The BICA Cognitive Decathlon:
A Test Suite for Biologically-Inspired Cognitive Agents

Shane T. Mueller, Ph.D.
Klein Associates Division/ARA Inc.
1750 Commerce Center Blvd
Fairborn, OH 45324
937-873-8166
smueller@ara.com

Matt Jones, Ph.D.
Klein Associates Division/ARA Inc.
and Department of Psychology, University of Texas at Austin
mattj@psy.utexas.edu

Brandon S. Minnery, Ph.D.
Julia M.H. Hiland
The MITRE Corporation
1750 Colshire Drive, McLean, VA
bminnery@mitre.org, jhiland@mitre.org

Keywords:
Artificial Intelligence, Model Evaluation

ABSTRACT: *BICA (Biologically-Inspired Cognitive Architectures) is a DARPA Phase-I program whose goal is to create the next generation of cognitive architecture models based on principles of psychology and neuroscience. This project is motivated by the belief that traditional artificial intelligence research has hit a wall in its quest to develop truly intelligent agents: although agents can be engineered to perform exceedingly well at specific tasks, they are typically quite brittle, unable to deal with unforeseen situations and unable to learn from others. This paper describes the BICA Cognitive Decathlon, Challenge Scenarios, and Biovalidity Assessment, a set of tests we designed to evaluate the performance of such agents in a variety of situations that cover a core set of cognitive, perceptual, and motor skills typical for a two-year-old human child. These include behavioral tasks related to search, navigation, manipulation, memory, language, and three pathways to procedural knowledge: instruction, demonstration, and reinforcement/exploration. The test suite has three distinct components: a set of four integrative challenge scenarios that support the goals of building coherent, systematic, integrated cognitive agents; a set of focused tasks that can better determine the extent to which the core cognitive competencies match the capabilities of humans; and a set of biovalidity assessments to determine the extent to which the agents architecture resembles the human brain. Ultimately, this three-level set of tests was designed to evaluate whether systems are flexible, comprehensive, and taskable in complex situations, while still performing tasks in ways similar to human performers. The test specification and the motivations for and background of individual tasks will be discussed.*

1. Overview

BICA (Biologically Inspired Cognitive Architectures) is a DARPA Phase I project administered through IPTO (Information Processing Technology Office) whose goal is to promote the next generation of artificial intelligence research, motivated by principles of psychology and neuroscience. As with all models in cognitive science modeling, it is important to evaluate and validate the behavior of the models against known behaviors of humans, to determine whether the goals of the modeling effort have been met. This process is normally done on an individual basis by individual scientists, either by showing competence in some domain or by comparing the model's behavior to human data. This creates obvious conflicts of interest, because the researcher is allowed to cherry-pick the best examples and ignore all cases in which the model produces inappropriate behavior. Because the BICA program proposes to support a number of parallel efforts at building AI agents, the task of testing the agents' behaviors was given to an independent team. This report summarizes the evaluation team's proposed test specification, which was sponsored during Phase I of BICA. It was intended to guide the modeling effort during Phase II and serve as the annual gateway test for evaluating progress. At the time of this writing, Phase II of the program has not been funded, and so the test design may not be exercised in its current form. However, future research efforts with similar goals may benefit from the work described here.

Paper read atthe Behavior Representation in Modeling and Simulation (BRIMS) conference, February 2007, Norfolk, VA

2. The Goals of BICA Evaluation

The primary goals of the BICA program are to develop comprehensive biological embodied cognitive agents that could learn and be taught like a human. A wide variety of tasks could be tested in such a program. The decision was made to make serious attempts at embedding the agents in a non-symbolic environment in which perception was done on raw input (unprocessed images and sound streams) and motor control was accomplished through micro-control of effectors. This limits the scope and difficulty of the tasks that could by accomplished in a five year program, and so we designated the target skillset to roughly map onto the capabilities of a two-year-old human child.

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 tasks tested in the BICA program were designed to exercise many of these core skills.

The program 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.

The test specification is 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 neuro-biological processes.

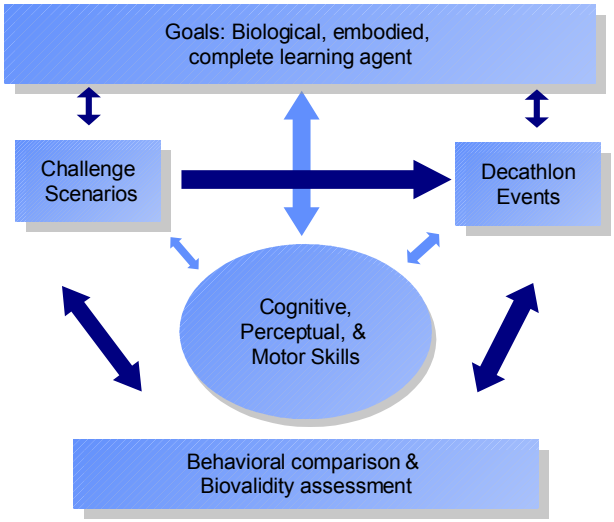

Figure 2.1. Depiction of test specification design. Program goals guide design of challenge scenarios, whose core skills are tested in the Decathlon. Performance on these tests is compared to robust qualitative behavioral and biological phenomena.

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.

Thus, as depicted in Figure 2.1, the program's goals guide the selection of the Challenge Scenarios; the scenarios require core cognitive, perceptual, and motor skills, which are tested in greater detail in the Decathlon Events. Performance in the Decathlon were planned to be compared to robust behavioral and neurological findings of humans.

3. The Challenge Scenarios

Over the proposed five-year scope of the program, versions of four distinct integrative challenge scenarios are planned to be tested. The tasks are designed with increasing levels of complexity, so that initial performance levels can be demonstrated year after year, but the tests maintain utility with sets of increasingly complex challenges presented as the agents attain competency.

Task Name	Description
Object Search Task	Search and object recognition; learning through instruction
Observational Language & Procedure Learning Task	Manipulation, action-object-language mappings; observational learning
Self-directed Search and Construction Task	Self-directed exploration and learning, search, construction, goal inference.
Open-ended Tasking	Open-ended taskability

Table 3.1 Overview of BICA Challenge Scenarios.

The Challenge Scenarios are designed to provide integrative tasks for embodied, learning agents to perform. In this section, we describe basic versions of each challenge scenario. These are, for the most part, novel tasks designed specifically for this program. Past empirical data on human subjects exists for some versions of some of these tasks, but we anticipated collecting new data on human subjects to compare to the agents' performance.

3.1 The Object Search Task

This task is designed to test navigation and search ability, together with the ability to learn through declarative instruction and from episodic memory. The task takes place in a connected series of rooms, and begins by showing the agent a probe object. The agent is asked to find and retrieve a copy of that object from somewhere in the environment. Verbal hints and constraints will be given to the agent to help guide search, and the agent should use knowledge of environment gained during earlier trials to guide later search paths.

3.2 Observational Language and Procedure Learning

This task is designed to test the agent's ability to manipulate its effectors and other objects by observational learning, and to learn the language constructs that describe these objects and events. The task takes place in a room with an instructor and a number of manipulable objects. The instructor will perform construction or manipulation tasks while describing them in words. The agent will be instructed to perform that task (or an earlier learned task) and be given feedback on its success. As language production skills improve, agents will be asked to describe actions it or the instructor is performing. Tasks could range from simple object-action events ("I am dropping the ball") to object construction ("I am building a tower."), coordinated action ("I am hitting the cup with the hammer."), and complex compound events. ("I am sweeping the floor").

3.3 Self-directed Search and Construction Task

An important aspect of human intelligence is the self-directed ability to explore the environment and learn from it. Yet most problems AI systems face are well defined with clear goal. Perhaps a more difficult problem is discovering these goals in the first place. This task attempts to replicate the notion of goal discovery by generating an environment populated with an ecology of rewards and punishments. The agent must explore the environment and discover useful behaviors on its own, or by observing other intelligent agents operating in the same environment. The task involves the construction of multi-component objects which can be redeemed for reward once completed. Components of the objects will be distributed in systematic ways probabilistically through the environment, and the agent will receive a reward when redeemed. Some objects may be easy to construct but produce small rewards; others may be difficult to construct be produce larger rewards. This task uses aspects of the Object Search Challenge and the Observational Language and Procedure Learning Task, but also requires an additional level of self-directed exploratory behavior that is adjusted through reinforcement to generate more valuable behaviors. are planned to actions, the agent should discover which constructed objects are more valuable, both in terms of their reward and the work required to find and construct them.

3.4 Open-ended tasking

Previous challenge scenarios offer agents opportunities to demonstrate the ability to learn in narrow domains of performance. Yet the flexibility of human cognition has substantial breadth in its ability to accomplish a wide range of tasks. This task extends the skill repertoire beyond the narrowly-defined tasks, and can contain any

task that can be taught through instruction, demonstration, and feedback. Such tasks will attempt to demonstrate the true flexibility and performance capability of the agents, and will hopefully allow the discovery of unique and serendipitous solutions. Possible tasks could include things like: playing simple games, sorting objects according to different rules, tasks requiring coordinated action with the teacher, the construction of novel artifacts, the learning of simple action-command associations, etc.

4. The BICA Cognitive Decathlon

Like the Olympic Decathlon, the BICA "Cognitive Decathlon" is 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.

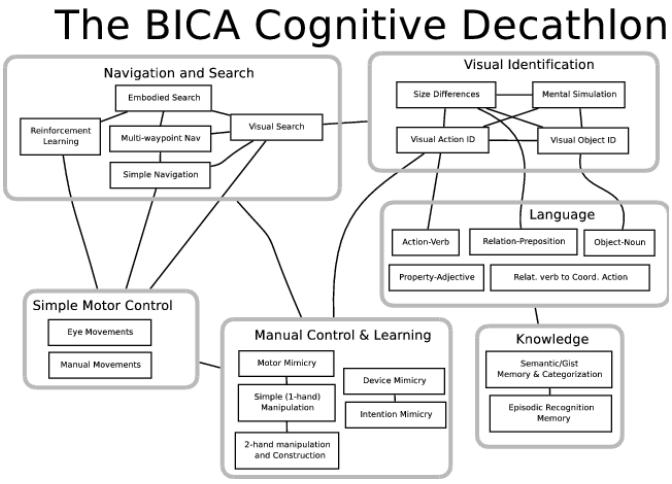

Figure 4.1. Graphical depiction of the BICA Cognitive decathlon. Grey rounded boxes indicate individual tasks that require the same basic procedural skills. Black rectangles indicate individual trial types or task variations. Lines indicate areas where there are strong relationships between tasks.

Table 4.1. The BICA Cognitive Decathlon Tasks

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
	Travelling Salesman Problem
	Embodied Search
	Reinforcement Learning
3. Manual Control & 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 & Concept Learning	Object-Noun Mapping
	Property-Adjective
	Relation-Preposition
	Action-Verb
	Relational Verb-Coordinated Action
6. Simple Motor Control	Eye Movements
	Aimed manual Movements

4.1. 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:

4.1.1. 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,

4.1.2. 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.

4.1.3. 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.

4.1.4. 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.

4.2. 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.

4.2.1. 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).

4.2.2. 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.

4.2.3 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.)

4.2.3. 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.

4.2.5. 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.

4.3. Manual Control & Learning

Along with visual and navigational skills, the agents will have ability to control its arms and graspers in order to manipulate the environment. Initial simple control of these effectors will be tested in the Simple Motor Control test (Section 4.6.3). This event incorporates for levels that go beyond simple control.

4.3.1. 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 ego-centric 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.

4.3.2. 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.

4.3.3. 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.

4.3.4. 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.

4.3.5. 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.

4.4. Knowledge Learning

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

4.4.1. 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.

4.4.2. 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.

4.5. 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.

4.5.1 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.

4.5.2. 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.

4.5.3. 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.

4.5.4. 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.

4.5.5. 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.

4.6. 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.

4.6.1. 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.

4.6.2. 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.

4.6.3. 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.

5. Biovalidity Assessment

As a complement to the Challenge Scenarios and Cognitive Decathlon, which are behavioral tests, a parallel evaluation framework was designed to ensure that the BICA program achieves its goal of developing models that incorporate brain-based design principles, computations, and mechanisms. These Biovalidity Assessments are structured to occur in three consecutive stages over a five-year period. During this timeframe, emphasis gradually shifts from evaluations that allow each team to define and test its own claims to bio-inspiration (thereby accommodating the diversity of approaches among different teams) towards evaluations that require all teams 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.

5.1 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 are to describe in detail the mapping of model components to brain structures and to comment on the connective topology of their model with respect to that of the brain. Assertions are to be backed by references to the neuroscience literature, including both human and animal studies. Although neurosimilitude could potentially encompass levels of detail ranging from single ion channels to cortical microcircuits to large-scale networks, BICA primarily seeks biological validity at the level of the brain's large-scale functional architecture (for example, cognitive control networks that recruit multiple neocortical and subcortical areas during task performance). Teams are not required to model neural information processing at finer scales; however, it is understood that principles and mechanisms operating at

one scale can enable functions at a larger scale, and that incorporating biological detail can lead to computational value added in surprising ways. To the extent that teams can demonstrate that modeling microcircuit-level details of neural systems contributes to behavioral success beyond what can be accomplished with more coarse-grained models, inclusion of such details is encouraged.

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

Stage 2, Year 2, affords each team the opportunity to compare the activity of their model, in a task-specific context, to data from the existing neuroscience literature. First, each team selects several cognitive functional domains, or skills, that feature prominently within a to-be-specified Challenge Scenario or Decathlon event. In some cases these domains might align with discrete subtasks within the Challenge Scenario/Decathlon event, or they might apply more generally across a range of episodes within the overarching task. It is expected that teams will select domains/subtasks 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 subtask 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 model develops the ability to perform a subtask via a temporal differences reinforcement learning algorithm might compare prediction error signaling in their model to that reported in neuroscience studies involving similar classes of tasks. It is not essential that teams perform a parametric fit to published data sets; rather, teams are to be assessed according to how well their models capture important qualitative features of the neural processes known to support key behavioral capabilities. Since, in the first year of Stage 2, teams are selecting for themselves the subtasks against which their models will be assessed, each team in effect has considerable influence over how its architecture is evaluated.

In Stage 2, Year 3, teams again compare model performance to existing neuroscience data in the context of the Challenge Scenarios and/or Decathlon tasks. This time, however, all teams are required to focus on the same set of subtasks, which are to be pre-selected by DARPA. 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.

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

In Stage 3, teams are to compare model activity to human functional neuroimaging (e.g., fMRI) data recorded from subjects performing actual BICA Challenge Scenarios and Decathlon events. Whereas Stage 2 involves comparisons to existing neuroscience data sets from the literature, Stage 3 allows for a more direct comparison between model and neural data, since models and humans will be performing very similar, if not identical, tasks.

To allow for comparisons with fMRI data, teams will generate a simulated BOLD signal using methods of their own choosing and will 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 classes of 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 are intended to provide a solid empirical platform from which teams can demonstrate the incorporation of biologically inspired principles, mechanisms and processes. Moreover, it is anticipated that Stage 3 comparisons will 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 requires each team to identify several subtasks/cognitive skill domains of their own choosing for which they will demonstrate a compelling relationship between model activity and neural data. Likewise, the second year of Stage 3 involves a common set of subtasks so as to facilitate comparisons across teams. In order to take advantage of the fact that fMRI techniques allow for access to *human* brain activity, selected subtasks are expected to differentially involve higher-order cognitive faculties associated with human intelligence (e.g., language acquisition, symbol manipulation). It is expected that there will 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 – on a subsystem-by-subsystem basis – the dynamics of model activity versus human brain activity as recorded in the same task environment.

6. Discussion

Alan Turing (1950) famously described a test for assessing artificial intelligence, in which a machine would be considered intelligent if its behavior cannot be distinguished from a human's. Interestingly, many of today's intelligent and robotic systems would fail this test because they perform the task so much better that the humans they replaced. For example, no librarian could

reproduce the breadth and speed of Google's knowledge retrieval, and no human assembly worker can rival the laser-guided accuracy and consistency of an industrial robot. Yet when one considers the flexibility that humans exhibit, no machine (and not even all machines put together) can currently come close.

The BICA Project is an ambitious attempt to promote the development of artificial intelligence that goes beyond the current failings of such systems. Current systems are often engineered to perform specific tasks well, whereas humans have evolved to be good at a wide range of tasks. Current systems are brittle and do not handle situations that were not anticipated; humans typically cope well with such situations and learn from them for future performance.

Despite the fact that humans are flexible learners, cognitive AI systems are typically evaluated in terms of the few specific situations they have been engineered to handle, and cognitive science models are typically targeted to a single experimental paradigm. For these types of evaluation, the best way to distinguish between models requires advanced statistical techniques that simultaneously examine model complexity and the complexity of the data set (cf. Myung, 2000 for a review of statistical techniques, and Gluck & Pew, 2005, for a summary of an effort to evaluate complex cognitive architectures). These techniques often punish models that are more flexible, because they can typically fit a greater variety of data. Although this is justified when dealing with a fine-grained model of a psychological process, the approach may backfire as models grow in complexity and become capable of performing a wider variety of tasks.

A central technique in model evaluation has always been generalizeability. This is often used to ensure models do not "over-fit" the data (cf. Busemeyer & Wang, 2000), but we advocate that it is especially well suited to evaluating AI models in cognitive science because of the flexibility humans repeatedly demonstrate. The test specification described here is motivated by this notion. For this program, if given the choice between two models: one that predicts human performance in a few tasks really well, and another that predicts robust qualitative phenomena in more tasks but with less accuracy in each task, the choice is clear—we prefer the more flexible model. We have designed this set of tests to evaluate this type of generalizeability. So, rather than requiring reasonable quantitative fits to new participants or different versions of a narrowly-scoped task, we require qualitative prediction of robust phenomena in new situations. We hope that the types of models that the BICA program hoped to encourage will soon be able to demonstrate this type of flexibility by performing tasks the model was never designed to perform, and tasks it learns to perform on its own.

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.

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.

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.

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.

Author Biographies

Dr. Shane T. Mueller is a Senior Research Scientist at Klein Associates Division of A.R.A. Inc. He received his Ph.D. in Cognitive Psychology at the University of Michigan. His research interests include model testing, and computational modeling of human memory and perceptual processes.

Dr. Matt Jones received his Ph.D. in Cognitive Psychology from the University of Michigan in 2003. He is currently a post-doctoral fellow at the University of Texas and a part-time researcher at Klein Associates Division of A.R.A., Inc. His research interests include mathematical and computational modeling of human learning and decision making.

Dr. Brandon Minnery is a Senior Artificial Intelligence Engineer in the Emerging Technologies Office at the MITRE Corporation. Dr. Minnery holds a B.S. in Physics from the University of Cincinnati and a Ph.D. in Neurobiology from the University of Pittsburgh. His early research examined the processing of tactile sensory information by neural circuits within the mammalian brainstem and thalamus. While at the MITRE Corporation, Dr. Minnery has provided technical support and guidance to several government agencies that fund neuroscience-related research programs, in particular programs that seek to apply neuroscience-inspired design principles and mechanisms to the development of novel approaches to computing, artificial intelligence, and human-computer interface technologies.

Julia Hiland is an Artificial Intelligence Engineer in the Emerging Technologies Office at the MITRE Corporation. She has a B.S. in psychobiology from SUNY Binghamton, where she examined developmental changes in the addictive properties of alcohol. Her ongoing Ph.D. work at University of California Merced focuses on the role of learning and memory in the formation and use of conceptual categories. During her time at the MITRE Corporation, she has supported government programs that emphasize on neuroscience, including programs that seek to apply neuroscience principles to traditional artificial intelligence methods, and programs that focus on using neuroimaging techniques to understand neural processes that underlie learning and cognition.