

Beyond the Turing Test: Performance Metrics for Evaluating a Computer Simulation of the Human Mind

Nancy Alvarado, Sam S. Adams, Steve Burbeck, Craig Latta
IBM, Thomas J. Watson Research Center
{nanalv, ssadams, sburbeck, crl}@us.ibm.com

Abstract

Performance metrics for machine intelligence (e.g., the Turing test) have traditionally consisted of pass/fail tests. Because the tests devised by psychologists have been aimed at revealing unobservable processes of human cognition, they are similarly capable of revealing how a computer accomplishes a task, not simply its success or failure. Here we propose the adaptation of a set of tests of abilities previously measured in humans to be used as a benchmark for simulation of human cognition. Our premise is that if a machine cannot pass these tests, it is unlikely to be able to engage in the more complex cognition routinely exhibited by animals and humans. If it cannot pass these sorts of tests, it will lack fundamental capabilities underlying such performance.

1. Introduction

What constitutes success for designers using computational autonomous development to create intelligent systems? The observation of change over time with experience in an environment is neither learning nor development unless that change results in closer approximation to some performance goal. Further, neither development nor learning necessarily indicates that a machine has behaved intelligently. While it may be interesting to observe how systems respond to different environmental challenges and how parameters affect a system's behavior, eventually some attempt to guide learning must be made if these systems are to shed much light on human cognitive processes or prove useful in applications.

This was the challenge we faced in developing a set of performance goals for the Joshua Blue Project. Joshua Blue applies ideas from complexity theory and evolutionary computational design to the simulation of a human mind on a computer. The goal is to enhance artificial intelligence by enabling the emergence of such capacities as common sense reasoning, natural language understanding, and emotional intelligence, acquired in the same manner as humans acquire them, through situated learning in a rich environment. Because our goal is to simulate human cognition, our performance goals arise from observing how humans behave in similar circumstances. Thus our project has been guided by findings in developmental and cognitive psychology.

If a system is conceived as a "black box," engineers work from inside the box to implement the capabilities that will achieve a desired outcome within a given set of constraints. In contrast, psychologists work from outside the box, reverse-engineering what already exists in the human mind to develop plausible explanations for observed behavior. These different interests converge in the need to specify performance goals for autonomous learning systems because, as in psychology, such testing must be performed from outside the box. Thus, the paradigms used by psychologists to test hypotheses about how the human mind functions provide a ready metric for assessing how closely computer behavior approximates human performance in well-defined contexts. In these tests, the observed behavior of an autonomous system becomes proof of what must be occurring inside a computer black box designed by an engineer. This is important for those systems that cannot be inspected directly – where the changes taking place are hidden from the system's designers or difficult to interpret, as is typically the case with architectures based on artificial neural networks.

For many purposes, it may not matter that a computer accomplishes a task in a manner similar to humans. In our case, because the goal of the Joshua Blue Project was to simulate human understanding, thinking in the same manner as humans has been an important design goal. Our reasoning has been that every human capacity exists within the interdependent context of the human mind for a specific reason. Thus, approximating human functioning as closely as possible in all respects seems necessary to success at modeling a mind with rich, robust human-like cognitive abilities. This goal is not as restrictive as it seems, given the variation that exists among humans. Experience may teach us which abilities are essential, which are not, and how much deviation is possible without sacrificing the cognitive qualities of the resulting system.

When a comparison between machine and human functioning is proposed, it seems natural to ask "which human shall we compare our machine with?" This question arises because humans do not all perform alike. Psychologists have addressed this by describing behavior in terms of distributions. The "normal" or bell curve for intelligence is one such description of the range of human performance on a specific measurement instrument (such as an IQ test). Such curves permit one to place a particular score or measured

behavior within the context of behavior for a large number of others. Thus it provides not simply pass/fail information but a basis for comparison with the range of human variation. Norms exist for most published tests. When machines become more autonomous in their learning and have greater flexibility in what they can do, when their learning experiences are richer and more varied, governed by the machine's own choices, the behavior of intelligent systems will also be less predictable and more variable. Statistical techniques for describing variance and comparing behavior, like those used by psychologists, will become more important in assessing machine performance. Such techniques will be needed to determine whether different observed behavior reflects learning or simply change.

One longstanding test of human-like computer intelligence has been the Turing Test [1], now conducted each year as the Loebner Competition [2]. A drawback of that test is that human functioning can be mimicked by systems that have little if any human-like cognitive capabilities. The performance goal can be met by simulating human behavior convincingly (to the satisfaction of human judges) for a relatively brief period of time in a restricted domain of discourse. Surface behavior is partially simulated but the underlying cognition producing that behavior is not. We believe that meaningful performance tests must incorporate the idea that it matters how behavior is accomplished, not simply that it occurs [3].

When a developmental process is considered important to acquiring human-like understanding, as we believe it to be, then a dimension of change in stages with milestones approximating human development is introduced. The time frame for a computer need not be the same as for a human child because its training experiences may differ, but the order of acquisition of abilities, the interrelationship of change in one domain with change in another, and so on, should be considered. Taking a modular approach, as Fodor suggests, temporarily simplifies evaluation of developmental sequence and inter-relationships among processes, but at some point separate processes become unified, domain-specific knowledge converges, sensory data is integrated, and more general or encompassing cognition emerges.

Ultimately, as a measure of success, we envision Joshua Blue acquiring the capacity to pass not an adult Turing Test, but a "Toddler Turing Test." Success would mean that our system would answer unrestricted questions posed by adults in the same manner as a three year old toddler would, with all of the common sense, naive physics, burgeoning language use, and social understanding of a preschool child. The true test would be whether Joshua Blue is distinguishable from human three year olds, not whether Joshua Blue fools an adult into believing it is not a computer, through faked typing errors, colloquialisms, illogic or scripted emotional responses.

2. Overview of the Test Suite

The proposed test suite is intended to assess whether certain essential, prerequisite cognitive abilities exist in a system. This required identification of foundational abilities without which a system cannot be considered a simulation of human cognition. The approach to testing for these abilities is twofold. First, changes in internal system values and other parameters indicating changes in mental state can be monitored and measured, and related to observed behavior. Second, well-understood paradigms from psychology can be adapted for use as performance benchmarks. In many cases, such paradigms have demonstrated existence of unobservable cognitive processes through observable and quantifiable behavior. By comparing system results to results obtained with animals and humans, designers can make stronger claims about the capabilities of that system.

The test suite is divided into three parts, as shown in Table 1. The goal for systems simulating the general properties of human cognition must be to pass all of the tests using a single, generalized system. Ultimately, passing an unrestricted toddler Turing test would require success in all three parts of this initial test suite, plus a great deal more.

3. Tests of Associative Learning

We believe that a system capable of learning from its environment must be endowed with motivation that gives it the impetus to explore its world, recognize reward and punishment, form goals and seek to satisfy them. While certain drive states may be innate to the system, it should also form acquired goals and use past experience to determine a course of action. This test suite is designed to assess a system's ability to direct its own behavior by recognizing environmental cues that signal reward or punishment. The tests are based upon animal testing in the field of behavior modification and learning. An excellent overview of this work is presented by Klein [4].

3.1 Tests Showing Formation of Associations

In classical conditioning, associations are formed between environmental cues (called conditioned stimuli) and elicitors of affective or motivational states (called unconditioned stimuli). Training occurs by presenting these two kinds of stimuli together until one becomes a predictor of the other and is responded to as if it were the other stimulus. Passing this test requires the ability to recognize salient features of the environment and associate them with the internal responses evoked by a co-occurring unconditioned stimulus.

Table 1. Overview of Performance Metrics for Evaluating a Simulation of Mind

Associative Learning	Classical Conditioning	Forms associations between a predictive environmental cue and an accompanying stimulus
	Presence of a Mind	Forms mental representation of two cues as a single predictor, demonstrates memory for previously encountered predictive cues
	Instrumental Conditioning	Demonstrates cause-effect learning and responds to changes in amount or rate of reward or punishment
	Purposeful Behavior	Acquires goals and expectations, shows “learned helplessness” with a lack of ability to achieve goals, shows escape/avoidance behaviors
Social Cognition	Social Encoding	Categorizes self and others by salient features and behavior, forms stereotypes, forms a self schema
	Social Inference	Forms expectations (heuristics) based on observed correlations and covariation among properties and behaviors of others, exhibits biases resulting from use of such heuristics
	Causal Attribution	Attributes motives to others, exhibits biases resulting from such attributions, generalizes its self schema to others
	Representation of Self	Forms a representation of self observable in self-preserving or enhancing behavioral choices in contexts that threaten self-image
	Empathy and Attachment	Generalizes the sense of self to others, demonstrates changes in affect depending upon affiliation with others, shows affect-guided helping behavior
Language Acquisition	Prelinguistic Structural Competences	Performs combinative operations on sets, acquires phoneme combinations appropriate to a specific language, progresses through cooing to babbling and prosody mimicking speech, acquires pragmatics of communication, understands communicative intentions of others
	Intentional Communication	Acquires and uses words to accomplish goals both through associative learning and imitation of observed others
	Word Acquisition	Sorts words by grammatical types (e.g., noun, verb), learns grammatical rules in same sequence as children, overgeneralizes grammatical rules
	Cross-Language Comparisons	Demonstrates language-specific competence in a language other than the first language taught, repeating the three tests above.

Given the capacity to form associations through experience in an environment, classical conditioning can be tested in a virtual environment using an embodied agent endowed with minimal sensors and effectors, motives, affect, and proprioception of both actions and affective states. In order to pass the classical conditioning tests, the system must form a classically conditioned association between an unconditioned stimulus (e.g., satisfaction of “hunger”) and a conditioned stimulus (cues signaling availability of “food”). This association should result in observed changes in behavior. If a pleasant stimulus is replaced by an aversive (unpleasant) stimulus, then fear-based conditioning should result in avoidance and escape behaviors.

A rich literature described in most learning theory textbooks [4, 5] specifies quantitative relationships between factors affecting learning and observed behavior. These experiments can be used as a highly specific model against which the behavior of a system can be compared.

3.2 Tests Demonstrating Mental Representation

A series of studies in the learning literature were used to demonstrate that certain classical conditioning effects are the result of expectations, which cannot exist unless experience is being represented mentally. Several phenomena can be used to show the existence of mental representations in a system. This use of behavioral paradigms to demonstrate existence of mental representations illustrates that the goal of a study that measures behavior can be far more than simply to demonstrate competence in performing that behavior. It can also demonstrate indirectly what must exist in the less observable parts of the system. The reasoning is that the observed behavior, although often uninteresting in its own right, could not be performed if certain important processing were not occurring. Two examples are presented here, but there are many more in the learning literature.

Normally, when two cues to an event are present, the more salient cue (the better predictor) will block conditioning of the lesser cue. However, if the two cues are perceived as part of a unified percept, instead of as two separate events, then the opposite effect occurs and presence of the highly salient cue potentiates the response to the less noticeable cue. This potentiation cannot occur without formation of a mentally unified percept incorporating both cues. Similarly, in backward blocking, a new aversive event becomes associated with a previous environmental cue not actually present during the unpleasant experience. If no memory of the cue were formed, it could not become associated with the aversive stimulus that came later. This provides an indirect way of confirming that a system is forming mental representations of its world and associating percepts with each other

3.3 Cause-Effect Learning Tests

Operant and instrumental conditioning (cause-effect learning) paradigms test whether associations are formed between environmental stimuli and voluntary behavior, governed by reward or punishment. Virtual or physical environments similar to the mazes and runways presented in

classic psychology experiments can be created to test such learning. These include runways along which an agent or robot can move, T-mazes and radial-arm mazes permitting choice of paths, two-compartment shuttle boxes to test escape and avoidance learning, more complex mazes like those used by Tolman [6] to demonstrate existence of cognitive maps, virtual or physical “Skinner boxes” (single and two-choice lever-pressing boxes). Equivalent behaviors can be devised, such as moving to a specific region of the environment to obtain reward.

To pass these tests, a system must show a reliable relationship between reward, punishment, and subsequent behavior. The system should increase its behavior in response to a reinforcer, decrease it in response to a punishment or in the absence of a reinforcer. It should also conform to effects of changes in size, rates and timing of reinforcement, as described in the literature on schedules of reinforcement [4, 5]. Further tests of complex contingencies can show that the system has the capacity to monitor longer chains of behavior and form more complex goals.

Learning theory includes descriptions of several more sophisticated learning phenomena involving expectations, multiple reinforcers and behavioral allocation (dividing attention between two sources of reward). We expect that a system should show similar behaviors, including Crespi’s depression effect, punishment effects, a partial reinforcement effect, adherence to Premack’s principle and Herrnstein’s matching law, all classic learning phenomena described in textbooks [4, 5]. Because stimulus control phenomena are closely related to ability to discriminate, identify relevant features, and form concepts and categories, they are not discussed here, but are described in Section 5.

3.4 Tests of Purposeful Behavior

Simple cause-effect learning does not necessarily imply that an organism has formed a mental representation of its goal states. Tolman’s [6] paradigms can be adapted to demonstrate existence of purpose, expectations, and mental maps in a system. A system must be capable of representing aspects of its environment for future use, even when those representations are not directly related to a current operation or goal state. Latent learning studies can be used to show that mental representations are being formed without reward.

Learned helplessness and phobias are both examples of phenomena demonstrating the existence of expectations. Learned helplessness occurs when an organism is subjected to punishment but can do nothing to escape it, and it results in a change in expectations about effectiveness of future activity. Phobias represent acquisition of a classically conditioned association between some environmental cue and an aversive event. Observation of phobias indicates formation of expectations about negative events, plus beliefs about what behaviors will prevent those events. Suitable paradigms for testing each of these can be adapted.

4. Tests of Social Cognition

Tests of two or more agents interacting in a shared environment permit application of a system's basic cognitive capacities to social interactions. In addition to the capacities needed to function alone in such an environment, the system must communicate, compete and form alliances with other agents, understand the behavior and intentions of others, and find ways to coexist with them while also attaining important goals. Fiske and Taylor [7] provide a thorough overview of paradigms for studying human social cognition.

4.1 Tests of Social Encoding Processes

Social encoding can be tested in a simplified, virtual multiple-agent environment by creating two types of embodied agents: (1) agents with similar bodies and behavior; (2) agents with dissimilar bodies and behavior. Scenarios can be manipulated so that these agents compete with, punish (show aggression), or cooperate with and help the system being tested. From this experience, the system should form an appropriate mental representation of both like and unlike agents and behave according to its previous experience with those agents. Stereotypes should be formed as mental representations of the characteristics of these groups of agents. Stereotypes can be demonstrated by observing an experienced system's behavior toward an ambiguous new agent that includes some but not all of the salient characteristics of previously learned social categories. Once social categories have been learned, increased ambiguity in classification of new agents should result in interruptions of behavior and affective distress in the system.

4.2 Tests of Social Inference

Social inferences are based on observed correlations and covariation among properties and behaviors of other agents and environmental events and features. Once learned, these become heuristics used in human reasoning. Existence of such heuristics is observed in specially devised contexts where they produce inference "errors." If a system has acquired heuristics, then it should demonstrate the same errors as humans, instead of strictly mathematical/statistical decision processes. These should result in the biases enumerated in most textbooks [7], including: (1) framing effects, use of extreme cases, over-reliance on small samples or biased samples; (2) salience of immediate experience; (3) inability to combine joint probabilities; (4) inability to identify diagnostic information; (5) inability to correct for regression artifact; (6) irregular or improper weighting of cues.

4.3 Tests of Causal Attribution

Humans recognize that living beings have agency – the ability to control their own actions in goal-directed ways. Thus they attribute motives to others and understand the actions of others in terms of intentions. While people understand that circumstances can dictate behavior, they prefer to attribute the actions of others to inherent motives, preserving a belief in free will. This is reflected in certain

attribution biases [7], including the fundamental attribution error, the actor-observer effect, and self-serving biases. Environments can be created for testing fundamental attribution error, the false consensus effect (where we consider our behavior to be more representative of the behavior of others than it really is), and self-serving bias.

4.4 Tests of a Representation of the Self

People tend to form representations of the self (self schemas) along dimensions that are important to them in their lives. Complexity of the self-schema is also determined by life experiences. Thus a system's self-schema should depend upon its experiences. In humans, tasks involving perception, memory and inference are used to demonstrate the characteristics of the self schema. These can be adapted to show that a system has formed a schema whose characteristics vary with experience [7]. For example, self-schemas change to more closely resemble those of others surrounding a person. Thus the existence of a self-schema can be demonstrated through changes in behavior that occur when the system is placed in a new environment, among different social agents than were present when the self-schema was formed. Maintenance or preservation of a self-schema motivates much human behavior. A cybernetic theory of self-attention and self-regulation has been proposed by Carver & Scheier [8]. This theory of regulation, with predicted consequences for failure, provides a detailed model for testing a system's self-schema in different contexts.

4.5 Tests of Empathy and Attachment

In humans, empathy and attachment are the two main mechanisms for regulating social interactions. Both depend upon the cognitive evaluations and interpretations made of social situations. When a system is endowed with affective responses, these should mediate social interactions in ways consistent with what is observed in humans. Tests of emotion-motivated helping behavior can be adapted to assess empathy [7]. Berscheid [9] proposes a theory that predicts how emotion changes in long-term relationships, readily testable in the simple social contexts described earlier.

5. Tests of Language Acquisition

The criteria for assessing machine language acquisition are that a computer's language must be used with the flexibility of a real language, but must not be so constrained that one language can be learned but not another (e.g., English but not Chinese). Two aspects of language acquisition are intertwined: (1) understanding of the structural properties of the representations comprising language; (2) understanding of the content of concepts represented by language. Understanding of concepts appears to precede language acquisition, although a child's preexisting concepts are in turn modified by later language learning. The complexities of the structure of language are so great that many theorists hypothesize innate mechanisms for acquiring structural

understanding, and even for acquiring conceptual understanding.

An important prerequisite to language learning is the ability to perform combinative operations on sets. Langer [10] provides convincing evidence that this ability is needed in order to form the class inclusion hierarchies essential to categorization. He also argues that these operations (recursive mapping of cognitions, correspondence mapping, substitutions and equivalence relations) are the foundation of abstraction and form the basis for linguistic rewrite rules.

Some theorists, such as Pinker, hypothesize a straightforward mapping of words to “mentalese” nonverbal representations, guided by innate mechanisms. Other theorists assert that general associationist learning can account for language acquisition, especially if social imitative learning directs a child’s attention to what is relevant during learning [11, 12]. Thus, a system must acquire language socially and must have the mechanisms for imitative social learning in place before acquiring language.

5.1 Tests of Prelinguistic Structural Competences

The first year of life is spent acquiring prelinguistic competences supportive of ultimate language development. These include a progression from cooing and laughing to vocal play, babbling, and utterances that mimic the prosody of speech, acquisition of the nonverbal pragmatics of communicative interactions, such as smiling, gaze orienting, turn-taking, pointing and showing, and mimicking behaviors. The child also acquires an understanding of the communicative intentions of adults [12]. A first task is to learn to segment a flow of speech sounds into meaningful combinations of phonemes (morphemes). The system’s developing language should show increasing inclusion of language-specific permissible combinations of phonemes. A set of tests requiring the system to classify items into sets and to perform the combinative operations on those sets described by Langer [10] can assess whether the conceptual cognitive abilities prerequisite to language learning exist.

5.2 Tests of Intentional Communication

Intentional communication requires use of language instrumentally to accomplish goals and an understanding of the ways in which others do so. The social paradigms described earlier can be adapted to create situations in which a system must learn to use words to accomplish specific goals. Placed into a new environment, the system must be able to learn to use words through observation and imitation of the behavior of other social agents inhabiting the same environment.

5.3 Tests of Word Acquisition

An understanding of grammar is necessary for word acquisition because grammar constrains the possible meanings of new words. While this grammatical knowledge may be innate, some theorists argue that the statistical frequencies in language call attention to the structural

properties of language sufficiently to permit grammatical knowledge to emerge [11]. Either way, tests of word acquisition begin with tests of grammatical knowledge. A system must be able to tell the difference between nouns, verbs, and adjectives, in a simple sorting task. Further, the system must use that knowledge to classify objects. This can be tested by adapting paradigms used with young children, such as those demonstrating shape bias with nouns [11]. Tests of more sophisticated grammatical knowledge demonstrate rule acquisition by observing occurrence of errors resulting from the overgeneralization of rules. These errors should gradually disappear with greater experience.

5.4 Cross-Language Comparisons

Once a system has demonstrated competence in one language, it must pass the same set of tests to demonstrate competence in a different language.

6. Conclusion

There are many other aspects of cognitive functioning that might be assessed. We propose this battery as a place to start. If a system cannot pass these tests, it is unlikely to be able to engage in the more complex cognition routinely exhibited by animals and humans. Because all of these tests focus upon abilities quantitatively measured in humans, they provide a ready-made benchmark for simulation of human cognition. The greater challenge is to create learning environments sufficient to enable machines to acquire these abilities if they are truly capable of doing so.

Notes

This paper was originally presented at the Second International Conference on Development and Learning, ICDL 2002, MIT, Cambridge, MA. Portions have been revised in response to comments by conference attendees. Nancy Alvarado may be reached at the University of California, San Diego, Center for Brain and Cognition (alvarado@psy.ucsd.edu).

References

- [1] A. M. Turing, “Computing Machinery and Intelligence,” In *Readings in Cognitive Science: A Perspective from Psychology and Artificial Intelligence*, A. Collins & E. Smith (Eds.), pp. 6-19, Morgan Kaufmann, San Francisco, CA, 1950/1991.
- [2] H. Loebner, “Loebner Prize,” described at: <http://www.loebner.net/Prizef/loebner-prize.html>. Breazeal, C. 2001. *Designing Sociable Machines*. The MIT Press. Forthcoming.
- [3] J. R. Searle, 1980. Minds, brains, and programs, *Behavioral and Brain Sciences*, 3, 417-424.
- [4] S. B. Klein, 1996. *Learning: Principles and Applications*. McGraw-Hill Inc., New York, NY.
- [5] G. Bower & E. Hilgard, 1981. *Theories of Learning*. Prentice-Hall, Inc., Englewood Cliffs, NJ.

- [6] E. C. Tolman and C. H. Honzik, 1930. "Insight" in Rats. University of California Publications in Psychology, 4, 215-232.
- [7] S. T. Fiske and S. E. Taylor, 1991. *Social Cognition*. McGraw-Hill Inc., New York, NY.
- [8] C. S. Carver & M. F. Scheier, 1981. *Attention and Self-regulation: A Control Theory Approach to Human Behavior*, Springer-Verlag, NY.
- [9] E. Berscheid, M. Snyder, & A. Omoto, 1989. Issues in studying close relationships: Conceptualizing and measuring closeness. In C. Hendrick (Ed.), *Review of Personality and Social Psychology: Vol 10. Close Relationships* (pp. 63-91). Sage, Newbury Park, CA.
- [10] J. Langer, 2001. The mosaic evolution of cognitive and linguistic ontogeny. In M. Bowerman and S. Levinson (Eds.), *Language Acquisition and Conceptual Development*. Cambridge University Press, New York, NY.
- [11] L. Smith, 2001. How domain-general processes may create domain-specific biases. In M. Bowerman and S. Levinson (Eds.), *Language Acquisition and Conceptual Development*. Cambridge University Press, New York, NY.
- [12] M. Tomasello, 2001. Perceiving intentions and learning words in the second year of life. In M. Bowerman and S. Levinson (Eds.), *Language Acquisition and Conceptual Development*. Cambridge University Press, New York, NY.