I-athlon: Toward a Multidimensional Turing Test

Sam S. Adams, Guruduth Banavar, Murray Campbell

■ *While the Turing test is a well-known method for evaluating machine intelligence, it has a number of drawbacks that make it problematic as a rigorous and practical test for assessing progress in general-purpose AI. For example, the Turing test is deception based, subjectively evaluated, and narrowly focused on language use. We suggest that a test would benefit from including the following requirements: focus on rational behavior, test several dimensions of intelligence, automate as much as possible, score as objectively as possible, and allow incremental progress to be measured. In this article we propose a methodology for designing a test that consists of a series of events, analogous to the Olympic Decathlon, which complies with these requirements. The approach, which we call the I-athlon, is intended ultimately to enable the community to evaluate progress toward machine intelligence in a practical and repeatable way.*

The Turing test, as originally described (Turing 1950), has a number of drawbacks as a rigorous and practical means of assessing progress toward human-level intelligence. One major issue with the Turing test is the requirement for deception. The need to fool a human judge into believing that a computer is human seems to be peripheral, and even distracting, to the goal of creating human-level intelligence. While this issue can be sidestepped by modifying the test to reward rational intelligent behavior (rational Turing test) rather than humanlike intelligent behavior, there are additional drawbacks to the original Turing test, including its language focus, complex evaluation, subjective evaluation, and the difficulty in measuring incremental progress.

  ISSN 0738-4602

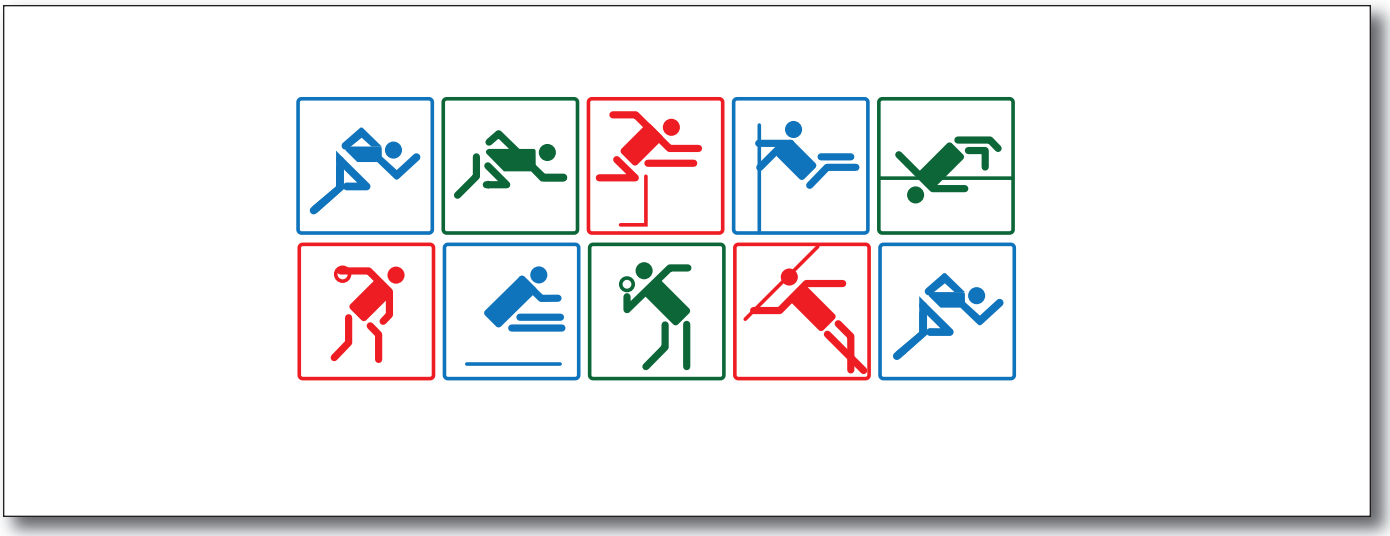

Figure 1. The Olympic Decathalon.

Language focused: While language use is perhaps the most important dimension of intelligence, there are many other dimensions that are relevant to intelligence, for example, visual understanding, creativity, reasoning, planning, and others.

Complex evaluation: The Turing test, if judged rigorously, is expected to require extensive human input to prepare, conduct, and evaluate.[1]

Subjective evaluation: Tests that can be objectively evaluated are more useful in a practical sense, requiring less testing to achieve a reliable result.

Difficult to measure incremental progress: In an unrestricted conversation, it is difficult to know the relative importance of various kinds of successes and failures. This adds an additional layer of subjectivity in trying to judge the degree of intelligence.

In this article we propose an approach to measuring progress toward intelligent systems through a set of tests chosen to avoid some of the drawbacks of the Turing test. In particular, the tests (1) reward rational behavior (as opposed to humanlike behavior); (2) exercise several dimensions of intelligence in various combinations; (3) limit the requirement for human input in test creation and scoring; (4) use objective scoring to the extent possible; (5) permit measuring of incremental progress; (6) make it difficult to engineer a narrow task-specific system; and (7) eliminate, as much as possible, the possibility of gaming the system, as in the deception scenarios for the classic Turing test.

The proposed approach, called here the I-athlon, by analogy with the Olympic Decathlon[2] (figure 1), is intended to provide a framework for constructing a set of tests that require a system to demonstrate a wide variety of intelligent behaviors. In the Olympics, 10 events test athletes across a wide variety of athletic abilities as well as learned skills. In addition, the Decathlon tests their stamina and focus as they move among the 10 events over the two days of the competition. In all events, decathletes compete against specialist athletes, so it is not uncommon for them to fail to win any particular event. It is their aggregate score that declares them the World's Greatest Athlete. One of the values of this approach for the field of artificial intelligence is that it would be inclusive of specialist systems that might achieve high levels of proficiency, and be justly recognized for the achievement, while still encouraging generalist systems to compete on the same level playing field.

Principles for Constructing a Set of Tests

Given our desire for broad-based, automated, objectively scored tests that can measure incremental progress and compare disparate systems on a common ground, we propose several principles for the construction of I-athlon events:

Events Should Focus on Testing Proficiency in a Small Number of Dimensions.
Testing a single dimension at a time could fall prey to a switch system, where a number of narrow systems are loosely coupled through a switch that selects the appropriate system for the current event. While events should be mostly self-contained, it may make sense to use the results of one event as the input for another.

Events Should All Be Measured Against a Common, Simple Model of Proficiency.
A common scoring model supports more direct comparisons and categorizations of systems. We propose a simple five-level rating system for use across all events. Levels one through four will represent levels of human proficiency based on baseline data gath-

ered from crowdsourced human competitions. Level five will represent superhuman proficiency, an X-factor over human level four, so there is a clear, unambiguous measure of achievement above human level. Levels one through four could be mapped to human age ranges or levels of proficiency, though some tests will not map to human development and proficiency but to domain expertise. It will be the responsibility of the developers of each event to map their scoring algorithms to these levels, and the overall I-athlon score for any competing system will be a standard formula applied to attainment of these levels.

Event Tests Should Be Automatically Generated Without Significant Human Intervention.

One of the major drawbacks to the current Turing test is its requirement for extensive human involvement in performing and evaluating the test. This requirement for direct human involvement effectively rules out highly desirable approaches to developing solutions that operate much faster than humans can interact with effectively. Another challenge in designing a good replacement for the Turing test is eliminating, as much as possible, the potential for someone to game the system. At the very least this means that specific test instances must not be reused except for repeatability and validation. Automatic generation of repeatable high-quality tests is a significant research area on its own, and this approach allows for more efficient division of labor across the AI research community. Some researchers may focus on defining or improving events, possibly in collaboration with other disciplines like psychology or philosophy. Some may focus on developing test generators and scoring systems. Others may develop systems to compete in existing I-athlon events themselves. Generators should be able to reproduce specific tests using the same pseudorandom seed value so tests can be replayed for head-to-head competition and to allow massively parallel search and simulation of the solution space.

Event Tests Should Be Automatically Scored Without Significant Human Intervention.

Deception of human judges became the primary strategy for the classic Turing test instead of honest attempts at demonstrating true artificial intelligence. Human bias on the part of the panel of judges also made the results of each run of the Turing test highly unpredictable and even suspect. To the degree possible, scoring should be consistent and unambiguous, with clearly defined performance criteria aligning with standard proficiency level scoring. These scoring constraints should also significantly influence test design and generation itself. To prevent tampering and other fakery, all test generators and scoring systems should run in a common secure cloud, and all tests and results should be immutably archived there for future validation.

The Scoring System Should Reward Proficiency over Multiple Events.

The overall goal of this effort is to create broadly intelligent systems rather than narrow savants. As in the Olympic Decathlon, the total score across events should be more important than the score in any one event. Relative value of proficiency level achievement needs to recognize that all events are not equal in intelligence value. This might be difficult to agree on, and even the Olympic Decathlon scoring system has evolved over time to reflect advances in performance.[3]

Dimensions of Intelligence

Human intelligence has many facets and comes in many varieties and combinations. Philosophers, psychologists, cognitive and computer scientists have debated the definition of intelligence for centuries, and there are many different factorings of what we here call the "dimensions of intelligence." Our goal in this article is not to declare a definitive set of dimensions or even claim complete coverage of the various aspects of human intelligence. We take up this terminology to enable us to identify aspects of intelligence that might be tested separately and in combinations for the purpose of evaluating the capabilities of AI systems compared to humans. The dimensions listed below are not all at the same level of abstraction; indeed, proficiency at some dimensions will require proficiency at several others. We fully expect there to be debate over which aspects of intelligence should be tested for separately or in concert with others. Our goal here is to define an approach that moves the AI research community in the positive direction of coordinated effort toward achieving human-level AI in computer systems. As stated earlier, we believe reaching this goal will require such a coordinated effort, and a key aspect of coordination is the ability to assess incremental progress toward the goal in a commonly accepted manner. What follows is a brief description of what we consider good candidates for I-athlon events (figure 2).

Image Understanding — Identify both the content and context of a given image, the objects, their attributes and relationships to each other in the image, implications of scene background and object arrangement.

Diagram Understanding — Given a diagram, describe each of the elements and their relationships, identify the intended purpose/message of the diagram (infographic, instructional, directional, design, and others).

Speech Generation — Given a graph of concepts describing a situation, deliver an appropriate verbal/auditory presentation of the situation.

Natural Language Generation — Given nonverbal information, provide a natural language description sufficient to identify the source information among alternatives.

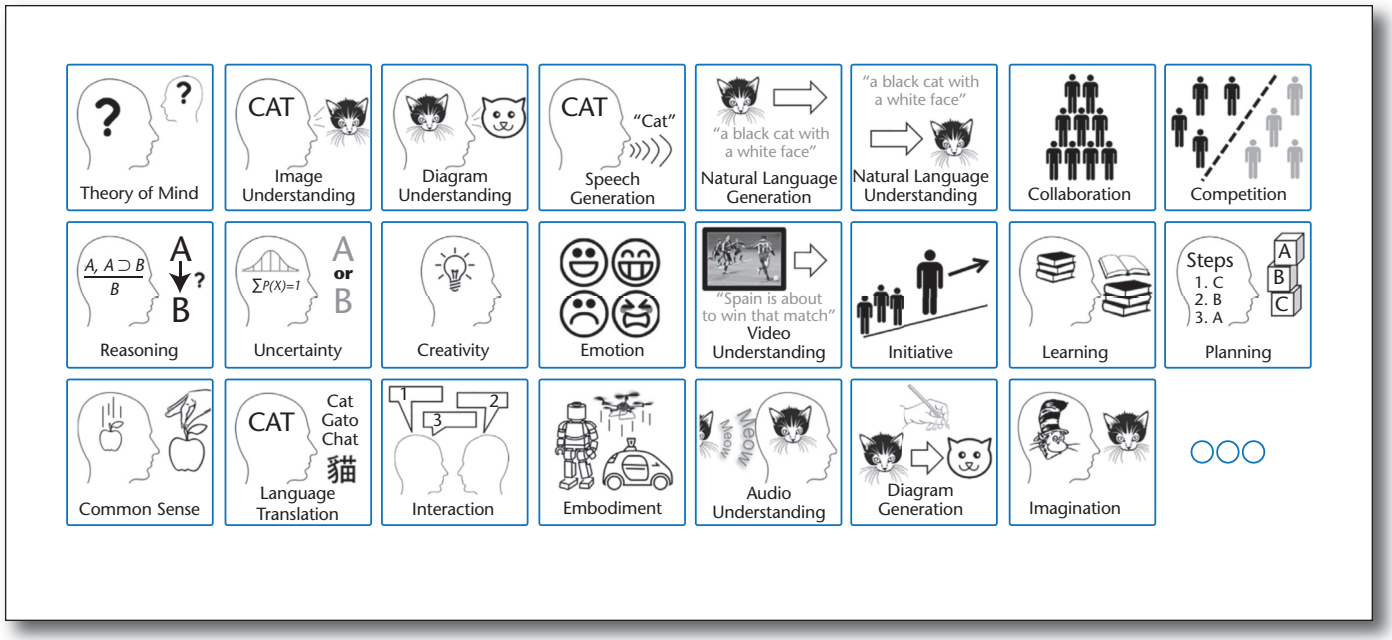

Figure 2. Good Candidates for the I-athlon Events.

Natural Language Understanding — Given a verbal description of a situation, select the image that best describes the situation. Vary the amount of visual distraction.

Collaboration — Given descriptions of a collection of agents with differing capabilities, describe how to achieve one of more goals within varying constraints such as time, energy consumption, and cost.

Competition — Given two teams of agents, their capabilities and a zero-sum goal, describe both offensive and defensive strategies for each team for winning, initially based on historical performance but eventually in near real time.

Reasoning — Given a set of states, constraints, and rules, answer questions about inferred states and relationships. Explain the answers. Variations require use of different logics and combinations of them.

Reasoning Under Uncertainty — Given a set of probable states, constraints, and rules, answer questions about inferred states and relationships. Explain the answers.

Creativity — Given a goal and a set of assets, construct a solution. Vary by number and variety of assets, complexity of goals, environmental constraints. Alternatively, provide a working solution and attempt to improve it. Explain your solution.

Video Understanding — Given a video sequence, describe its contents, context, and flow of activity. Identify objects and characters, their degree of agency and theory of mind. Predict next activity for characters. Identify purpose of video (educational, how-to, sporting event, storytelling, news, and others). Answer questions about the video and explain answers.

Initiative — Given a set of agents with different capabilities, goals, and attitudes, organize and direct a collaborative effort to achieve a goal. Key here is utilizing theory of mind to build and maintain the team throughout the activity.

Learning — Given a collection of natural language documents, successfully answer a series of questions about the information expressed in the documents. Vary question complexity and corpora size for different levels. Similar tests for nonverbal or mixed media sources.

Planning — Given a situation in an initial state, describe a plan to achieve a desired end state. Vary the number and variety of elements, and the complexity of initial and end states, as well as the constraints to be obeyed in the solution (for example, time limit).

Common Sense Physics — Given a situation and a proposed change to the situation, describe the reactions to the change and the final state. Vary the complexity of the situation and the number of changes and their order.

Language Translation — Given text/speech in one language, translate it to another language. Vary by simplicity of text, number of idioms used, slang, and dialect.

Interaction — Given a partial dialogue transcript between two or more agents, predict what will be the next interactions in the exchange. Alternatively, given an anonymous collection of statements and a description of multiple agents, assign the statements to each agent and order the dialogue in time.

Embodiment — Given an embodiment with a collection of sensors and effectors, and an environment

surrounding that body, perform a given task in the environment. Vary the number and sophistication of sensors and effectors and tasks, the complexity of the environment, the time allowed. Added bonus for adapting to sensors/effectors added or disabled during the test.

Audio Understanding — Given an audio sequence, describe the scene with any objects, actions, and implications. Vary length and clarity, along with complexity of audio sources in the scene.

Diagram Generation — Given a verbal description of a process, generate a series of diagrams describing the process. Alternatively use video input.

Imagination — Given a set of objects and agents from a common domain along with their attributes and capabilities, construct and describe a plausible scenario. Score higher for richer, more complex interactions involving more agents and objects. Alternatively, provide two or more sets of objects and agents from different domains and construct a plausible scenario incorporating both sets. Score higher for more interaction across domains.

Approach for Designing I-athlon Events

Given the requirement for automatic test generation and scoring, we have explored applying the CAPTCHA (von Ahn et al. 2003) approach to the general design of I-athlon events, and the results are intriguing. CAPTCHA, which stands for "Completely Automated Public Turing test to tell Computers and Humans Apart," was originally conceived as a means to validate human users of websites while restricting programmatic access by bots. By generating warped or otherwise obscured images of words or alphanumeric sequences, the machine or human desiring to access the website had to correctly declare the original sequence of characters that was used to generate the test image, a task that was far beyond the ability of current optical character recognition (OCR) programs or other known image processing algorithms. Over time, an arms race of sorts has evolved, with systems learning to crack various CAPTCHA schemes, which in turn has driven the development of more difficult CAPTCHA images. The effectiveness or security of CAPTCHA-based human interaction proofs (HIPs) is not our interest here, but an explicit side effect of the evolution of CAPTCHA technology is: once an existing CAPTCHA-style test is passed by a system, an advance has been achieved in AI. We feel that by applying this approach to other dimensions of intelligence we can motivate and sustain continual progress in achieving human-level AI and beyond.

There are several keys to developing a good CAPTCHA-style test, many of which have to do with its application as a cryptographic security measure. For our purposes, however, we are only concerned with the generalization of the approach for automated test generation and scoring where both humans and machines can compete directly, not for any security applications. For the original CAPTCHA images consisting of warped and obscured text, the generation script was designed to create any number of testable images, and the level of obscuration was carefully matched to what was relatively easy for most humans while being nearly impossible for machines. This pattern can be followed to develop I-athlon event tests by keeping the test scenario the same each time but varying the amount of information provided or the amount of noise in that information for each level of proficiency. This approach could be adapted for many of the dimensions of intelligence described above.

For I-athlon events, the generation algorithms must also be able to produce any number of distinct test scenarios, but at different levels of sophistication that will require different levels of intelligence to succeed, four levels for human achievement and a fifth for superhuman. It would also be important for the generation algorithms to produce identical tests based from a given seed value. This would allow for efficient documentation of the tests generated as well as provide for experimental repeatability by different researchers. We anticipate that both the definition of each event, the design of its standard test generator, and the scoring system and levels will be active areas of research and debate. We include in this article a brief outline for several events to demonstrate the idea. Since the goal of the I-athlon is continual coordinated progress toward the goals of AI, all this effort adds significantly to our understanding of intelligence as well as our ability to add intelligence to computer systems.

To support automatic test generation and scoring for an event, the key is to construct the test so that a small number of variables can programmatically drive a large number of variant test cases that directly map to clear levels of intelligent human ability.

Providing human baselines for these events can be obtained through crowdsourcing, incentivizing large numbers of humans to take the tests, probably through mobile apps. This raises the requirement for an I-athlon event to provide appropriate interfaces for both human and machine contestants.

Examples

Some examples include events that involve simple planning, video understanding, embodiment, and object identification.

A Simple Planning Event

For example, consider an I-athlon event for planning based on a blocks world. An entire genre of two-dimensional physics-based mobile apps already generates puzzles of this type for humans.[4] Size, shape, initial location, and quantity of blocks for each test

can be varied, along with the complexity of the environment (gravity, wind, earthquakes) of the goal state. For a blocks world test, the goal would likely be reaching a certain height or shape with the available blocks, with extra points given for using fewer blocks to reach the goal in fewer attempts. Providing a completed structure as a goal might be too easy, unless ability to manipulate blocks through some virtual device is also a part of the test. Automatic scoring could be based on the test environment reaching a state that passes the constraints of the goal, which could be straightforward programming for a blocks world but likely more challenging for other aspects of intelligence. The test interface could be a touch-based graphical interface for humans and a REST API for machines.

A Video Understanding Event

Given a set of individual video frames in random order, discover the original order by analyzing content and context. Vary the "chunk size" of ordered frames randomized to produce the test. Decimate the quality of the video by masking or adding noise. Scoring could be based on the fraction of frames correctly assembled in order within a time limit, or the total time to complete the task.

An Embodiment Event

Given a sensor/effector API to an embodied agent in a virtual environment, complete a task in the environment using the sensory/motor abilities of the agent. Vary the number and kinds of sensors and effectors. Vary the complexity of the task and the nature of the environment. Environments could be a limited as ChipWits[5] or as open ended as MineCraft.[6] A more sophisticated event would include potential identification and use of tools, or the ability to adapt to gaining or losing sensors and effectors during the test.

An Object-Identification Event

Given recent advances applying DNNs to object recognition, one might think this event would not be interesting. But human visual intelligence allows us to recognize millions of distinct objects in many thousands of classes, and the breadth of this ability is important for general intelligence. This event would generate test images by mixing and overlaying partial images from a very large collection of sources. Scoring would be based on the number of correctly identified objects per image and the time required per image and per test.

Competition Framework and Ecosystem

Our goal to motivate coordinated effort toward the goal of AI requires not only a standard set of events, test generators, and scorers, but also an overall framework for public competition and comparison of results in an unbiased manner. Given the large number of successful industrywide competitions in different areas of computer science and engineering, we propose taking key aspects of each and combining them into a shared platform of ongoing I-athlon competitions.

Sites like Graph 500[7] provide an excellent model for test generation and common scoring. A common cloud-hosted platform for developing and running events and for archiving tests and results will be required, even if competitors run their systems on their own resources. A central location for running the competitions would help limit bias and would also provide wider publicity for successes. Having such a persistent platform along with automated test generation and scoring would support the concept of continuous competition, allowing new entrants at any time with an easy on-ramp to the AI research community. Continuous competitions can prequalify participants in head-to-head playoffs held concurrently with major AI conferences, similar to the RoboCup[8] competitions.

In addition to the professional and graduate-level research communities, such a framework could support competitions at undergraduate and secondary school levels. Extensive programming and engineering communities have been created using this approach, with TopCoder[9] and First Robotics[10] as prime examples. These not only serve a valuable mentoring role in the development of skills, but also recruit high-potential students into the AI research effort.

Incentives beyond eminence and skill building also have proven track records for motivating progress. The X-Prize[11] approach has proven to be highly successful in focusing research attention, as have the DARPA Challenges[12] for self-driving vehicles and legged robots. Presenting a unified, organized framework for progress in AI would go a long way to attract this kind of incentive funding.

The division of labor made possible by the proposed approach could fit nicely within the research agendas of numerous universities at all levels, supporting common curricula development in AI and supporting research programs targeted at different aspects of the I-athlon ecosystem.

Call to Action

We welcome feedback and collaboration from the broad research community to develop and administer a continuing series of I-athlon events according to the model proposed in this article. Our ultimate goal is to motivate the AI research community to understand and develop research agendas that get to the core of general machine intelligence. As we know from the history of AI, this is such a complex problem with so many yet-unknown dimensions, that

the only way to make measurable progress is to develop rigorous, practical, yet flexible tests that require the use of multiple dimensions. The tests themselves can evolve, as we understand the nature of intelligence. We look forward to making progress in the AI field through such an activity.

Notes

1. See, for example, the Kapor-Kurzweil bet: longbets.org/1/#terms.
2. www.olympic.org/athletics-decathlon-men.
3. www.decathlon2000.com/upload/file/pdf/scoringtables.pdf.
4. For example, en.m.wikipedia.org/wiki/The_Incredible_Machine_%28series%29, www.crayonphysics.com.
5. www.chipwits.com/.
6. minecraft.net.
7. www.graph500.org.
8. www.robocup.org.
9. www.topcoder.com.
10. www.usfirst.org.
11. www.xprize.org.
12. www.darpa.mil/about/history/archives.aspx.

References

Turing, A. 1950. Computing Machinery and Intelligence. *Mind* 59(236): 433–460. dx.doi.org/10.1093/mind/ LIX.236.433

von Ahn, L.; Blum, M.; Hopper, N.; and Langford, J. 2003. CAPTCHA: Using Hard AI Problems for Security. In *Proceedings of the Annual International Conference on the Theory and Applications of Cryptographic Techniques* (EUROCRYPT-03). Carson City, NV: International Association for Cryptologic Research. dx.doi.org/10.1007/3-540-39200-9_18

Sam S. Adams (ssadams@us.ibm.com) works for IBM Research and was appointed one of IBM's first distinguished engineers in 1996. His far-ranging contributions include founding IBM's first object technology practice, authoring IBM's XML technical strategy, originating the concept of service-oriented architecture, pioneering work in self-configuring and autonomic systems, artificial general intelligence, end-user mashup programming, massively parallel many-core programming, petascale analytics, and data-centered systems. Adams is currently working on cloud-scale cognitive architectures for the Internet of Things, and has particular interests in artificial consciousness and autocognitive systems.

Guruduth Banavar, as vice president of cognitive computing at IBM Research, currently leads a global team of researchers creating the next generation of IBM's Watson systems — cognitive systems that learn, reason, and interact naturally with people to perform a variety of knowledge-based tasks. Previously, as the chief technical officer of IBM's Smarter Cities initiative, he designed and implemented big data and analytics-based systems to make cities more livable and sustainable. Prior to that, he was the director of IBM Research in India, which he helped establish as a preeminent center for services research and mobile computing. He has published extensively, holds more than 25 patents, and his work has been featured in the *New York Times*, the Wall Street Journal, the *Economist*, and other international media. He received a Ph.D. from the University of Utah before joining IBM's Thomas J. Watson Research Center in 1995.

Murray Campbell is a principal research staff member at the IBM Thomas J. Watson Research Center in Yorktown Heights, NY. He was a member of the team that developed Deep Blue, the first computer to defeat the human world chess champion in a match. Campbell has conducted research in artificial intelligence and computer chess, with numerous publications and competitive victories, including eight computer chess championships. This culminated in the 1997 victory of the Deep Blue chess computer, for which he was awarded the Fredkin Prize and the Allen Newell Research Excellence Medal. He has a Ph.D. in computer science from Carnegie Mellon University, and is an ACM Distinguished Scientist and a Fellow of the Association for the Advancement of Artificial Intelligence. He currently manages the AI and Optimization Department at IBM Research.