

Adapting the Turing Test for Embodied Neurocognitive Evaluation of Biologically-Inspired cognitive agents.

Shane T. Mueller, Ph.D.*

Klein Associates Division, ARA Inc.
Fairborn, OH 45324
smueller@ara.com

Brandon S. Minnery, Ph.D.†

The MITRE Corporation
1750 Colshire Drive, McLean, VA
bminnery@mitre.org

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. We 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 a human (who we agree is intelligent), and if their verbal behaviors and interactions are indistinguishable from one another, the machine might 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. However, although linguistics remains an important branch of modern AI, the field has expanded into many non-verbal domains related to embodied intelligent behavior. These include 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.

*Part of the research reported here was conducted as part of the U.S. DARPA program, contract FA8650-05-C-7257, *Biologically Inspired Cognitive Architectures*, and presented at the 2007 BRIMS conference and the 2008 MAICS conference. Approved for Public Release, Distribution Unlimited.

†Part of the research reported here was conducted as part of the U.S. DARPA program *Biologically Inspired Cognitive Architectures*. Approved for public release, distribution unlimited. No. 07-0258.

Copyright © 2008, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Indeed, it is fair to say that almost no cutting-edge research in cognitive science or AI has a goal of passing the VTT. Some observers have suggested the VTT is a stunt or a joke (e.g., Sundman, 2004), or an impossible goal that is not useful for current research (Shieber, 1994). Yet some have argued that the test is indeed relevant for the types of research that is being produced today. For example, Harnad (1989, 1990, 2000, 2004) argued that an embodied version of the Turing test is consistent with Turing's original thought experiment, which matches the domains of today's research. This argument (expanded in the next section) suggests that we can still look to the Turing Test as a way to measure intelligence, but it presents a challenge as well. Given that even the VTT seems to be an impossible goal, embodied versions of the Turing test (which are supersets of the VTT) would seem an even greater challenge. Yet, perhaps by relaxing some of the properties of the Turing Test, a version that is both relevant and useful to today's researchers can be framed.

Adapting the Turing Test for Modern Artificial Intelligence

A general statement of the Turing test has three important aspects, each of which are somewhat ambiguous:

A machine can be considered intelligent if its behavior in (1) a specified domain is (2) indistinguishable from (3) human behavior.

The Domain of the Turing Test. The first aspect describes the domain of the test. Harnad (2004) argued that Turing's writings are consistent with the domain being a *sliding scale*, and he described five levels of Turing Test domains: 1. For limited tasks; 2. For verbal context; 3. For sensori-motor 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 test (which is the most common interpretation) is susceptible to gaming. A more powerful and relevant version consistent with Turing's argument is Turing-3: a sensori-motor Turing Test. 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 than Turing-2, because it would be an even more difficult test to pass. Yet, the other two aspects of the test may suggest ways to design and implement a useful version of the test.

The Meaning of Indistinguishable. A second aspect of the Turing Test is that it looks for “indistinguishable” 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 but limited tasks. So, we might also specify a number of levels of “indistinguishable”: at the minimum, consider the criterion of competence: the artificial system produces behavior that is at least as good as (and possibly better than) a typical human. This is a useful criterion in many cases (and is one that has placed humans and machines in conflict at least since John Henry faced off against the steam hammer.) 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 higher fidelity 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 might be viewed as somewhat contentious, because an artificial agent that is smarter/stronger/better than its human counterpart might be considered to exhibit embodied intelligence. There are a number of contexts in which one might prefer verisimilitude over competence. For example, if one’s goal is to develop an artificial agent that can replace a human as a teammate or adversary (e.g., for training, design, or planning), 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, this criterion can provide some tests for *how* an agent processes information and reasons: for example, if one’s goal is to create a system that processes information like the human brain, verisimilitude can improve the chances of developing the right algorithms without having to understand exactly how the brain achieves the processing.

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

The Target of Intelligent Behavior. A third important aspect of a 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 who exhibit intelligent behavior. So, one way to frame a useable Turing-3 test is to choose a target that might be easier than an adult able-bodied human expert. The different version of 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 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 VTT.

In the remainder of this report, we describe such a plan for testing embodied intelligence of artificial agents. It was an attempt 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 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). Phase I of the BICA program was the design phase, during which the set of tests described here were selected. Later phases of the program were not funded, and so these tests have not been used as a comprehensive evaluation suite for embodied intelligence. The simulated BICA agents were planned to be embodied either in a photorealistic virtual environment or on 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.

Goals

The primary goals of the BICA program were to develop comprehensive biological embodied cognitive agents that could learn and be taught like a human. The test specification was designed to promote these goals, encouraging the construction of models that were capable of a wide range of tasks, but approached them as a coherent system rather than a loose collection of subsystems designed to solve each individual task. 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 semantic 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

Table 1: Variations on three aspects of the Turing Test.

Target	Fidelity	Domain (Harnad, 2000)
1. Lower animals	1. Competence: can accomplish task target achieves	1. Local indistinguishability 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

embodiment inspired by human biology; and (4) Promote and assess the use of processing and control algorithms inspired by neuro-biological processes.

To achieve these goals, we designed three types of tests: the Cognitive Decathlon (which is the focus of this report); integrative “challenge scenarios”, and a set of Biovalidity Assessments. The Challenge Scenarios were designed to require interaction between different subsystems in order to achieve a high-level task. The biovalidity assessments were designed to determine the extent to which the artificial systems used computation systems inspired by neurobiology. The “Cognitive Decathlon” was intended to provide detailed tests of core cognitive functions, and provide stepping stones along the way to achieving the more complex challenge scenario tasks.

Design of 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. To enable an achievable capability level within the scope of the program, target behavior of a two-year old human toddler was selected. There were many motivations for this target, but one central notion is that if one could design a system with the capabilities of a two-year-old, it might be possible to essentially grow a three-year-old, given realistic experiences in a simulated environment. The tasks we chose covered a broad spectrum of verbal, perceptual, and motor tasks, and attempt to cover many of the intelligent behaviors of a toddler, and the assessment criteria were planned to require competence in early years, and reproduction of robust qualitative trends in later years.¹

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 in-

clude 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).

To develop the decathlon, we first began by examining hundreds of empirical tasks studied by psychologists in research laboratories. From these, we selected a set of specific tests for which (1) human performance on the tasks were fairly well understood; (2) there typically existed computational or mathematical models accounting for these behaviors; (3) were related to the core abilities of a two-year-old child; and (4) were components that are frequently integrated to accomplish more complex tasks. Basic descriptions of these tasks are provided below, along with some information regarding human performance on the tasks.

Our basic taxonomy of tasks is shown in Table 2. We identified six taxons that describe basic skill types, and which are tied back to distinct neural or biological systems. A number of other taxonomies of cognitive skill have been used in other contexts. For example, the Hawaii Early Learning Profile (Parks, 2006) describes six taxons: cognitive, language, gross motor, fine motor, social, and self-help. Our taxons focused on the first four of these, and view social interactions as a ubiquitous manner of interacting outside the scope of the taxonomy. As another example, the Army’s IMPRINT tool recognizes nine taxons: Visual, numerical, cognitive, fine motor discrete, fine motor continuous, gross motor heavy, gross motor light, communication–reading and writing, and communication–oral (Allender, Sali, & Promisel, 1997). Our taxonomy covers more than half of these domain, avoiding reading and writing and numerical skills. Our selection of tasks was guided by the desire to have fairly comprehensive coverage of low-lever cognitive core skills, while highlighting tasks in which standard AI approaches would accomplish in ways fundamentally different from human performers.

Visual Identification

The ability to identify visual aspects of the environment is a critical skill used for many tasks faced by humans. In the decathlon, this skill is captured in a graded series tests that

¹In the scope of the BICA program, the ability of agents to achieve specific performance criteria was a requirement for continued funding in subsequent phases.

Table 2: Component tasks of the cognitive decathlon.

Task	Level
1. Vision	Invariant Object Identification Object ID: Size discrimination Object ID with rotation Object ID: relations 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

determine if an agent can tell whether two objects or events are identical.

The notion of sameness is an ill-defined and perhaps socially constructed concept (cf. French, 1995), and this ambiguity helped structure a series of graded tests related to visual identification. Typically, objects used for identification should 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 should be deemed 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.

The object recognition tasks are important tests of biological intelligence because they are a fairly important means by which we interact with the world, and the machine vision community has developed many successful algorithms that are not inspired by biological structures.

In the basic task, the agent should be shown two objects, and be required to determine whether the objects are the same or different. For each variation, both “same” and “different” trials should be presented. The different variations include:

Invariant Object Recognition. The goal of this trial type is to provide a simple task that rudimentary visual systems can accomplish. On “same” trials, the objects should be oriented in the same direction. On “different” trials, objects should differ along color, visual texture, or shape properties.

Size Differences. An object is perceived as maintaining a constant size even when its distance to the observer distance (and thus the size of its proximal stimulus) changes. In fact, neural mechanisms have developed that are sensitive to shape similarities regardless of the size (Wallis & Rolls, 1997). This type of trial should test 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 some way to detect sameness. This skill can often be accomplished by adult humans through “mental rotation” (Shepard & Metzler, 1971), although other strategies (physical rotation or even moving to different viewing positions) can also succeed. On these trials, identical objects should be rotated along two orthogonal axes, so that physical or mental rotation is required to correctly identify whether they are the same or different. Typical human performance response times for both same and different trials increase as the angle of rotation is increased, a result that may be diagnostic of the computational representations used by the agent.

Relation Identification. As described earlier, the objects used in these tasks should have multiple components, which requires an understanding of the relations between these components. As a greater challenge, simple spatial relations among sets objects should also be tested. These should map onto the prepositions tested in the language skills tasks.

Event Recognition. Perceptual identification is not just static in time; it also includes events that occur as a sequence of movements along a trajectory in time. This trial type examines the agent’s ability to represent and discriminate such events. The two objects should repeat through a short equally-timed event loop (e.g., rotating, moving, bouncing, etc.) and the agent would be required to determine whether the two events are the same or different.

Search and Navigation.

A critical skill for embodied agents is the ability to navigate through and learn about its environment. Search and navigation tasks form a fundamental cognitive skillset used by lower animals and adult humans alike. Furthermore many automated search and navigation systems employ optimization techniques or require GPS navigation or terrain databases to be succeed. A fundamental property of human navigation is that we don’t require these external navigation points, and indeed we learn the terrain by experiencing it.

Thus, search and navigation tasks can be useful in discriminating biological from non-biological spatial reasoning systems. A graded series of decathlon events tests these abilities.

Visual Search. A core skill required for many navigation tasks is the spatial localization of a target. In the visual search task, the agent should view a visual field containing a number of objects, including a well-learned target. The agent should determine whether the target is or is not present. Behavior similar to human performance for simple task manipulations should be expected (e.g., both color-based pop-out and deliberate search strategies should be observed; cf. Treisman & Gelade, 1980).

Simple Navigation. In this task, the agent should have the goal of finding and moving to a target (e.g., a red light) in a room containing obstacles. Obstacles of different shapes and sizes should be present in the room (to allow landmark-based navigation), and should change from trial to trial (to prevent learning specific configurations). For simple versions of the task, the target should be visible to the agent from its starting point, but difficulty can be increased by allowing obstacles to occlude the target either at the beginning of the trial or at intermediate points. Agents should 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 spatial reasoning tasks is the ability to navigate to multiple locations in an efficient search path through multiple points of interest. This skill has been studied in humans in the context of the Traveling Salesman Problem (TSP).

The TSP belongs to a class of problems that are “NP-Complete”, which means that algorithmic solutions potentially 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 problem solving approaches that rely on search through a problem space. Such approaches could produce solution times that scale as a power of the number of cities, and would never succeed at finding efficient solutions to large 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) suggesting human solutions to the task are fundamentally different from traditional approaches in computer science. Recent research (e.g., Pizlo, et al., 2006) has suggested that the multi-layered pyramid structure of the visual system enables efficient solutions of the task, and that such skills may form the basis of many human navigation abilities.

For this task, the agent should have the goal of visiting a set of target locations in a room. Once visited, each target light can disappear, to enable task performance without needing to remember all past visited locations. The agents’ performance should 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. Search ability requires some amount of metaknowledge, such as the ability remember locations that have already been searched. In this task, the agent must find a single target light occluded in such a way that it can only be seen when approached. Multiple occluders not containing the target should be present in the search area. Performance should 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 our 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, this task requires the agents to perform a modified search task that closely resembles tasks such as the N-armed bandit (e.g., Sutton & Barto, 1998) or the Iowa Gambling Task (e.g., Bechara et al., 1994).

The task is similar to the Embodied Search Task, but the target light should be hidden probabilistically in different locations on each trial. Different locations should 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.

Reinforcement learning goes beyond just spatial reasoning, and indeed is an important skill in its own right. Although machine learning has long been tied closely with psychological and biological learning theory (cf. Bush & Mosteller, 1951; Rescorla & Wagner, 1972), advances in machine learning have provided systems that can outlearn humans in limited domains. Thus, tests of learning can provide a good test of biological inspiration and discriminate between biological and non-biological mechanisms.

Simple Motor Control

A critical aspect of embodied intelligence is the ability to control motor systems. These tests are designed to compare some aspects of low-level motor control to human counterparts; later tests (in the section “Manual Control and Learning”) require more complex motor skills. The motivation for these tasks is that low-level task performance constraints imposed by these control mechanisms can have cascading effects that impact performance on many higher-level tasks. These biological factors place strong constraints on task performance that are not necessarily faced by robotic or engineered control mechanisms, and so they offer discriminative tests of biological inspiration.

Saccadic and Smooth Pursuit Eye Movements. Humans use two basic forms of voluntary eye movement (cf. Krauzlis, 2005): saccades, which are ballistic movements to a specific location or targets occurring with low latency and brief duration; and pursuit movements, which are smooth continuous movements following specific targets. Saccadic movements should be tested by presenting target objects in the visual periphery, to which the agent should shift its eyes in discrete movements, with time and accuracy profiles similar to humans. Pursuit movements should be tested by requiring the agent to track objects with its eyes 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 should 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.

Manual Control & Learning

Building on these simple motor skills, embodied agents should have ability to control arms and graspers to manipulate the environment. The following tasks evaluate these skills in a series of more and more complex tests.

Motor Mimicry. One pathway to procedural skill is the ability to mimic the actions of others. This task tests this skill by evaluating the agents ability to copy manual movements. For this task, the agent should replicate hand movements of an instructor (with identical embodiment), including moving fingers, rotating hands, moving arms, touching a locations, etc. This test should not include the manipulation of artifacts or the requirement to move two hands/arms in a coordinated manner. Mimicry should be ego-centric and not driven by shared attention to absolute locations in space, but errors related to left-right symmetries can be relaxed. Agents should 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 mimicry involves interacting with objects in a dexterous fashion. The agent should be expected to grasp, pick up, rotate, move, put down, push, or otherwise manipulate objects, copying the actions of an instructor. Given the possibility of substantial skill required to coordinate two hands, all manipulations in this version of the task should involve a single arm/grasper. The agent should 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 should be assessed on their ability to mimic these novel manipulations, and the complexity of the actions they are able to produce.

Two-hand Manipulation. With enough skill, an agent should be able to mimic 2-hand coordinated movement and construction. Actions could include picking up objects that requiring two hands, assembling or breaking two-piece objects; etc. Evaluation should be similar to the Simple Manipulation task, but for these more complex objects and actions.

Device Mimicry. Although the ability to mimic the actions of a similar instructor is a critical sign of intelligence, human observational learning allows for more abstract mimicry. For example, a well-engineered mirror neuron system might be able to 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, or if the objects the teacher is manipulating differs from the one the learner is using. This task goes beyond direct mimicry of action to tasks that require the mimicry of complex tools and devices, and (in a subsequent task) the teacher's intent.

The task involves learning how a novel motor action maps onto a physical effect in the environment. The agent should 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 should be given opportunity to explore how the actions control the device. When it has sufficiently explored the control of the device, the agent should 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 should be visible to the agent, so that it can repeat the operations exactly if it chooses. The instructor should demonstrate the action, and should 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 intent and goals of the teacher. The agent should 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 *should not* be visible to the agent, but the agent should be expected to (1) achieve the same goal in a way (2) similar to how the instructor did. Performance success and deviation from standard should be assessed.

Knowledge Learning

Humans learn incidentally about their environment, without needing to explicitly decide that objects and events need to be committed to memory. The tests described next include several memory assessments that 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. For this test, an agent should be allowed to explore a room containing a series of configurations of objects. After

a short break, the agent should be shown a new set of object configurations and be required to determine which of them had been seen during the learning period. Agents should display robust qualitative trends exhibited by humans in such tasks. For example, they should be better at identifying objects that were given more study time; and increase false alarms for new configurations of previously-seen objects.

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 should view a series of objects formed from a small set of primitive components. Each object should be labeled verbally by the instructor, and the objects should fall into a small number of categories (e.g., 3–5). No two objects should be identical, and the distinguishing factors should 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 should be shown novel objects and be asked whether it belongs to a specific category (Is this a DAX?). Category membership should not be exclusive, should be hierarchically structured, and could depend upon probabilistically on the presence of features and the co-occurrence and relationship between features. Agent should 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 language ability opens up the domain of tasks that can be performed by the agents. Furthermore, traditional versions of the Turing Test were solely linguistic, which makes it an important skill for intelligent agents. Language grounding is a critical aspect of language acquisition (cf. Landau et al., 1998), and the following series of tests evaluates an agents ability to learn mappings between physical objects or events and the words used to describe them. For each test type, the agent should be shown examples with verbal descriptions, and later be tested on yes-no transfer trials. Brief descriptions of each test type are given below.

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

Property-Adjective 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 (e.g., Smith & Gasser) and typically requires more repetitions to master. This test examines the ability of an agent to learn adjectives, and recognize their corresponding properties in novel objects.

Spatial Relation-Preposition 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.

Action-Verb 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.

Multi-object Action to Relational Verb Mapping. The most complex linguistic structure tested should 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 they relate to events in the visual world.

Connections between tasks

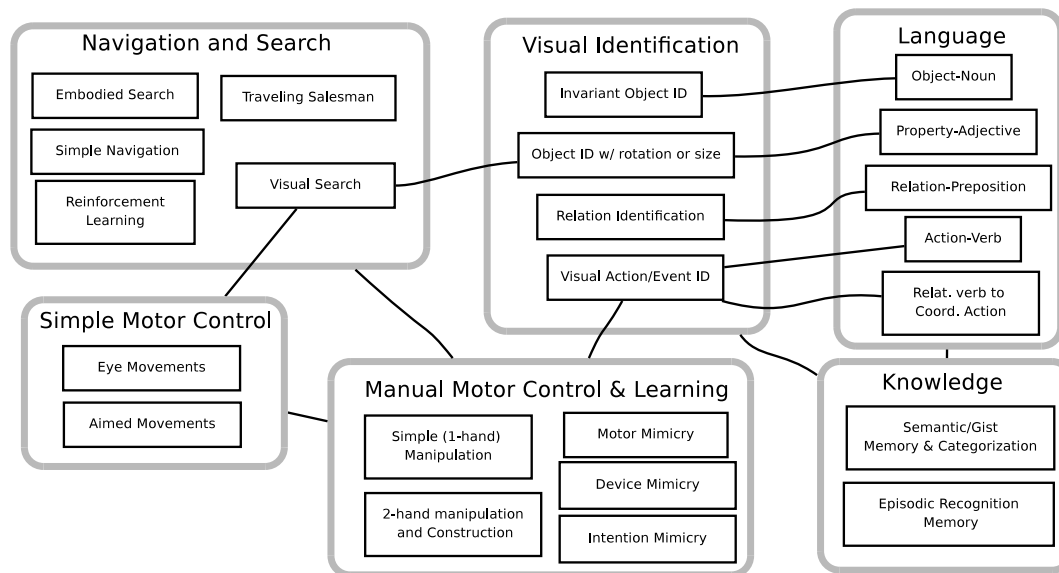
The previous section provided a very elementary description of a set of empirical tasks that we proposed to use for measuring comprehensive embodied intelligence of cognitive agents. Within each group, there are obvious relations between tasks, and many sub-tasks are simply elaborations or variations on other sub-tasks. However, an important aspect of human intelligence is how we use multiple systems together. For example, research on “active vision” has shown the importance of understanding how visual processing and motor control together provide simple accounts of phenomena that appear complex when approached from traditional visual processing perspectives.

Figure 1 depicts some of the strong connections between tasks in different domains. For example, there is a strong correspondence between the visual identification of objects, relations, and events, and the use of linguistic forms such as nouns, adjectives, and verbs. As a result, a strong emphasis was placed on language tasks that were grounded in the environment, or could be used as a means to instruct the agent to perform specific tasks.

The connections between tasks are best illustrated by describing some of the integrated ‘challenge scenarios’ that were also part of the BICA evaluation but not described here. For example, one scenario was called “The Egg Hunt”, and the agent was expected to be able to search for an object in a set of rooms with obstacles. For advanced variations of the task, the agent would be given a verbal instruction describing the object (“Bring me the red basket”). A surprising number of core decathlon tasks would be required

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.

The BICA Cognitive Decathlon



accomplish this fairly simple task. For example, in the language tasks, the agent would have learned the color property red, the name basket, and perhaps the meaning of the word “find”; in the knowledge tasks the agent may have learned the basic shape category of a basket; searching the rooms requires skills tested in the visual search task, embodied search, simple navigation, and the TSP task. To identify the basket, it would draw on skills required for invariant object recognition as well as object identification requiring rotation or size differences. Along with eye movements required to perform visual search, the agent would require at least the skill of simple manipulation, and possibly aspects of motor mimicry and device mimicry if it needed to be taught how to carry a basket.

Biovalidity Assessment

As a complement to the Challenge Scenarios and Cognitive Decathlon, a parallel evaluation plan was developed for the BICA program to assess the degree to which an agent’s cognitive architecture reflects brain-based design principles, computations, and mechanisms. These Biovalidity Assessments are not intended so much as a “Neural Turing Test”, but rather as a means to 1)compel teams to explore neurobiologically-inspired design solutions, and 2) enable comparisons between an agent’s cognitive architecture and that of a mammalian brain. The Biovalidity Assessments are structured to occur in three consecutive stages over the course of a five-year program, with the idea being that teams will continually refine their architectures based on insights from biological comparisons. During this time-

frame, emphasis gradually shifts from evaluations that permit each team to define and test its own claims to biological validity towards evaluations that require all teams to test their architectures against common neural data sets, including functional neuroimaging data recorded from human subjects as they perform Challenge Scenario and Decathlon tasks. The use of common neural data sets is intended to facilitate comparison across teams and to better focus discussion as to which approaches are most successful on certain tasks and why.

Stage 1: Overall Neurosimilitude (Year 1)

Neurosimilitude refers to the degree to which a model incorporates the design principles, mechanisms, and computations characteristic of neurobiological systems. To effectively demonstrate neurosimilitude, teams should describe in detail the mapping of model components to brain structures and comment on the connective topology of their model with respect to that of the brain. Assertions should be backed by references to the neuroscience literature, including both human and animal studies. Teams should not be required to capture neurobiological details at very fine scales (e.g., multi-compartment Hodgkin-Huxley type models); however, to the extent that teams can demonstrate that modeling micro-level details of neural systems contributes to behavioral success beyond what can be accomplished with more coarse-grained models, inclusion of such details should be encouraged.

Stage 2: Task-Specific Assessments (Years 2-3)

Stage 2, Year 2, affords each team the opportunity to compare the activity of their model to data from the existing neuroscience literature in a task-specific context. First, each team should select several cognitive functions, or skills, that feature prominently within one of the Challenge Scenarios or Decathlon events. It would be expected that teams would select those skills/tasks that highlight the biologically inspired capabilities of their own architecture. For instance, a team whose architecture includes a detailed model of the hippocampus might choose a task involving spatial navigation and might choose to show that path integration in their model occurs via the same mechanisms as in the rat hippocampus. Similarly, a team whose architecture employs a temporal differences reinforcement learning algorithm to perform a task might compare prediction error signaling in their model to that reported in neuroscience studies involving similar tasks. It should not be required for teams to perform parametric fits to published data sets; rather, teams should be assessed according to how well their models capture important qualitative features of the neural processes known to support key behaviors. Since, in the first year of Stage 2, teams would select for themselves the sub-tasks against which their models will be assessed, each team would in effect have considerable influence over how its architecture is evaluated.

In Stage 2, Year 3, teams would again compare model performance to existing neuroscience data in the context of the Challenge Scenarios and/or Decathlon tasks. This time, however, all teams should be required to focus on the same set of tasks, which would be selected by the evaluation team. The emphasis on a common set of tasks is meant to facilitate comparison across models and to compel each team to begin thinking about biological inspiration in domains other than those at which their models already excel.

Stage 3: Human Data Comparisons (Years 4-5)

In Stage 3, teams should compare model activity to human functional neuroimaging (e.g., fMRI) data recorded from subjects performing actual Challenge Scenarios and Decathlon tasks. Whereas Stage 2 involves comparisons to previously published neuroscience data, Stage 3 would allow for a more direct comparison between model and neural data, since models and humans would be performing very similar, if not identical, tasks.

To allow for comparisons with fMRI data, teams should generate a simulated BOLD signal using methods of their own choosing and should compare the performance profile of their model to that of the human brain during discrete task elements, with a focus on identifying which model components are most strongly correlated with which brain areas during which tasks, and on how variations in the patterns of correspondence between model and brain activity predict performance across a range of tasks. (For examples of simulated brain imaging studies, see Arbib et al., 2000 and Sohn et al., 2005). Such comparisons would provide a solid empirical platform from which teams could demonstrate the incorporation of neurobiological design principles. Moreover,

it is anticipated that Stage 3 comparisons would generate new insights as to how teams might further incorporate biologically inspired ideas to enhance the functionality of their models.

As in Stage 2, the first year of Stage 3 would require each team to identify several cognitive skills/tasks of their own choosing for which they would demonstrate a compelling relationship between model activity and neural data. Likewise, the second year of Stage 3 would involve a common set of tasks so as to facilitate comparisons across teams. In order to take advantage of access to human brain data, selected tasks would be expected to differentially involve higher-order cognitive faculties associated with human intelligence (e.g., language, symbol manipulation). It is expected that there would be significant methodological challenges involved in parsing and interpreting data from tasks that are as open-ended as the Challenge Scenarios, in which a subject may select from a near infinite repertoire of actions at any point within a continuum of events. However, the risks involved in this approach are outweighed by the potential insights that may be gained from the ability to compare the dynamics of model activity versus human brain activity in the same task environment.

Discussion

This report describes the motivation and design for the “Cognitive Decathlon”, an embodied 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. Although the program for which the test was developed was not funded, it is hoped that this work (1) provides new approach that allows the Turing Test to be useful and relevant for today’s researchers; (2) Suggests a comprehensive set of skills that cover a wide range of embodied cognitive skills; and (3) Identifies ways in which how these core skills are interrelated, providing rationale for tests of embodied intelligence.

References

- Allender, L., Salvi, L., Promisel, D. (1997). Evaluation of human performance under diverse conditions via modeling technology. *In Proceedings of workshop on emerging technologies in human engineering testing and evaluation, NATO Research Study Group 24, Brussels, Belgium, June 1997.*
- 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.

- Bush, R. R. & Mosteller, F. (1951). A mathematical model of simple learning. *Psychological Review*, 58, 313–323.
- 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 acquisition 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.
- Krauzlis, R. J. (2005). The control of voluntary eye movements: New perspectives. *Neuroscientist*, 11, 124–137.
- 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/>
- Rescorla, R. A., & Wagner, A. R. (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement, *Classical Conditioning II*, A. H. Black and W. F. Prokasy, Eds., pp. 64-99. Appleton-Century-Crofts.
- 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.
- Shepard, R & Metzler, J. (1971). Mental rotation of three dimensional objects, *Science* 171, 701–703.
- 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. (2003), *Artificial Stupidity*. Salon, Feb. 2003.
- Sutton, R. S. & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.
- Turing, A. (1950). Computing machinery and intelligence. *Mind*, LIX, 433-460.
- Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, 12, 97-136.
- Wallis, G. & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology*, 51, 167-194.