations, and the complexity of the actions they are able to produce.

Two-hand Manipulation. Based on simple verbal instructions ("Copy Me."), the agent will mimic 2-hand coordinated movement and construction. Actions might include picking up objects that requiring two hands, assembling or breaking two-piece objects; etc. Evaluation will be similar to the Simple Manipulation task.

Device Mimicry. Although the ability to mimic the actions of a similar instructor is critical, human observational learning allows for more abstract mimicry. A well-engineered mirror neuron system could possibly map observed actions onto the motor commands used to produce them, but might fail if the observed actions are produced by a system that physically differs from the agent, or if substantial motor noise exists. This task goes beyond direct mimicry of action to tasks that require the mimicry of complex tools and devices, and (in a subsequent task) intentions.

The task involves learning how a novel motor action maps onto a physical effect in the environment. The agent will control a novel mechanized device (e.g., an articulated arm or a remote control vehicle) by pressing several action buttons with the goal of accomplishing some task. The agent will be given opportunity to explore how the actions control the device. When it has sufficiently explored the control of the device, the agent will be tested by an instructor who controls the device to achieve a specific goal (e.g., moving to a specific location). The instructor's control operations will be visible to the agent, so that it can repeat the operations exactly if it chooses. The instructor will demonstrate the action, and will repeat the sequence if requested.

Intention Mimicry. This task is based on the device mimicry task, but tests more abstract observational learning, in order to promote understanding of intention and goal inference. The agent will observe a controlled simulated device (robot arm/remote control vehicle) accomplish a task that requires solving a number of sub-goals. The instructor's operator sequence will not be visible to the agent, but the agent will be expected to (1) achieve the same goal in a way (2) similar to how the instructor did. Performance success and deviation from standard will be assessed.

Knowledge Learning

A major goal of the BICA program was to develop agents that learn ubiquitously and incidentally about their environment and can use this to solve later tasks. We included several memory assessments to determine the extent to which the knowledge memory system produces results resembling robust human behavioral findings.

Episodic Recognition Memory. A key type of information required for episodic memory is the ability to remember a specific occurrence of known objects or events in a specific context. To ensure a basic familiarity with all objects to be used in testing, the agent will begin in a small "familiarization" room containing a number of objects that can be observed and examined. After a short pre-determined period of time, the agent will move to a new room (a testing room) and be shown a series of configurations of objects. After a short break, the agent will be shown another series of objects or events and be asked "Did you see this here before?" All the objects in the test episodes will have been present in the familiarization room, but only some (the targets) will have been shown in the testing room. Agents should interpret the instructions to mean a specific combination of objects in a specific arrangement in the specific room the test is occurring in. Agents should produce strength effects, (i.e., be better at identifying objects that were given more study time). A secondary phenomenon to be produced is the strength-based mirror effect, in which hits are greater and false alarms are fewer when the stimuli are given more study.

Semantic Gist/Category Learning. An important aspect of human semantic memory is the ability to extract the basic gist or meaning from complex and isolated episodes. This skill is useful in determining where to look for objects in search tasks, and the ability to form concept ontologies and fuzzy categories.

The agent will view a series of objects formed from a small set of primitive components. Each object will be labeled verbally by the instructor, and the objects will fall into a small number of categories (e.g., 3–5). No two objects will

be identical, and the distinguishing factors will be both qualitative (e.g., the type of component or the relation between two components) and relative (e.g., the size of components). Following study, the agent will be shown novel objects and be asked whether it belongs to a specific category (Is this a DAX?). Category membership will not be exclusive, may be hierarchically structured, and may depend upon probabilistically on the presence of features and the co-occurrence and relationship between features. Agent will be expected to categorize novel objects in ways similar to human categorization performance.

Language/Concept Learning

Language understanding plays a central role for instruction and tasking, and opens up the domain of tasks that can be performed by the agents. Language grounding is a critical aspect of language acquisition (cf. Landau et al., 1998), and we will use a series of five tests evaluate the agents ability to learn mappings between physical objects or events and the words used to describe them. For each test type, the agent will be shown examples with verbal descriptions, and later be tested on yes-no transfer trials. Brief descriptions of each test type are given below.

Noun-Object Mapping. One of the first language skill developed by children is the ability to name objects (Smith & Gasser, 1998), and even small children can form object-name mappings quickly and permanently with a few examples. This test examines the ability to learn the names of objects.

Adjective-Property Mapping. A greater challenge is learning how adjectives refer to properties of objects, and can apply to a number of objects. Such skill follows object naming, and typically requires many repetitions to master. This test examines the ability of an agent to learn adjectives, and recognize their corresponding properties in novel objects.

Preposition-Spatial Relation Mapping. Research has suggested that many relational notions are tied closely to the language used to describe them. Spatial relations involve relations of objects, and so rely not just on presence of components but their relative positions. This test examines the ability of an agent to infer the meaning of a relation, and recognize that relation in new episodes.

Verb-Action Mapping. Recognition is not static in time, but also involves events occurring in time. Furthermore, verbs describing these events are abstracted from the actor objects performing the event, and represent a second type of relation that must be learned about objects (Gentner, 1978). This test examines the ability of the agent to represent such events and the verb labels given to them, and recognize the action taking place with new actors in new situations.

Relational Verbs-Multi-object actions. The most complex linguistic structure tested will involve relational verbs, which can describe multi-object actions whose relationship is critical to the correct interpretation For example, in the statement, "The cat chased the dog.", the mere co-presence of dog and cat do not unambiguously define the relationship. This test examines the ability of the agents to understand these types of complex linguistic structures and how the relate to events in the visual world.

Simple Motor Control

Because fairly complex motor control will be required, the low-level components of this control will be tested in comparison to robust human behaviors. Arguably, low-level gross locomotion and manipulation are tested in other tasks; the following tasks focus on properties of how eyes and other effectors are moved.

Saccadic Eye Movements. One form of eye movement is known as a saccade, which is typically a ballistic movement occurring with low latency and durations to a specific location in visual space. This ability will be tested by presenting target objects in the visual periphery, to which the agent will shift its eyes in saccadic movements, with time and accuracy profiles similar to humans.

Smooth Pursuit Eye Movements. Additionally, humans are able to smoothly track moving objects. Such a skill relies on close linkage between the ocular, motor, vestibular, and perceptual processes, and presents a useful test of their integration. Agents will be expected to smoothly track objects moving in trajectories and velocities similar to those humans are capable of tracking.

Aimed Manual Movement. Fitts's (1954) law states that the time required to make an aimed movement is proportional to the log of the ratio between the distance moved and the size of the target. Agents will be tested in their ability to make aimed movements to targets of varying sizes and distances, and are expected to produce Fitts's law at a qualitative level.

Plan for Testing

Although the tests here are presented as a complete set, many individual components form natural progressions of complexity. For example, the language mapping tasks progress from simplest (object-noun) to most complex (requiring complex relations and abstract labels.) Our intent for the program was to stagger the testing requirements so that the more primitive tasks were tested earlier in the program, and more complicated tests built up later.

Discussion

This report describes the motivation and design for the "Cognitive Decathlon", a version of the Turing test designed to be useful and relevant for the current domains of study in

Artificial Intelligence. The goal was to design a comprehensive set of tests that could be accomplished by a single intelligent agent using available technology in the next five years.

References

Arbib, M.A., Billard, A., Iacoboni, M. & Oztop E. (2000). Synthetic brain imaging: grasping, mirror neurons and imitation. Neural Networks, 13, 975-997.

Bechara A, Damasio AR, Damasio H, Anderson SW (1994). Insensitivity to future consequences following damage to human prefrontal cortex. Cognition, 50: 7-15.

Berg, E. A. (1948). A simple objective technique for measuring flexibility in thinking J. Gen. Psychol. 39: 15-22.

Busemeyer, J. & Wang, Y. (2000). Model Comparisons and Model Selections Based on Generalization Criterion Methodology. Journal of Mathematical Psychology, 44, 171-189.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. Journal of Experimental Psychology, 47, June 1954, pp. 381-391. (Reprinted in Journal of Experimental Psychology: General, 121(3):262–269, 1992.

French, R. M. (1995). The Subtlety of Sameness. Cambridge, MA: The MIT Press, ISBN 0-262-06810-5.

Gasser, M. & Smith, L. B. (1998). Learning nouns and adjectives: A connectionist account. Language and cognitive processes,13, 269-306.

Gentner, D. (1978) On relational meaning: The acquisiton of verb meaning. Child Development, 48, 988-998.

Gluck, K. A. & Pew, R. W. (2005). Modeling human behavior with integrated cognitive architectures. Mahwah, New Jersey: Lawrence Erlbaum.

Harnad, S. (1990), The Symbol Grounding Problem, *Physica D 42*, 335–346.

Harnad, S. (1991), Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem, *Minds and Machines 1*, 43–54.

Harnad, S. (2001). Minds, Machines and Turing: The Indistinguishability of Indistinguishables. Journal of Logic, Language, and Information.

Harnad, S. (2004). The Annotation Game: On Turing (1950) on Computing, Machinery, and Intelligence. in Epstein, R. & Peters, G Eds.) *The Turing Test Sourcebook: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Kluwer.

Landau, B., Smith, L., & Jones, S. (1998). Object shape, Object Function, and Object Name. Journal of Memory and Language, 38, 1-27.

Myung, I. J.. (2000). The Importance of complexity in model selection. Journal of Mathematical Psychology, 44,190-204.

Parks, S. Inside HELP, Administrative and Reference Manual. Palo Alto, CA: VORT Corp, ISBN 0-89718-097-6.

Pizlo, Saalweachter, & Stefanov. (2006) "Visual solution to the traveling salesman problem". Journal of Vision (6). http://www.journalofvision.org/6/6/964/

Sandini, G., Metta, G. & Vernon, D. (2004). RobotCub: An open framework for research in embodied cognition. International Journal of Humanoid Robotics, 8, 1-20.

Shieber, S (1994). Lessons form a Restricted Turing Test. Communications of the Association for Computing Machinery, 37, 70–78.

Sohn, M.H., Goode, A., Stenger, V.A., Jung, K.J., Carter, C.S. & Anderson, J.R. (2005). An information-processing model of three cortical regions: evidence in episodic memory retrieval. NeuroImage, 25, 21-33.

Sundman, J. (200x), Artificial Stupidity. Salon, Feb. 2003.

Turing, A. (1950). Computing machinery and intelligence. Mind, LIX, 433-460.

Wallis, G. & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. Progress in Neurobiology, 51, 167-194.