My Computer Is an Honor Student — But How Intelligent Is It? Standardized Tests as a Measure of AI

Peter Clark, Oren Etzioni

■ *Given the well-known limitations of the Turing test, there is a need for objective tests to both focus attention on, and measure progress toward, the goals of AI. In this paper we argue that machine performance on standardized tests should be a key component of any new measure of AI, because attaining a high level of performance requires solving significant AI problems involving language understanding and world modeling — critical skills for any machine that lays claim to intelligence. In addition, standardized tests have all the basic requirements of a practical test: they are accessible, easily comprehensible, clearly measurable, and offer a graduated progression from simple tasks to those requiring deep understanding of the world. Here we propose this task as a challenge problem for the community, summarize our state-of-the-art results on math and science tests, and provide supporting data sets (www.allenai.org).*

Alan Turing (Turing 1950) approached the abstract question *can machines think?* by replacing it with another, namely can a machine pass the *imitation game* (the Turing test). In the years since, this test has been criticized as being a poor replacement for the original enquiry (for example, Hayes and Ford [1995]), which raises the question: what would a better replacement be? In this article, we argue that standardized tests are an effective and practical assessment of many aspects of machine intelligence, and should be part of any comprehensive measure of AI progress.

While a crisp definition of machine intelligence remains elusive, we can enumerate some general properties we might expect of an intelligent machine. The list is potentially long (for example, Legg and Hutter [2007]), but should at least include the ability to (1) answer a wide variety of questions, (2) answer complex questions, (3) demonstrate commonsense and world knowledge, and (4) acquire new knowledge scalably. In addition, a suitable test should be clearly measurable, graduated (have a variety of levels of difficulty), not gameable, ambitious but realistic, and motivating.

There are many other requirements we might add (for example, capabilities in robotics, vision, dialog), and thus any comprehensive measure of AI is likely to require a battery of different tests. However, standardized tests meet a surprising number of requirements, including the four listed, and thus should be a key component of a future battery of tests. As we will show, the tests require answering a wide variety of questions, including those requiring commonsense and world knowledge. In addition, they meet all the practical requirements, a huge advantage for any component of a future test of AI.

Science and Math as Challenge Areas

Standardized tests have been proposed as challenge problems for AI, for example, Bringsjord and Schimanski (2003), Bringsjord (2011), Beyer et al. (2005), Fujita et al. (2014), as they appear to require significant advances in AI technology while also being accessible, measurable, understandable, and motivating. They also enable us easily to compare AI performance with that of humans.

In our own work, we have chosen to focus on elementary and high school tests (for 6–18 year olds) because the basic language-processing requirements are surmountable, while the questions still present formidable challenges for solving. Similarly, we are focusing on science and math tests, and have recently achieved some baseline results on these tasks (Seo et al. 2015, Koncel-Kedziorski et al. 2015, Khot et al. 2015, Li and Clark 2015, Clark et al. 2016). Other groups have attempted higher level exams, such as the Tokyo entrance exam (Strickland 2013), and more specialized psychometric tests such as SAT word analogies (Turney 2006), GRE word antonyms (Mohammad et al. 2013), and TOEFL synonyms (Landauer and Dumais 1997).

We also stipulate that the exams are taken exactly as written (no reformulation or rewording), so that the task is clear, standard, and cannot be manipulated or gamed. Typical questions from the New York Regents 4th grade (9–10 year olds) science exams, SAT math questions, and more are shown in the next section. We have also made a larger collection of challenge questions drawn from these and other exams, available on our web site.[1]

We propose to leverage standardized tests, rather than synthetic tests such as the Winograd schema (Levesque, Davis, and Morgenstern 2012) or MCTest (Richardson, Burges, and Renshaw 2013), because they provide a natural sample of problems and more directly suggest real-world applications in the areas of education and science.

Exams and Intelligence

One pertinent question concerning the suitability of exams is whether they are gameable, that is, answerable without requiring any real understanding of the world. For example, questions might be answered with a simple search-engine query or through simple corpus statistics, without requiring any understanding of the underlying material. Our experience is that while some questions are answerable in this way, many are not. There is a continuum from (computationally) easy to difficult questions, where more difficult questions require increasingly sophisticated internal models of the world. This continuum is highly desirable, as it means that there is a low barrier to entry, allowing researchers to make initial inroads into the task, while significant AI challenges need to be solved to do well in the exam. The diversity of questions also ensures a variety of skills are tested for, and guards against finding a simple shortcut that may answer them all without requiring any depth of understanding. (This contrasts with the more homogeneous Winograd schema challenge [Levesque, Davis, and Morgenstern 2012], where the highly stylized question format risks producing specialized solution methods that have little generality). We illustrate these properties throughout this article.

In addition, 45–65 percent of the regents science exam questions (depending on the exam), and virtually all SAT geometry questions, contain diagrams that are necessary for solving the problem. Similarly, the answers to algebraic word problems are typically four numbers (see, for example, table 1). In all these cases, a Google search or simple corpus statistics will not answer these questions with any degree of reliability.

A second important question, raised by Davis in his critique of standardized tests for measuring AI (Davis 2014), is whether the tests are measuring the right thing. He notes that standardized tests are authored for people, not machines, and thus will be testing for skills that people find difficult to master, skipping over things that are easy for people but challenging for machines. In particular, Davis conjectures that "standardized tests do not test knowledge that is obvious for people; none of this knowledge can be assumed in AI systems." However, our experience is generally contrary to this conjecture: although questions do not typically test basic world knowledge directly, basic commonsense knowledge is frequently required to answer them. We will illustrate this in detail throughout this article.

The New York Regents Science Exams

One of the most interesting and appealing aspects of elementary science exams is their graduated and multifaceted nature: Different questions explore different types of knowledge and vary substantially in difficulty (for a computer), from a simple lookup to those requiring extensive understanding of the world. This allows incremental progress while still demanding significant advances for the most difficult questions. Information retrieval and bag-of-words methods work well for a subset of questions but eventually reach a limit, leaving a collection of questions requiring deeper understanding. We illustrate some of this variety here, using (mainly) the multiple choice part of the New York Regents 4th Grade Science exams[2] (New York State Education Department 2014). For a more detailed analysis, see Clark, Harrison, and Balasubramanian (2013). A similar analysis can be made of exams at other grade levels and in other subjects.

Basic Questions

Part of the New York Regents exam tests for relatively straightforward knowledge, such as taxonomic ("isa") knowledge, definitional (terminological) knowledge, and basic facts about the world. Example questions include the following.

> (1) Which object is the best conductor of electricity? (A) a wax crayon (B) a plastic spoon (C) a rubber eraser (D) an iron nail

> (2) The movement of soil by wind or water is called (A) condensation (B) evaporation (C) erosion (D) friction

> (3) Which part of a plant produces the seeds? (A) flower (B) leaves (C) stem (D) roots

This style of question is amenable to solution by information-retrieval methods and/or use of existing ontologies or fact databases, coupled with linguistic processing.

Simple Inference

Many questions are unlikely to have answers explicitly written down anywhere, from questions requiring a relatively simple leap from what might be already known to questions requiring complex modeling and understanding. An example requiring (simple) inference follows:

> (4) Which example describes an organism taking in nutrients? (A) dog burying a bone (B) A girl eating an apple (C) An insect crawling on a leaf (D) A boy planting tomatoes in the garden

Answering this question requires knowledge that eating involves taking in nutrients, and that an apple contains nutrients.

More Complex World Knowledge

Many questions appear to require both richer knowledge of the world, and appropriate linguistic knowledge to apply it to a question. As an example, consider the following question:

> (5) Fourth graders are planning a roller-skate race. Which surface would be the best for this race? (A) gravel (B) sand (C) blacktop (D) grass

Strong cooccurrences between sand and surface, grass and race, and gravel and graders (road-smoothing machines), throw off information-retrieval-based guesses. Rather, a more reliable answer requires knowing that a roller-skate race involves roller skating, that roller skating is on a surface, that skating is best on a smooth surface, and that blacktop is smooth. Obtaining these fragments of world knowledge and integrating them correctly is a substantial challenge.

As a second example, consider the following question:

> (6) A student puts two identical plants in the same type and amount of soil. She gives them the same amount of water. She puts one of these plants near a sunny window and the other in a dark room. This experiment tests how the plants respond to (A) light (B) air (C) water (D) soil

Again, information-retrieval methods and word correlations do poorly. Rather, a reliable answer requires recognizing a model of experimentation (perform two tasks, differing in only one condition), knowing that being near a sunny window will expose the plant to light, and that a dark room has no light in it.

As a third example, consider this question:

> (7) A student riding a bicycle observes that it moves faster on a smooth road than on a rough road. This happens because the smooth road has (A) less gravity (B) more gravity (C) less friction (D) more friction

A reliable processing of this question requires envisioning and comparing two different situations, overlaying a simple qualitative model on the situations described (smoother → less friction → faster). It also requires basic knowledge that bicycles move, and that riding propels a bicycle.

All the aforementioned examples require general knowledge of the world, as well as simple science knowledge. In addition, some questions more directly test basic commonsense knowledge, such as the following:

> (8) A student reaches one hand into a bag filled with smooth objects. The student feels the objects but does not look into the bag. Which property of the objects can the student most likely identify? (A) shape (B) color (C) ability to reflect light (D) ability to conduct electricity

This question requires, among other things, knowing that touch detects shape, and that sight detects color.

Some questions require selecting the best explanation for a phenomenon, requiring a degree of metareasoning. For example, consider the following question:

> (9) Apple trees can live for many years, but bean plants usually live for only a few months. This statement suggests that (A) different plants have different life spans (B) plants depend on other plants (C) plants produce many offspring (D) seasonal changes help plants grow

This requires not just determining whether the statement in each answer option is true (here, several of them are), but whether it explains the statement given in the body of the question. Again, this kind of question would be challenging for a retrieval-based solution.

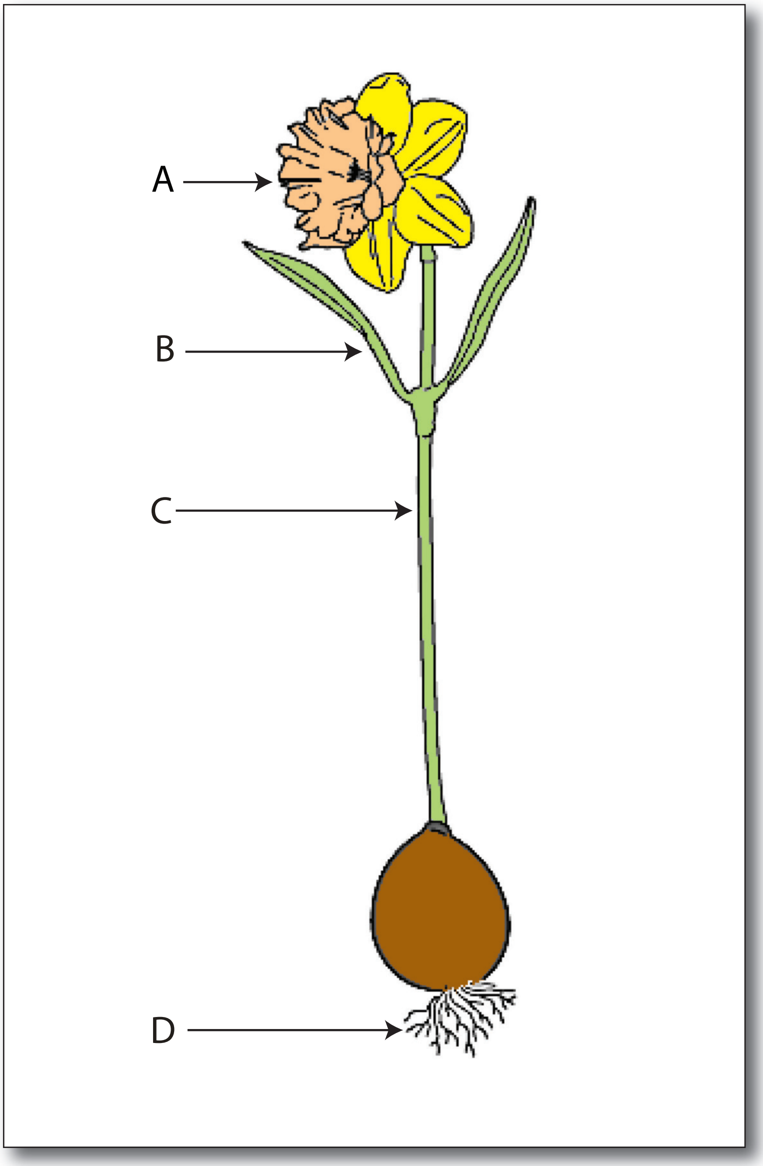


Figure 1. Question 11.

(11) Which letter in the diagram points to the plant structure that takes in water and nutrients?

As a final example, consider the following question from the Texas Assessment of Knowledge and Skills exam[3] (Texas Education Agency 2014):

> (10) Which of these mixtures would be easiest to separate? (A) Fruit salad (B) Powdered lemonade (C) Hot chocolate (D) Instant pudding

This question requires a complex interplay of basic world knowledge and language to answer correctly.

Diagrams

A common feature of many elementary grade exams is the use of diagrams in questions. We choose to include these in the challenge because of their ubiquity in tests, and because spatial interpretation and reasoning is such a fundamental aspect of intelligence. Diagrams introduce several new dimensions to question-answering, including spatial interpretation and correlating spatial and textual knowledge. Diagrammatic (nontextual) entities in elementary exams include sketches, maps, graphs, tables, and diagrammatic representations (for example, a food chain). Reasoning requirements include sketch interpretation, correlating textual and spatial elements, and mapping diagrammatic representations (graphs, bar charts, and so on) to a form supporting computation. Again, while there are many challenges, the level of difficulty varies widely, allowing a graduated plan of attack. Two examples are shown. The first, question 11 (figure 1), requires sketch interpretation, part identification, and label/part correlation. The second, question 12 (figure 2), requires recognizing and interpreting a spatial representation.

Mathematics and Geometry

We also include elementary mathematics in our challenge scope, as these questions intrinsically require mapping to mathematical models, a key requirement for many real-world tasks. These questions are particularly interesting as they combine elements of language processing, (often) story interpretation, mapping to an internal representation (for example, algebra), and symbolic computation. For example (from ixl.com):

> (13) Molly owns the Wafting Pie Company. This morning, her employees used 816 eggs to bake pumpkin pies. If her employees used a total of 1339 eggs today, how many eggs did they use in the afternoon?

Such questions clearly cannot be answered by information retrieval, and instead require symbolic processing and alignment of textual and algebraic elements (for example, Hosseini et al. 2014; Koncel-Kedziorski et al. 2015; Seo et al. 2014, 2015) followed by inference. Additional examples are shown in table 1.

Note that, in addition to simple arithmetic capabilities, some capacity for world modeling is often needed. Consider, for example, the following two questions:

> (14) Sara's high school won 5 basketball games this year. They lost 3 games. How many games did they play in all?

> (15) John has 8 orange balloons, but lost 2 of them. How many orange balloons does John have now?

Both questions use the word *lost,* but the first question maps to an addition problem (5 + 3) while the second maps to a subtraction problem (8 – 2). This illustrates how modeling the entities, events, and event sequences is required, in addition to basic algebraic skills.

Finally we also include geometry questions, as these combine both arithmetic and diagrammatic reasoning together in challenging ways. For example,

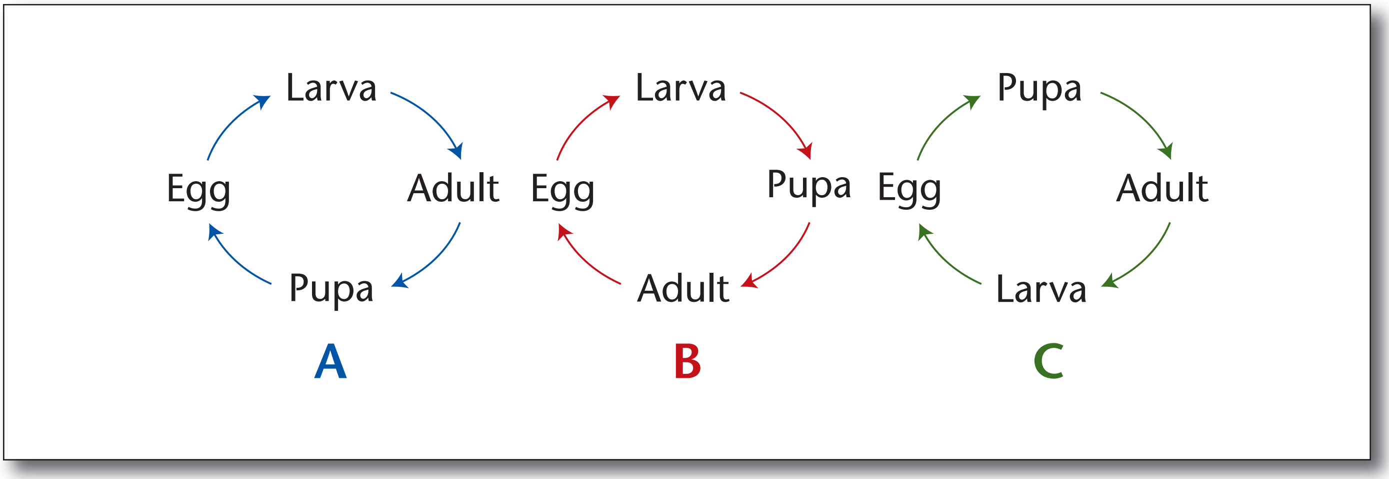


Figure 2. Question 12.

(2) Which diagram correctly shows the life cycle of some insects?

Problems and Equations
John had 20 stickers. He bought 12 stickers from a store in the mall and got 20 stickers for his birthday. Then John gave 5 of the stickers to his sister and used 8 to decorate a greeting card. How many stickers does John have left? $((20 + ((12 + 20) - 8)) - 5) = x$
Maggie bought 4 packs of red bouncy balls, 8 packs of yellow bouncy balls, and 4 packs of green bouncy balls. There were 10 bouncy balls in each package. How many bouncy balls did Maggie buy in all? $x = (((4 + 8) + 4) * 10)$
Sam had 79 dollars to spend on 9 books. After buying them he had 16 dollars. How much did each book cost? $79 = ((9 * x) + 16)$
Fred loves trading cards. He bought 2 packs of football cards for \$2.73 each, a pack of Pokemon cards for \$4.01, and a deck of baseball cards for \$8.95. How much did Fred spend on cards? $((2 * 2.73) + (4.01 + 8.95)) = x$

Table 1. Examples of Problems Solved By Alges with the Returned Equation.

(From Koncel-Kedziorski et al. [2015])

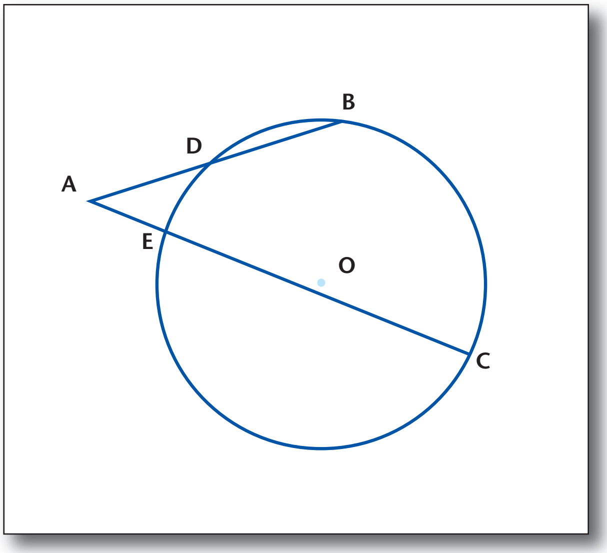


Figure 3. Question 16.

(16) In the diagram, AB intersects circle O at D, AC intersects circle O at E, AE = 4, AC = 24, and AB = 16. Find AD.

question 16 (figure 3) requires multiple skills (text processing, diagram interpretation, arithmetic, and aligning evidence from both text and diagram together). Although very challenging, there has been significant progress in recent years on this kind of problem (for example, Koncel-Kedziorski et al. [2015]). Examples of problems that current systems have been able to solve are shown in table 2.

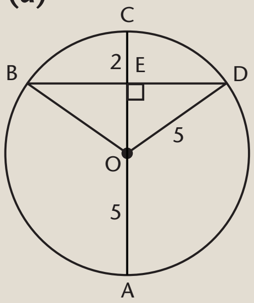
Testing for Commonsense

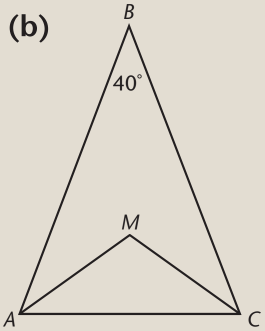
Possessing and using commonsense knowledge is a central property of intelligence (Davis and Marcus 2015). However, Davis (2015) and Weston et al. (2015) have both argued that standardized tests do not test "obvious" commonsense knowledge, and hence are less suitable as a test of machine intelligence. For instance, using their examples, the following questions are unlikely to occur in a standardized test:

> Can you make a watermelon fit into a bag by folding the watermelon?
>
> If you look at the moon then shut your eyes, can you still see the moon?
>
> If John is in the playground and Bob is in the office, then where is John?
>
> Can you make a salad out of a polyester shirt?

However, although such questions may not be

Questions		Interpretations

(a)

	In the diagram at the left, circle O has a radius of 5, and CE = 2. Diameter AC is perpendicular to chord BD. What is the length of BD?	*Equals(RadiusOf(O), 5)*
		IsCircle(O)
		Equals(LengthOf(CE), 2)
		IsDiameter(AC)
		IsChord(BD)
		Perpendicular(AC), BD)
		Equals(what, Length(BD))
		correct
		a) 12 b) 10 [✓ c) 8] d) 6 e) 4

(b)

	In isosceles triangle ABC at the left, lines AM and CM are the angle bisectors of angles BAC and BCA. What is the measure of angle AMC?	*IsIsoscelesTriangle(ABC)*
		BisectsAngle(AM, BAC)
		IsLine(AM)
		CC(AM, CM)
		CC(BAC, BCA)
		IsAngle(BAC)
		IsAngle(AMC)
		Equals(what, MeasureOf(AMC))
		correct
		[✓ a) 110] b) 115 c) 120 d) 125 e) 130

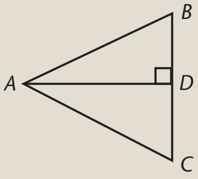
(c)

	In the figure at left, the bisector of angle BAC is perpendicular to BC at point D. If AB = 6 and BD = 3, what is the measure of angle BAC?	*IsAngle(BAC)*
		BisectsAngle(line, BAC)
		Perpendicular(line, BC)
		Equals(LengthOf(AB), 6)
		Equals(LengthOf(BD), 3)
		IsAngle(BAC)
		Equals(what, MeasureOf(BAC))
		correct
		a) 15 b) 30 c) 45 [✓ d) 60] e) 75

Table 2. Examples of Problems That Current Systems Have Solved.

Questions (left) and interpretations (right) leading to correct solution by GEOS. From Seo et al. (2015).

directly posed in standardized tests, many questions indirectly require at least some of this commonsense knowledge in order to answer. For example, question (6) (about plants) in the previous section requires knowing (among other things) that if you put a plant near X (a window), then the plant will be near X. This is a flavor of blocks-world-style knowledge very similar to that tested in many of Weston et al.'s examples. Similarly question (8) (about objects in a bag) requires knowing that touch detects shape, and that not looking implies not being able to detect color. It also requires knowing that a bag filled with objects contains those objects; a smooth object is smooth; and if you feel something, you touch it. These commonsense requirements are similar in style to many of Davis's examples. In short, at least some of the standardized test questions seem to require the kind of obvious commonsense knowledge that Davis and Weston et al. call for in order to derive an answer, even if the answers themselves are less obvious. Conversely, if one authors a set of synthetic commonsense questions, there is a significant risk of biasing the set toward one's own preconceived notions of what *commonsense* means, ignoring other important aspects. (This has been a criticism sometimes made of the Winograd schema challenge.) For this reason we feel that the natural diversity present in standardized tests, as illustrated here, is highly beneficial, along with their other advantages.

Other Aspects of Intelligence

Standardized tests clearly do not test all aspects of intelligence, for example, dialog, physical tasks, speech. However, besides question-answering and reasoning there are some less obvious aspects of intelligence they also push on: explanation, learning and reading, and dealing with novel problems.

Explanation

Tests (particularly at higher grade levels) typically include questions that not only ask for answers but also for explanations of those answers. So, at least to some degree, the ability to explain an answer is required.

Learning and Reading

Reddy (1996) proposed the grand AI challenge of reading a chapter of a textbook and answering the questions at the end of the chapter. While standardized tests do not directly test textbook reading, they do include question comprehension, including sometimes long story questions. In addition, acquiring the knowledge necessary to pass a test will arguably require breakthroughs in learning and machine reading; attempts to encode the requisite knowledge by hand have to date been unsuccessful.

Dealing with Novel Problems

As our examples illustrate, test taking is not a monolithic skill. Rather it requires a battery of capabilities and the ability to deploy them in potentially novel and unanticipated ways. In this sense, test taking requires, to some level, a degree of versatility and the ability to handle new and surprising problems that we would expect of an intelligent machine.

State of the Art on Standardized Tests

How well do current systems perform on these tests? While any performance figure will be exam specific, we can provide some example data points from our own research.

On nondiagram, multiple choice science questions (NDMC), our Aristo system currently scores on average 75 percent (4th grade), 63 percent (8th grade), and 41 percent (12th grade) on (previously unseen) New York Regents science exams (NDMC questions only, typically four-way multiple choice). As can be seen, questions become considerably more challenging at higher grade levels. On a broader multistate collection of 4th grade NDMC questions, Aristo scores 65 percent (unseen questions). The data sets are available at allenai.org/aristo.html. Note that these are the easier questions (no diagrams, multiple choice); other question types pose additional challenges as we have described. No system to date comes even close to passing a full 4th grade science exam.

On algebraic story problems such as those in table 1, our AlgeS system scores over 70 percent accuracy on story problems that translate into single equations (Koncel-Kedziorski et al. 2015). Kushman et al. (2014) report results on story problems that translate to simultaneous algebraic equations. On geometry problems such as those in table 2, our GeoS system achieves a 49 percent score on (previously unseen) official SAT questions, and a score of 61 percent on a data set of (previously unseen) SAT-like practice questions. The relevant questions, data, and software are available on the Allen Institute's website.[4]

Summary

If a computer were able to pass standardized tests, would it be intelligent? Not necessarily, but it would demonstrate that the computer had several critical skills we associate with intelligence, including the ability to answer sophisticated questions, handle natural language, and solve tasks requiring extensive commonsense knowledge of the world. In short, it would mark a significant achievement in the quest toward intelligent machines. Despite the successes of data-driven AI systems, it is imperative that we make progress in these broader areas of knowledge, modeling, reasoning, and language if we are to make the next generation of knowledgable AI systems a reality. Standardized tests can help to drive and measure progress in this direction as they present many of these challenges, yet are also accessible, comprehensible, incremental, and easily measurable, To help with this, we are releasing data sets related to this challenge.

In addition, in October 2015 we launched the Allen AI Science Challenge,[5] a competition run on kaggle .com to build systems to answer eighth-grade science questions. The competition attracted over 700 participating teams, and scores jumped from 32.5 percent initially to 58.8 percent by the end of January 2016. Athough the winner is not yet known at press time, this successful impact demonstrates the efficacy of standardized tests to focus attention and research on these important AI problems.

Of course, some may claim that existing data-driven techniques are all that is needed, given enough data and computing power; if that were so, that in itself would be a startling result. Whatever your bias or philosophy, we encourage you to prove your case, and take these challenges!

AI2's data sets are available on the Allen Institute's website.[5]

Notes

1. www.allenai.org.

2. www.nysedregents.org/Grade4/Science/home.html .

3. tea.texas.gov/student.assessment/taks/released-tests/.

4. www.allenai.org/euclid.html.

5. www.allenai.org/2015-science-challenge.html.

6. www.allenai.org/data.html.

References

Bayer, S.; Damianos, L.; Doran, C.; Ferro, L.;

Fish, R.; Hirschman, L.; Mani, I.; Riek, L.; and Oshika, B. 2005. Selected Grand Challenges in Cognitive Science, MITRE Technical Report 05-1218. Bedford MA: The MITRE Corporation.

Bringsjord, S. 2011. Psychometric Artificial Intelligence. *Journal of Experimental and Theoretical Artificial Intelligence* (JETAI) 23(3): 271–277.

Bringsjord, S., and Schimanski, B. 2003. What Is Artificial Intelligence? Psychometric AI as an Answer. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence,* 887–893. San Francisco: Morgan Kaufmann Publishers. dx.doi.org/10.1080/0952813X.2010.502314

Clark, P.; Harrison, P.; and Balasubramanian, N. 2013. A Study of the Knowledge Base Requirements for Passing an Elementary Science Test. In AKBC'13: *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*. New York: Association for Computing Machinery. dx.doi.org/10.1145/2509558.2509565

Clark, P.; Etzioni, O.; Khashabi, D.; Khot, T.; Sabharwal, A.; Tafjord, O.; Turney, P. 2016. Combining Retrieval, Statistics, and Inference to Answer Elementary Science Questions. In *Proceedings of the Thirtieth Conference of the Association for the Advancement of Artificial Intelligence*. Menlo Park, CA: AAAI Press.

Davis, E. 2014. The Limitations of Standardized Science Tests as Benchmarks for AI Research. Technical Report, New York University. arXiv Preprint arXiv:1411.1629. Ithaca, NY: Cornell University Library.

Davis, E., and Marcus, G. 2015. Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence. *Communications of the ACM* 58(9): 92–103. dx.doi.org/10.1145/2701413

Fujita, A.; Kameda, A.; Kawazoe, A.; and Miyao, Y. 2014. Overview of Todai Robot Project and Evaluation Framework of its NLP-based Problem Solving. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2014). Paris: European Language Resources Association.

Hayes, P., and Ford, K. 1995. Turing Test Considered Harmful. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence.* San Francisco: Morgan Kaufmann Publishers.

Hosseini, M.; Hajishirzi, H.; Etzioni, O.; and Kushman, N. 2014. Learning to Solve Arithmetic Word Problems with Verb Categorization. In *EMNLP 2014: Proceedings of the Empirical Methods in Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics. dx.doi.org/10.3115/v1/D14-1058

Khot, T.; Balasubramanian, N.; Gribkoff, E.; Sabharwal, A.; Clark P.; and Etzioni, O. 2015. Exploring Markov Logic Networks for Question Answering. In *EMNLP 2015: Proceedings of the Empirical Methods in Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics. dx.doi.org/10.18653/v1/D15-1080

Koncel-Kedziorski, R.; Hajishirzi, H.; Sabharwal, A.; Ang, S. D.; and Etzioni, O. 2015. Parsing Algebraic Word Problems into Equations. *Transactions of the Association for Computational Linguistics,* Volume 3 (2015).

Kushman, N.; Artzi, Y.; Zettlemoyer, L.; and Barzilay, R. 2014. Learning to Automatically Solve Algebra Word Problems. In *EMNLP 2014: Proceedings of the Empirical Methods in Natural Language Processing Conference*. Stroudsburg, PA: Association for Computational Linguistics. dx.doi.org/ 10.3115/v1/p14-1026

Landauer, T. K., and Dumais, S. T. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge. *Psychological Review* 104(2): 211–240. dx.doi.org/10.1037/0033-295X.104.2.211

Legg, S., and Hutter, M. A. 2007. Collection of Definitions of Intelligence. In *Advances in Artificial General Intelligence: Concepts, Architectures, and Algorithms*. Frontiers in Artificial Intelligence and Applications Volume 157. Amsterdam, The Netherlands: IOS Press.

Levesque, H.; Davis, E.; and Morgenstern, L. 2012. The Winograd Schema Challenge. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Thirteenth International Conference* (KR2012), 552–561. Palo Alto: AAAI Press.

Li, Y., and Clark, P. 2015. Answering Elementary Science Questions via Coherent Scene Extraction from Knowledge Graphs. In *EMNLP 2015: Proceedings of the Empirical Methods in Natural Language Processing Conference*. Stroudsburg, PA: Association for Computational Linguistics. dx.doi.org/10.18653/v1/D15-1236

Mohammad, S. M.; Dorr, B. J.; Hirst, G.; and Turney, P. D. 2013. Computing Lexical Contrast. *Computational Linguistics* 39(3): 555–590. dx.doi.org/10.1162/COLI_a_00143

New York State Education Department. 2014. *The Grade 4 Elementary-Level Science Test.* Albany, NY: University of the State of New York.

Reddy, R. 1996. To Dream the Possible Dream. *Communications of the ACM* 39(5): 105–112. dx.doi.org/10.1145/ 229459.233436

Richardson, M.; Burges, C.; and Renshaw, E. 2013. MCTest: A Challenge Dataset for the Machine Comprehension of Text. In *EMNLP 2013: Proceedings of the Empirical Methods in Natural Language Processing Conference*. Stroudsburg, PA: Association for Computational Linguistics. dx.doi.org/10.18653/v1/D15-1171

Seo, M.; Hajishirzi, H.; Farhadi, A.; and Etzioni, O. 2014. Diagram Understanding in Geometry Questions. In *Proceedings of the Twenty-Eighth Conference on Artificial Intelligence* AAAI 2014. Palo Alto, CA: AAAI Press.

Seo, M.; Hajishirzi, H.; Farhadi, A.; Etzioni, O.; and Malcolm, C. 2015. Solving Geometry Problems: Combining Text and Diagram Interpretation. In *EMNLP 2015: Proceedings of the Empirical Methods in Natural Language Processing Conference*. Stroudsburg, PA: Association for Computational Linguistics.

Strickland, E. 2013. Can an AI Get into the University of Tokyo? *IEEE Spectrum* 21 August. dx.doi.org/10.1109/ mspec.2013.6587172

Texas Education Agency. 2014. *Texas Assessment of Knowledge and Skills*. Austin, TX: State of Texas Education Agency.

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

Turney, P. D. 2006. Similarity of Semantic Relations. *Computational Linguistics* 32(3): 379–416. dx.doi.org/10.1162/ coli.2006.32.3.379

Weston, J.; Bordes, A.; Chopra, S.; Mikolov, T.; and Rush, A. 2015. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. arXiv Preprint arXiv:1502.05698v6. Ithaca, NY: Cornell University Library.

Peter Clark is the senior research manager for Project Aristo at the Allen Institute for Artificial Intelligence. His work focuses on natural language processing, machine reasoning, and large knowledge bases, and the interplay among these three areas. He has received several awards including a AAAI Best Paper (1997), Boeing Associate Technical Fellowship (2004), and AAAI Senior Member (2014).

Oren Etzioni is chief executive officer of the Allen Institute for Artificial Intelligence. Beginning in 1991, he was a professor at the University of Washington's Computer Science Department. He has received several awards, including the Robert Engelmore Memorial Award (2007), the IJCAI Distinguished Paper Award (2005), AAAI Fellow (2003), and a National Young Investigator Award (1993).