International Journal of Machine Consciousness
Vol. 2, No. 2 (2010) 273–288
© World Scientific Publishing Company
DOI: 10.1142/S1793843010000497

# A PARTIAL IMPLEMENTATION OF THE BICA COGNITIVE DECATHLON USING THE PSYCHOLOGY EXPERIMENT BUILDING LANGUAGE (PEBL)

SHANE T. MUELLER

*Klein Associates Division, Applied Research Associates, Inc.,*
*1750 Commerce Center Blvd, Fairborn OH 45324, United States*
*smueller@ara.com*

The Cognitive Decathlon is a proposed set of tasks that can be tested on both human and artificially intelligent agents, and which constitutes a modern specification for the Turing Test. In this paper, a partial implementation of the Cognitive Decathlon is described using the Psychology Experiment Building Language (PEBL). The tasks focus not simply on generic human abilities, but on critical skills that highlight aspects of human performance that are at odds with common artificial intelligence approaches. The differences between human and algorithmic behavior in such tasks can reveal properties of the human cognitive architecture, and production of similar behavior by artificial systems can help constrain and validate biologically-inspired systems.

*Keywords*: Turing Test; Cognitive Decathlon.

## 1. Retrospective and Reflection on the BICA Cognitive Decathlon

The Cognitive Decathlon was developed as a test plan during the initial stage of DARPA's Biologically-Inspired Cognitive Architectures (BICA) program. Although a full research program was not funded following that initial stage, the effort nonetheless fostered numerous collaborations that have already begun to bear fruit. The goals of the original Cognitive Decathlon were centered on the particular needs and motivations of BICA (see Mueller *et al.* [2006] for a detailed description), which aimed to develop embodied biologically-inspired cognitive agents and robots. Thus, the test was targeted toward a somewhat lower skill level than had it been designed for no-holds-barred artificial intelligence, see [Mueller and Minnery, 2008], but it still encompassed a broad and useful range of target skills. Although the original effort did not result in testable empirical methodologies, it assembled a plan for testing a broad set of skills that could serve as a target for future researchers developing general human-level artificial intelligence.

In the time since, subsequent researchers have suggested ways in which the Decathlon might be revised using factor-analytic approach to human intelligence

[Simpson and Twardy, 2008]. In addition, research by myself and my collaborators has extended the notions of the BICA Cognitive Decathlon to evaluate human cognitive performance over a range of skills [Mueller *et al.*, 2009; 2010]. Each of these efforts have reconceptualized the Cognitive Decathlon in the domain of psychological testing and test battery construction. However, the Cognitive Decathlon does not make for a complete cognitive test battery that can be used to assess human skill, and available test batteries (or even test "armories", see Hunt [1990]; O'Donnell *et al.* [2005]) would not necessarily make for a reasonable Cognitive Decathlon. Another step is needed that goes beyond simply identifying a broad set of competencies. One must also identify behavioral measures that can help distinguish biological from non-biological intelligence. Then, to the extent that an agent produces behavior that is indistinguishable from a human's, it demonstrates similar intelligence. Thus, the logic of the Decathlon is identical to that of a traditional concept for measuring artificial intelligence: the Turing Test.

## 2. Cognitive Decathlon: A Modern Version of the Turing Test

Testing artificial intelligence has both similarities to and important distinctions from testing human intelligence. Modern psychometric theory has helped researchers develop many scales and metrics (see Camara *et al.* [2000]) that are used to assess the psychological capabilities and intelligences of children and adult humans. The rationale of much of this work has been to discriminate amongst different strata of people, so that one might perhaps know who should be selected for training, treatment, education, or therapy, often focusing on individuals at the extreme ends. This makes human intelligence measures inadequate for testing artificial intelligences, because for the most part, an artificial intelligence that reaches even low-functioning levels of general human intelligence would be a substantial accomplishment.

But just as modern intelligence testing has a goal of separating low intelligence from high intelligence, the goal of the BICA Cognitive Decathlon was to separate simulated intelligent behavior with biological underpinnings from simulated intelligent behavior with non-biological bases. This goal is different from psychological testing because for many well-framed problems, general machine intelligence solutions can often outperform human solvers, both in time and in quality. But this performance often comes at a cost, so that typical machine solutions have limited ability to generalize beyond the problems they were designed to solve.

At first, Turing's [1950] Test seems to offer little for this goal. He suggested that a reasonable test for artificial machine intelligence is to compare the machine's behavior to a human (who we can agree is intelligent), and if the machine's (primarily verbal) behaviors and interactions over an extended period of time (maybe years) is indistinguishable from a human's, the machine might as well be considered intelligent (or we must admit that humans are not intelligent). Turing's original proposal was limited to verbal interactions alone, and this is how the test is typically interpreted in common usage. However, testing biologically-inspired intelligence may go beyond verbal intelligence, since embodied perception and action are critical aspects of this type of intelligence. Thus, a Verbal Turing Test (VTT) is not an appropriate test of biologically-inspired artificial intelligence.

However, Harnad [1989; 1990; 2000; 2004] argued that an embodied version of the Turing Test is consistent with Turing's original thought experiment. Furthermore, Mueller and Minnery [2008] argued that in addition to the domain (whether verbal or embodied), there are two additional aspects of the Turing Test that can be modified or relaxed to form a test relevant for assessing biologically-inspired artificial intelligence. Together, these three parts include: (1) the domain (as argued by Harnad); (2) the measure of similarity, and (3) the target intelligence.

Just as Harnad [2004] argued that Turing did not require verbal-only communication (and so an embodied Turing Test is possible), Mueller and Minnery [2008] argued that the intelligent target is left unspecified by Turing, as are the measures of similarity. So, once a domain is specified, one could conceive of a test comparing an artificial intelligence to the skills of a dog, a child, an average adult, or a genius or superstar in a domain. As long as we can come to an agreement that the target is intelligent, we do not need to define intelligence and simply have to examine whether performance of the artificial system is indistinguishable from the target. Yet once a target is specified, one also needs to specify a measure of similarity: what is meant by "indistinguishable". The most common measures used AI amongst researchers is probably one of *competence* (can the artificial intelligence accomplish a task?), but to assess biological validity, one probably needs to use a stronger measure, in which at least robust data trends are reproduced. Turing's version of the test was incredibly ill-specified in its measure of similarity, and the test is often viewed as one based on fooling or confusing a judge about the identity of real versus artificial intelligence. However, even this intuitive judgment is a measure of similarity. In contrast, I argue that the critical aspect of the Turing Test is *not* the ability to fool a judge, but rather to provide a given standard for measuring behavior, and determining whether there is a difference between human and artificial agent.

These notions provided guidance for the design of the original BICA Cognitive Decathlon. First, it is a broad set of tasks (like the Olympic Decathlon) scoped to mostly be simple enough for a target intelligence of a human child to perform, ranging over a broad set of input and output modalities. In addition, tasks were chosen so that performance which reproduced robust trends of human performers would provide insight into whether the underlying implementations faithfully replicated aspects of biological intelligence.

The proposed taxonomy of tasks in the Cognitive Decathlon is shown in Table 1. We organized these into basic categories or taxons, which are similar to a number of past taxonomies of human skill (see [Fleishman *et al.*, 1968; Fleishman, 1975; Allender *et al.*, 1997; Parks, 2006]). The similarity of our approach to these past taxonomies suggests that we have captured the same breadth of human intelligent behavior that others have considered important.

Table 1. Component tasks of the Cognitive Decathlon.

| Task | Level | PEBL Implementation |
|---|---|---|
| 1. Vision | Invariant Object Identification | Object Judgment Test |
| | Object ID: Size Discrimination | Object Judgment Test |
| | Object ID: With Rotation | Object Judgment Test |
| | Object ID: Relations | |
| | Visual Action/Event Recognition | |
| 2. Search | Simple Navigation | Trail-Making Test |
| | Visual Search | Visual Search Test |
| | Traveling Salesman Problem | Traveling Salesman Problem |
| | Embodied Search | |
| | Reinforcement Learning | Bechara's Gambling Task |
| 3. Manual Control and Learning | Motor Mimicry | |
| | Simple (One-Hand) Manipulation | Compensatory Tracking |
| | Two-Hand Manipulation | |
| | Device Mimicry | Device Mimicry Task |
| | 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-Action | |
| 6. Simple Motor Control | Eye Movements | |
| | Aimed Manual Movements | Aimed Movement Test |

The Decathlon tasks may be useful for researchers in artificial intelligence in a number of ways. First, few researchers in cognitive science look across broad sets of skills, either from an empirical or modeling perspective. Thus, the Decathlon may provide guidance to AI specialists about the breadth of human skill. Second, the Decathlon aims to identify behavioral phenomena that are performed in ways fundamentally different from current AI approaches. These differences may reveal important ways in which artificial systems could be improved. Third, the Turing Test has become associated with futile attempt to build chatbots. In contrast, the Decathlon provides a very specific implementation of the Turing Test, parts of which may be useful in much broader research communities. And finally, insights into human and artificial task performance are sometimes required for careful observation and behavioral study, both on oneself and others. Thus, implemented versions of the Cognitive Decathlon may help facilitate understanding in these areas, as it could provide useful data sources for comparison to artificial agents.

Even if fully implemented tasks were available, I would not envision the Cognitive Decathlon forming the basis for competition akin to the Robocup robotic soccer competition. Rather, it serves as a useful concept for thinking about human-level intelligence across a broad range of capabilities, and it identifies particular skills that were believed to be especially critical for discriminating between biologically-inspired

AI and other systems. Yet there is value in having reference implementations of these tasks. The rest of the paper will describe a subset of the Cognitive Decathlon that is implemented and available for researchers to use, either to gain better introspective insight into their own performance of a task, or to collect data from human subjects to test hypotheses about how humans accomplish the task.

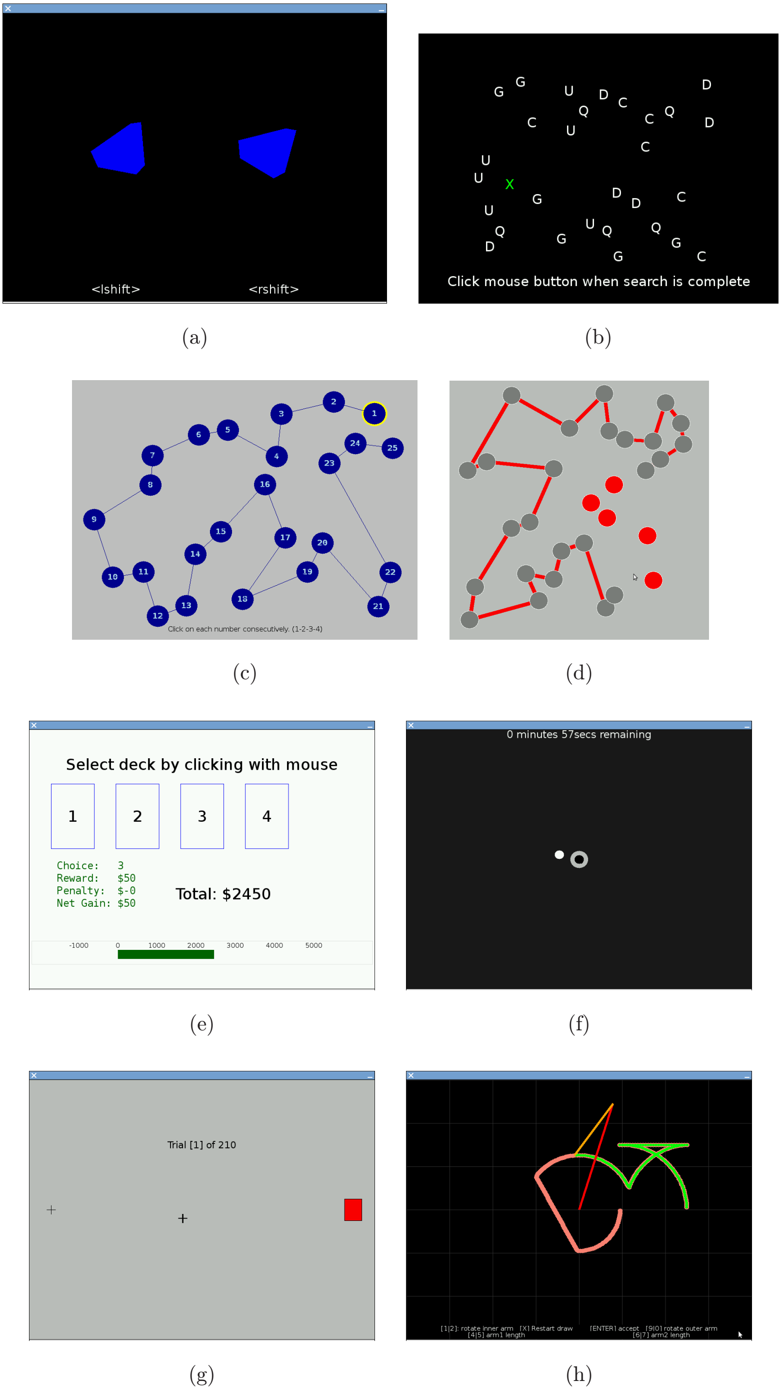
## 3. Implementation of the PEBL Cognitive Decathlon

Although the BICA Cognitive Decathlon was never implemented or tested, subsequent research by myself and my colleagues has continued to develop and test many of these same ideas. One central research project has involved developing and testing the impact that heat strain places on a broad range of cognitive skills [Mueller *et al.*, 2009]. Furthermore, parts of the Cognitive Decathlon incorporate fairly common tasks used frequently in general cognitive, aptitude, and occupational skills testing (e.g., [Perez *et al.*, 1987]).

In the time since the BICA program, I have implemented many tests suggested by or similar to those proposed in the BICA Cognitive Decathlon, using the Psychology Experiment Building Language (PEBL, [Mueller, 2009]). PEBL is cross-platform computer programming language designed for running computer-based psychology experiments. Its source code and executables are available for free, and researchers can use it to design and run experiments, and share them freely with one another. Although it currently operates only as a human testing platform, long-range plans for the system include enabling its use as an artificial experiment lab that can drive the task environment of AI systems. As part of the PEBL Project, I distribute a set of common psychological tests called the PEBL Test Battery. The test battery consists of close to 50 common psychology experiment paradigms, many of which map onto the tests proposed for the Cognitive Decathlon. I will next describe a subset of these tasks that I will refer to as the PEBL Cognitive Decathlon, which are tests that have been implemented in PEBL and are close or exact matches to the tests proposed in the BICA Cognitive Decathlon. Software archives of the PEBL Cognitive Decathlon are available from the author and the PEBL website (http://pebl.sourceforge.net). Roughly eight tasks implemented so far cover several areas of the original Cognitive Decathlon. Screenshots of these tasks are shown in Fig. 1.

### 3.1. *Visual identification*

The ability to identify visual aspects of the environment is a critical skill used for many tasks faced by humans. The Cognitive Decathlon proposed to test this skill with a graded series tests that determine if an agent can tell whether two objects or events are identical. Detailed tests of object recognition capabilities are important because the machine vision community has developed many highly successful algorithms that are not inspired by biological structures. Consequently, tasks that humans find simple may be difficult for generic machine vision algorithms, and tasks that humans find difficult may be simple.



Click mouse button when search is complete

Click on each number consecutively. (1-2-3-4)

Select deck by clicking with mouse

Choice: 3
Reward: $50
Penalty: $-0
Net Gain: $50

Total: $2450

0 minutes 57secs remaining

Trial [1] of 210

[1|2]: rotate inner arm [X] Restart draw [ENTER] accept [9|0] rotate outer arm
[4|5] arm1 length [6|7] arm2 length

(a) (b)

(c) (d)

(e) (f)

(g) (h)

Fig. 1. Screenshots from six of the implemented tasks described in the text. (a) Object judgment task; (b) Visual search; (c) Trail-making test; (d) Traveling salesman problem. (e) Bechara's Gambling task; (f) Compensatory tracker; (g) Aimed movement task; (h) Device mimicry.

For example, for humans, the size and orientation of an object can be used to help make discriminations (because they can be useful in judging distance and approach angles), but they also need to be ignored frequently (because the visual size and orientation of an object can change rapidly as one approaches it). For humans, identifying small differences in an object's size or orientation can be very difficult. In contrast, many machine vision systems require rectified and scaled objects as input, as even tiny differences between stimuli can be discriminated (see Latecki *et al.* [2005] for shape-based alternatives that are more robust).

Several tests within the broad PEBL Test Battery involve pattern recognition skills, including a pattern comparison test, a delayed match-to-sample test, a matrix rotation test, and the polygon rotation test. In addition, the Object Judgment Task (Panel A of Fig. 1) was designed to implement the graded series of tests proposed in the BICA Cognitive Decathlon. This test briefly displays a randomly-generated Attneave shape (see Attneave and Arnoult [1956]; Collin and McMullen [2002]), and after a brief interval, the participant must choose which of two alternative shapes was presented: the original or a slightly altered foil.

Foils are created in three ways; either by rotating the shape 5, 10, or 20 degrees, by scaling the shape up or down by 5, 10, or 20%, or by perturbing 1, 2, or 4 points up to ten pixels in any direction. Figure 2 shows example accuracy results (50% was chance) from one human subject in these conditions. For object rotations there is an area of indifference for small degrees of rotation where accuracy is only slightly above chance. These rotations are large enough to detect when compared directly, but not large enough to remember after a brief interval. Similarly for scaling, there is a fairly large range (roughly 90 to 110%) in which changes are detected relatively poorly. In

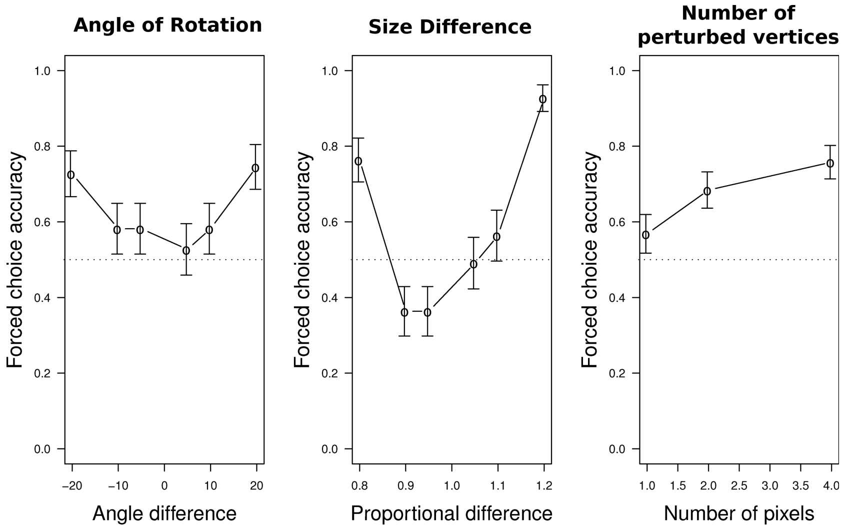


Fig. 2. Accuracy for forced-choice judgments about options where target and foil differed along three dimensions (angle of rotation, proportional size, or number of vertices perturbed).

addition, a response bias favoring the smaller target was present in the data shown. Finally, trials in which foils have perturbed vertices produce higher accuracy as more vertices are perturbed.

Another widely-reported phenomena for object comparison is termed the "mental rotation" effect. In comparisons between an object and a rotated comparison which may be identical or a mirror image, the time needed to choose is highly correlated with the angle of rotation [Shepard and Metzler, 1970]. This implies that comparison involves an analog visual rotation process, rather than a propositional coding of object parts that might be done by an artificially intelligent system. This robust behavior has potential as an important test for biological vision, and indeed forms the basis of some biologically-inspired models [Goebel, 1990]. However, it may be a mistake to view mental rotation as a core primitive function of the vision architecture, because a substantial number of studies have failed to find mental rotation effects when the comparison tests were not mirror images (see Förster *et al.* [1996], for a review). Furthermore, Förster *et al.* [1996] argued that the rotation strategy is dependent on the difficulty of the comparison, and showed the time taken to make difficult direct comparisons between two objects increases as the rotation increases, and as the objects become more similar. This is consistent with the absolute judgment effects discussed earlier, because it means that human vision is blind to small changes, and must then rely on deliberate comparisons to make the discrimination.

Consequently, a second version of the object judgment task was designed to identify the robust trends related to object comparison under rotation and transformation. In this test (modeled after those described by Larsen [1985]), a target polygon is presented, and after a blank delay a transformed polygon is shown. The new shape is either the same original shape or one with perturbed vertices, but always at a new angle of rotation or scale. This test differs from the previous task (which had participants identify the object that was identical to the standard in every way) because it required the participant to determine whether the comparison object had an identical shape, ignoring orientation and size. In conditions where mirror-image shapes are compared, rotation effects can be seen; for most other conditions, flat comparison times are obtained across angle of rotation and size ratio. These effects establish ways in which object recognition is essentially invariant to large changes in orientation and size, and small distortions in the object itself.

These basic object recognition tests embed several critical comparisons set forth in the original Cognitive Decathlon. More complex comparisons were also proposed (e.g., multi-component objects, and dynamic scenes), and these tests will require further research and experimentation to identify the robust effects that highlight how human biological vision systems excel and fail at such identification tasks.

### 3.2. *Search and navigation*

A critical set of skills for an embodied agent is the ability to navigate through and learn about its environment. Search and navigation form a fundamental cognitive skill set used by lower animals and adult humans alike. Yet many automated search and navigation systems employ optimization techniques or require GPS navigation or terrain databases to succeed. In contrast, humans typically do not require these external navigation tools, although we can benefit from them. Thus, search and navigation tasks can be useful in discriminating biological from non-biological spatial reasoning systems. A graded series of tests in the BICA Cognitive Decathlon was proposed to test these abilities, including Visual Search, Navigation, and the Traveling Salesman Problem. Several of these tasks, which will be described next, have been selected for the PEBL Cognitive Decathlon.

**Visual search:** A core skill required for many navigation tasks is the spatial localization of a known target. In many respects, human visual search may resemble typical computer-based visual search algorithms, but one frequent difference is that human search methods can depend on the similarity of the cue to the background. For different combinations of target and background, either parallel or serial search may be exhibited (cf. [Treisman and Gelade, 1980]). In contrast, computational algorithms might always perform parallel search to identify patterns (for example, comparing how well a template matches each possible screen coordinate) or might always perform serial search matching a template to critical locations, but may not produce the proper mix of serial and parallel search exhibited by human vision. To do so requires the presence of global feature-matching systems that can perform parallel detection of certain feature classes, together with focal visual attention and eye movement theory (e.g., Wolfe *et al.* [1989]; Itti and Koch [2001]). Consequently, biologically-inspired vision systems should exhibit both serial and parallel visual search in at least a few critical comparison conditions. Example conditions include: (1) search for a target with a unique color amongst distractors. This should exhibit pop-out: flat response times as the number of distractors are increased; (2) search for a target with a unique feature (curved versus horizontal/vertical line segments). This should also produce pop-out; (3) search for conjunctions or configurations of features (color-orientation combinations; L versus T, etc.). This should produce serial search (increasing response time as number of distractors increase). A flexible visual search task implemented in PEBL (Panel B of Fig. 1) allows one to specify and test each of these conditions.

**Navigation and route planning:** The original Cognitive Decathlon proposed a task which would assess the efficiency of simple navigation tasks: direct and indirect movement to a target in the presence of obstacles. This was intended as a simple demonstration of basic route-planning capabilities that nonetheless presents challenges to modern robotic systems (e.g., [Best, 2004]). Study of simple route navigation is probably best left to tests in physical environments, and so no specific tests of this capability are present in the PEBL Cognitive Decathlon. However, some related tasks include the PEBL trail-making task (Panel C of Fig. 1), a version of Reitan's

[1958] trail-making task. This task is a classic and frequently-used test of spatial navigation, in which the participant must "connect the dots", following a prespecified path through the locations on the screen.

A related but more complex task is the route-planning task called the "Traveling Salesman Problem" (TSP). The task involves finding an efficient path through multiple locations on a plane. 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 presents an interesting challenge for problem-solving approaches that rely on search through a problem space (a hallmark of traditional AI approaches). 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 very 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 humans solve the task in ways fundamentally different from traditional AI approaches. 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.

A reference implementation of the TSP was developed for the PEBL Cognitive Decathlon (Panel D of Fig. 1). Basic results from a study using this test are shown in Fig. 3. Consistent with past research, solution times were linear with problem size, whereas inefficiencies remained between 5−10% of optimal, growing as the problem size increased. The solid and dashed lines in the figure represent two conditions in the experiment which produced no reliable impact on performance, and which are irrelevant for the current presentation.

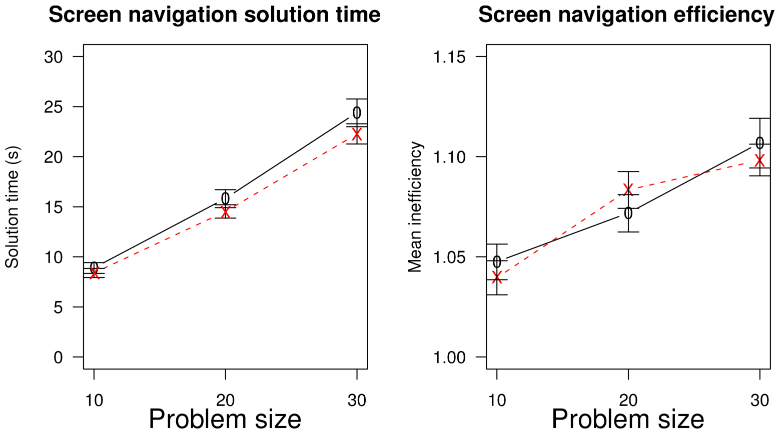


Fig. 3. Sample results of human solutions to the Traveling Salesman Problem. Left panel shows solution's time (which are linear with problem size). Right panel shows solution inefficiency, which increases slightly with problem size, but remains between 5−10% longer than optimal.

**Reinforcement learning:** The proposed 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. The original Cognitive Decathlon proposed a test that resembled N-armed bandit (e.g., Sutton and Barto [1998]; Rescorla and Wagner [1972]) or the Iowa Gambling Task [Bechara *et al.*, 1994].

The Iowa Gambling task is a simple option selection task in which a monetary or point-based reward or punishment is given after each choice. Each of four piles has a different reward schedule, two with a long-term net gain, and two with a long-term net loss. Each pair of gain/loss decks also differ in their variability, with either fewer numbers of larger loss cards or more cards with smaller losses. Typically humans converge on the net gain decks after a few dozen trials, although some patients with specific brain injuries persevere on net-loss decks. In its basic form, the task may not discriminate well between human and artificial intelligence, but by systematically varying payoffs and reward probabilities, one can test whether the artificial systems reproduce several of the robust decision-making behaviors of humans in this area, such as loss aversion, contextual effects of unchosen alternatives, and anchoring effects (see [Busemeyer and Johnson, 2004]). A version of the Iowa Gambling Task (called the "Bechara Gambling Task", Panel E of Fig. 1) is included in the PEBL Cognitive Decathlon.

### 3.3. *Motor control and learning*

The BICA Cognitive Decathlon specified a number of tests that evaluated simple and complex motor control, including eye movements, manual tracking, aimed movements, and object manipulation, and higher-order learning of more abstract use of devices. Several of these tasks have been implemented for the PEBL Cognitive Decathlon, allowing behavioral measures of human performance to be made.

**Manual tracking:** Simple motor tracking requires tight coupling between visual input, visual motor processing, and manual motor processing. The PEBL Cognitive Decathlon includes a Compensatory Tracking Task (similar to Makeig and Jolley's [1996] Task), shown in Panel F of Fig. 1. The task requires the participant to move a circular cursor onto a ring-like target. The cursor movement is influenced by the subject's mouse movements, but also by independent sinusoidal noise and repulsion from the target. For such a task, reasonable tests of artificial intelligence would evaluate first-order similarities in deviations from the target to those of humans.

**Aimed manual movement:** Fitts [1954] famously discovered a fundamental relationship between target distance and target size for rapid aimed movements, described by the relationship $t = a + b \times \log(d/s)$. Here, $d$ is the distance to the target, and $s$ is the size of the target, each in a common measurement unit. This relationship has several important properties. First, the critical baseline effects are that time is proportional to the logarithm of both the target distance and the target size. Yet robotic systems are often engineered to such a level of precision that they would likely produce linear relationships with distance, and possibly no effect of target size. The ability to produce these basic logarithmic relationships reveals an important property of biological motor control. In addition, the coefficient of the relationships between size and distance are typically identical. This remarkable property is still not completely understood or accounted for, but may provide insight into hypothesized mechanisms of aimed movement.

A version of the aimed movement test (Panel G of Fig. 1) is available within the PEBL Cognitive Decathlon. In this test, single movements from a base position to a known target are made via a computer mouse (although the software can be easily adapted to be used with touch-screen systems or alternate pointing devices). On different trials, targets of different sizes (10, 20, or 40 pixels) and different distances (between 50 and 700 pixels) are displayed, and the time needed to move the mouse between a starting position until it rests without moving on the target is recorded. Results typical to this test are shown in Fig. 4, which produced an inferred Fitts relationship of $t = 155 + 235 \times \log(d/s)$.

**Device learning and mimicry:** One primary avenue for learning tasks is through mimicry and observation. For the PEBL Cognitive Decathlon, a special device mimicry test (Panel H of Fig. 1) was implemented to serve as a platform for investigating mimicry and observational learning effects in motor control. A four degree-of-freedom virtual articulated device was created, which consisted of an arm rotating about the middle of space with a second arm attached to the first arm. The angles of rotation and lengths of each arm can be controlled via computer keyboard.

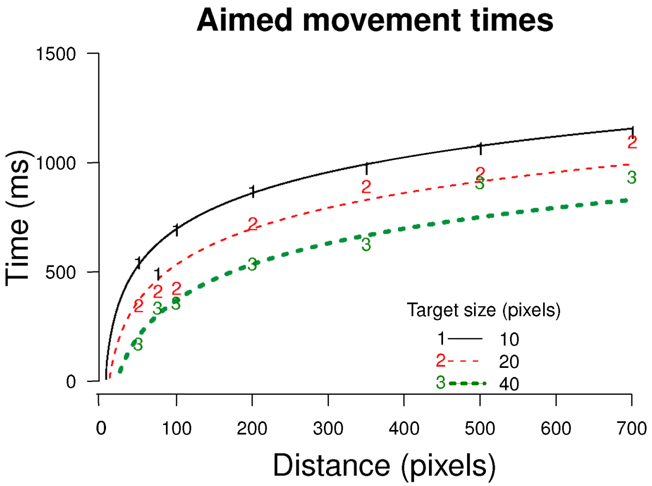


Fig. 4. Sample data from the Aimed Movement Task, for seven target distances and three target sizes. Mean response time is proportional to the logarithm of the ratio of distance and size.

The task involves learning how a novel motor action maps onto a physical effect in the environment, by reproducing traces created by a teacher. The software is able to record and replay sample paths created prior to an experiment. Sample paths can be shown as either complete, animated as it was drawn, or animated while showing the positions of the articulated arms. Basic effects expected to be displayed by humans includes initial key-mapping errors, improved ability to reproduce paths when the path animation and arm positions used to reproduce the path are shown.

### 3.4. *Task areas currently not implemented*

The originally-proposed BICA Cognitive Decathlon encompassed a broad range of behavioral tasks that could be carried out on both human performers and virtual or robotic agents. The tasks implemented here focus primarily on visual/spatial reasoning and somewhat on action control. A number of tasks were also proposed which focused on simple language interactions and their mappings between linguistic concepts and real-world concepts. These remain important skills, but do not appear in the version of the PEBL Cognitive Decathlon described here. Future versions of the PEBL Cognitive Decathlon may expand to cover those areas as well.

## 4. Discussion

This paper describes a set of implemented freely-available tasks that serve as a partial implementation of the Cognitive Decathlon. The basic goal of the Cognitive Decathlon is to provide a set of tasks that measure skills important for general artificial intelligence, but also delineate ways in which artificial behavior should be indistinguishable from human behavior if it is to be considered intelligent. Searching for Turing-indistinguishable behavior is important, because evaluations of artificial intelligence based purely on capability encourage focused but brittle systems. This is one of the challenges of developing biologically-inspired artificial intelligence: it can lag behind special-purpose artificial systems that are not restricted to be biologically-inspired. After all, generic AI can always incorporate biologically-inspired ideas when they make sense. Yet the BICA movement has grown based on the notion that these shortcuts taken by traditional AI, which perform narrow tasks well, have ultimately held the field back (see [Samsonovich and Mueller, 2008]).

It might be argued that it is a mistake to attempt to reproduce the foibles alongside the strengths of human intelligence. If the approach to doing so is just mimicry — developing an optimal system and introducing error processes to simulate human error — there is probably not much value. But human intelligence has evolved over time to be robust and flexible; and the flexibility of general-purpose intelligence introduces limitations for special-purpose problems. For example, the fact that humans do not have a "calibrated eyeball"[1] that can measure fine distinctions in size may

[1] The author thanks and acknowledges Mr. David McDowell for first bringing the conundrum of the calibrated eyeball to his attention.

seem like an evolutionary weakness that can be sidestepped by an artificial system. And, for example, today's robotic quality control systems use modern sensors to make very fine size discriminations in order to find faulty parts, in a way that an unaided human cannot. But our lack of a "calibrated eyeball" seems to be a direct consequence of living in a world in which our perspective is always changing. The lack of calibration turns out to be incredibly useful because we are able to perform shape-based matching to objects regardless of size or orientation. This skill, which is trivial for humans, can be remarkably difficult for some AI approaches, and AI systems that have taken this biologically-inspired skill seriously have exhibited impressive performance against complex classification sets (see [Latecki *et al.*, 2005]). This phenomena is likely to recur across a wide range of human skills: when consistent errors or imprecisions are found in human performance, it may go hand-in-hand with a flexibility and robustness, such that if our performance were better, it would also be more brittle.

The original BICA Cognitive Decathlon contained a plan for a number of ways in which artificial agent behavior could be compared to human behavior. Our intent was to evaluate the validity of agents based on detailed comparisons of performance with respect to robust trends, and in some ways drive the types of problems which we believed had the greatest chance of improving the state-of-the-art in general artificial intelligence. The (partial) implementation of the Decathlon described here is intended to further this goal, insofar as it enables human behavior in representative tasks to be examined and measured simply, especially by researchers who have strong backgrounds in artificial intelligence but less exposure to the many robust phenomena studied by cognitive psychology.

## References

Allender, L., Salvi, L. and Promisel, D. [1997] "Evaluation of human performance under diverse conditions via modeling technology," *Proceedings of Workshop on Emerging Technologies in Human Engineering, Testing and Evaluation, NATO Research Study Group 24*, Brussels, Belgium.

Attneave, F. and Arnoult, M. D. [1956] "The quantitative study of shape and pattern perception," *Psychological Bulletin* **53**, 452–471.

Bechara, A., Damasio, A. R., Damasio, H. and Anderson, S. W. [1994] "Insensitivity to future consequences following damage to human prefrontal cortex," *Cognition* **50**, 7–15.

Best, B. [2004] "Route planning and threat avoidance through cognitive robotics," *Proceedings of the 2004 Winter Simulation Conference.*

Busemeyer, J. R. and Johnson, J. G. [2004] "Computational models of decision making," *Blackwell Handbook of Judgment and Decision Making* (Blackwell Publishing Co., Oxford, UK), pp. 133–154.

Camara, W. J., Nathan, J. S. and Puente, A. E. [2000] "Psychological test usage: Implications in professional psychology," *Professional Psychological Research and Practice* **31**(2), 141–154.

Collin, C. A. and McMullen, P. A. [2002] "Using MATLAB to generate families of similar Attneave shapes," *Behavior Research Methods* **34**(1), 55–68.

Fitts, P. M. [1954] "The information capacity of the human motor system in controlling the amplitude of movement," *Journal of Experimental Psychology* **47**, 381–391.

Fleishman, E. A. [1975] “Toward a taxonomy of human performance,” *American Psychologist* **30**(12), 1127–1149.

Fleishman, E. A., Kinkade, R. G. and Chambers, A. N. [1968] “Development of a taxonomy of human performance: A review of the first year’s progress,” *DTIC Document AD0684583.*

Förster, B., Gebhardt, R., Lindlar, K., Siemann, M. and Delius, J. D. [1996] “Mental-rotation effect: A function of elementary stimulus discriminability?” *Perception* **25**, 1301–1316.

Goebel, R. P. [1990] “The mathematics of mental rotations,” *Journal of Mathematical Psychology* **34**(4), 435–444.

Harnad, S. [1989] “The Symbol Grounding Problem,” *Physica D* **42**, 335–346.

Harnad, S. [1990] “Other bodies, other minds: A machine incarnation of an old philosophical problem,” *Minds and Machines* **1**, 43–54.

Harnad, S. [2000] “Minds, machines and Turing,” *Journal of Logic, Language and Information* **9**(4), 425–445.

Harnad, S. [2004] “The annotation game: On Turing (1950) on computing, machinery, and intelligence,” in *The Turing Test Sourcebook: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, R. Epstein and G. Peters (eds.) (Kluwer).

Hunt, E. [1990] “A modern arsenal for mental assessment,” *Educational Psychologist* **25**(3), 223–241.

Itti, L. and Koch, C. [2001] “Computational modelling of visual attention,” *Nature Reviews Neuroscience* **2**(3), 194–203.

Larsen, A. [1985] “Pattern matching: Effects of size ratio, angular difference in orientation, and familiarity,” *Perception & Psychophysics* **38**(1), 63–68.

Latecki, L. J., Lakaemper, R. and Wolter, D. [2005] “Optimal partial shape similarity,” *Image and Vision Computing Journal (IVC)* **23**, 227–236.

Makeig, S. and Jolley, M. [1996] “COMPTRACK: A compensatory tracking task for monitoring alertness,” *Technical Document 96-3C Naval Health Research Center*, San Diego.

Mueller, S. T. [2009] “The Psychology Experiment Building Language (PEBL),” Version 0.10, `http://pebl.sf.net.`

Mueller, S. T., Anno, G., Fallon, C., Price, O. and McClellan, G. [2010] “Adapting the Task-Taxon-Task methodology to model the impacts of chemical protective gear,” in *Proceedings of the 19th Beh. Rep. in Modeling and Sim*, Charleston, SC.

Mueller, S. T. and Minnery, B. [2008] “Adapting the Turing test for embodied neurocognitive evaluation of biologically-inspired cognitive agents,” *AAAI Fall Symposium on Biologically Inspired Cognitive Architectures*, Arlington, VA.

Mueller, S. T., Zimmerman, L., Cheng, K., Crary, D., Simpkins, B., Gabbard, S. and McClellan, G. [2006] “Establishing a set of cognitive tests to predict the effect of novel protection suit design concepts,” in *Proceedings of Chem. Bio. Defense (CBD) Science and Technology Conference*, Dallas, TX.

Mueller, S. T., Zimmerman, L. Crandall, B., Cheng, K., Crary, D. and McClellan, G. [2009] “A qualitative analysis of MOPP gear and the warfighter,” in *Proceedings of the Chem. Bio. Defense (CBD) Science and Technology Conference*, Dallas, TX.

O’Donnell, R. D., Moise, S. and Schmidt, R. M. [2005] “Generating performance test batteries relevant to specific operational tasks,” *Aviation, Space, and Environmental Medicine* **76**, 24–30.

Parks, S. [2006] *Inside HELP, Administrative and Reference Manual* (VORT Corp., Palo Alto).

Pizlo, Z., Stefanov, E., Saalweachter, J., Li, Z., Haxhimusa, Y. and Kropatsch, W. [2006] “Traveling Salesman Problem: A foveating pyramid model,” *Journal of Problem Solving* **1**(1), 83–101.

Reitan, R. M. [1958] "Validity of the trail making test as an indicator of organic brain damage," *Perc. Mot. Skills* **8**, 271–276.

Rescorla, R. A. and Wagner, A. R. [1972] "A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement," *Classical Conditioning II*, (Appleton-Century-Crofts), pp. 64–99.

Samsonovich, A. and Mueller, S. T. [2008] "Toward a growing computational replica of the human mind," Preface to the *Papers from the AAAI Fall Symposium, Biologically Inspired Cognitive Architectures* (Menlo Park, AAAI Press).

Shepard, R. and Metzler, J. [1970] "Mental rotation of three dimensional objects," *Science* **171**, 701–703.

Simpson, R. L. and Twardy, C. R. [2008] "Refining the cognitive decathlon," *Proceedings of Perf. Metrics for Int. Systems (PERMIS)*, Gaithersburg, MD.

Sutton, R. S. and Barto, A. G. [1998] *Reinforcement Learning: An Introduction* (MIT Press).

Treisman, A. and Gelade, G. [1980] "A feature integration theory of attention," *Cog. Psych.* **12**, 97–136.

Turing, A. [1950] "Computing machinery and intelligence," *Mind* **LIX**, 433–460.

Wolfe, J. M., Cave, K. R. and Franzel, S. L. [1989] "Guided search: An alternative to the feature integration model for visual search," *Journal of Experimental Psychology: Human Perception and Performance* **15**(3), 419–433.