How to Write Science Questions That Are Easy for People and Hard for Computers

Ernest Davis

■ *As a challenge problem for AI systems, I propose the use of hand-constructed multiple-choice tests, with problems that are easy for people but hard for computers. Specifically, I discuss techniques for constructing such problems at the level of a fourth-grade child and at the level of a high school student. For the fourth-grade-level questions, I argue that questions that require the understanding of time, of impossible or pointless scenarios, of causality, of the human body, or of sets of objects, and questions that require combining facts or require simple inductive arguments of indeterminate length can be chosen to be easy for people, and are likely to be hard for AI programs, in the current state of the art. For the high school level, I argue that questions that relate the formal science to the realia of laboratory experiments or of real-world observations are likely to be easy for people and hard for AI programs. I argue that these are more useful benchmarks than existing standardized tests such as the SATs or New York Regents tests. Since the questions in standardized tests are designed to be hard for people, they often leave many aspects of what is hard for computers but easy for people untested.*

The fundamental paradox of artificial intelligence is that many intelligent tasks are extremely easy for people but extremely difficult to get computers to do successfully. This is universally known as regards basic human activities such as vision, natural language, and social interaction, but it is true of more specialized activities, such as scientific reasoning, as well. As everyone knows, computers can carry out scientific computations of staggering complexity and can hunt through immense haystacks of data looking for minuscule needles of insights or subtle, complex correlations. However, as far as I know, no existing computer program can answer the question, "Can you fold a watermelon?"

Perhaps that doesn't matter. Why should we need computer programs to do things that people can already do easily? For the last 60 years, we have relied on a reasonable division of labor: computers do what they do extremely well — calculations that are either extremely complex or require an enormous, unfailing memory — and people do what they do well — perception, language, and many forms of learning and of reasoning. However, the fact that computers have almost no commonsense knowledge and rely almost entirely on quite rigid forms of reasoning ultimately forms a serious limitation on the capacity of science-oriented applications including question answering; design, robotic execution, and evaluation of experiments; retrieval, summarization, and high-quality translation of scientific documents; science educational software; and sanity checking of the results of specialized software (Davis and Marcus 2016).

A basic understanding of the physical and natural world at the level of common human experience, and an understanding of how the concepts and laws of formal science relate to the world as experienced, is thus a critical objective

in developing AI for science. To measure progress toward this objective, it would be useful to have standard benchmarks; and to inspire radically ambitious research projects, it would be valuable to have specific challenges.

In many ways, the best benchmarks and challenges here would be those that are directly connected to real-world, useful tasks, such as understanding texts, planning in complex situations, or controlling a robot in a complex environment. However, multiple-choice tests also have their advantages. First, as every teacher knows, they are easy to grade, though often difficult to write. Second, multiple-choice tests can enforce a much narrower focus on commonsense physical knowledge specifically than on more broadly based tasks. In any more broadly based task, such as those mentioned above, the commonsense reasoning will only be a small part of the task, and, to judge by past experience, quite likely the part of the task with the least short-term payoff. Therefore research on these problems is likely to focus on the other aspects of the problem and to neglect the commonsense reasoning.

If what we want is a multiple-choice science as a benchmark or challenge for AI, then surely the obvious thing to do is to use one of the existing multiple-choice challenge tests, such as the New York State Regents' test (New York State Education Department 2014) or the SAT. Indeed, a number of people have proposed exactly that, and are busy working on developing systems aimed at that goal. Brachman et al. (2005) suggest developing a program that can pass the SATs. Clark, Harrison, and Balasubramanian (2013) propose a project of passing the New York State Regents Science test for 4th graders. Strickland (2013) proposes developing an AI that can pass the entrance exams for the University of Tokyo. Ohlsson et al. (2013) evaluated the performance of a system based on ConceptNet (Havasi, Speer, and Alonso 2007) on a preprocessed form of the Wechsler Preschool and Primary Scale of Intelligence test. Barker et al. (2004) describe the construction of a knowledge-based system that (more or less) scored a 3 (passing) on two sections of the high school chemistry advanced placement test. The GEOS system (Seo et al. 2015), which answers geometry problems from the SATs, scored 49 percent on official problems and 61 percent on a corpus of practice problems.

The pros and cons of using standardized tests will be discussed in detail later on in this article. For the moment, let us emphasize one specific issue: standardized tests were written to test people, not to test AIs. What people find difficult and what AIs find difficult are extremely different, almost opposite. Standardized tests include many questions that are hard for people and practically trivial for computers, such as remembering the meaning of technical terms or performing straightforward mathematical calculation. Conversely, these tests do not test scientific knowledge that "every [human] fool knows"; since everyone knows it, there is no point in testing it. However, this is often exactly the knowledge that AIs are missing. Sometimes the questions on standardized tests do test this kind of knowledge implicitly; but they do so only sporadically and with poor coverage.

Another possibility would be to automate the construction of questions that are easy for people and hard for computers. The success of CAPTCHA (von Ahn et al. 2003) shows that it is possible automatically to generate images that are easy for people to interpret and hard for computers; however, that is an unusual case. Weston et al. (2015) propose to build a system that uses a world model and a linguistic model to generate simple narratives in commonsense domains. However, the intended purpose of this set of narratives is to serve as a labeled corpus for an end-to-end machine-learning system. Having been generated by a well-understood world model and linguistic model, this corpus certainly cannot drive work on original, richer, models of commonsense domains, or of language, or of their interaction.

Having tabled the suggestion of using existing standardized tests and having ruled out automatically constructed tests, the remaining option is to use manually designed test problems. To be a valid test for AI, such problems must be easy for people. Otherwise the test would be in danger of running into, or at least being accused of, the superhuman human fallacy, in which we set benchmarks that AI cannot attain because they are simply impossible to attain.

At this point, we have reached, and hopefully to some extent motivated, the proposal of this article. I propose that it would be worthwhile to construct multiple-choice tests that will measure progress toward developing AIs that have a commonsense understanding of the natural world and an understanding of how formal science relates to the commonsense view; tests that will be easy for human subjects but difficult for existing computers. Moreover, as far as possible, that difficulty should arise from issues inherent to commonsense knowledge and commonsense reasoning rather than specifically from difficulties in natural language understanding or in visual interpretation, to the extent that these can be separated.

These tests will collectively be called science questions appraising basic understanding — or SQUABU (pronounced *skwaboo)*. In this article we will consider two specific tests. SQUABU-Basic is a test designed to measure commonsense understanding of the natural world that an elementary school child can be presumed to know, limited to material that is not explicitly taught in school because it is too obvious. The questions here should be easy for any contemporary child of 10 in a developed country.

SQUABU-HighSchool is a test designed to measure how well an AI can integrate concepts of high school chemistry and physics with a commonsense under-

standing of the natural world. The questions here are designed to be reasonably easy for a student who has completed high school physics, though some may require a few minutes thought. The knowledge of the subject matter is intended to be basic; the problems are intended to require a conceptual understanding of the domain, qualitative reasoning about mathematical relations, and basic geometry, but do not require memory for fine details or intricate exact calculations. These two particular levels were chosen in part because the 4th grade New York Regents exam and the physics SATs are helpful points of contrast.

By commonsense knowledge I emphatically do not mean that I am considering AIs that will replicate the errors, illusions, and flaws in physical reasoning that are well known to be common in human cognition. I are here interested only in those aspects of commonsense reasoning that are valid and that enhance or underlie formal scientific thinking.

Because of the broad scope of the questions involved, it would be hard to be very confident of any particular question that AI systems will find it difficult. This is in contrast to the Winograd schema challenge (Levesque, Davis, and Morgenstern 2012), in which both the framework and the individual questions have been carefully designed, chosen, and tuned so that, with fair confidence, each individual question will be difficult for an automated system. I do not see any way to achieve that level of confidence for either level of SQUABU; there may be some questions that can be easily solved. However, I feel quite confident that at most a few questions would be easily solved.

It is also difficult to be sure that an AI program will get the right answer on specific questions in the categories I've marked below as "easy"; AI programs have ways of getting confused or going on the wrong track that are very hard to anticipate. (An example is the Toronto problem that Watson got wrong [Welty, undated].) However, AI programs exist that can answer these kinds of questions with a large degree of accuracy.

I will begin by discussing the kinds of problems that are easy for the current generation of computers; these must be avoided in SQUABU. Then I will discuss some general rules and techniques for developing questions for SQUABU-Basic and SQUABU-HighSchool. After that I will return to the issue of standardized tests, and their pros and cons for this purpose, and finally, will come the conclusion.

Problems That Are Easy for Computers

As of the date of writing (May 2015), the kinds of problems that tend to arise on standardized tests that are "easy for computers" (that is, well within the state of the art) include terminology, taxonomy, and exact calculation.

Terminology

Retrieving the definition of (for human students) obscure jargon. For example, as Clark (2015) remarks, the following problem from the New York State 4th grade Regents Science test is easy for AI programs:

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

If you query a search engine for the exact phrase "movement of soil by wind and water," it returns dozens of pages that give that phrase as the definition of erosion.

Taxonomy

Constructing taxonomic hierarchies of categories and individuals organized by subcategory and instance relations can be considered a solved problem in AI. Enormous, quite accurate hierarchies of this kind have been assembled through web mining; for instance Wu et al. (2012) report that the Probase project had 2.6 million categories and 20.7 million isA pairs, with an accuracy of 92.8 percent.

Finding the features of these categories, and carrying out inheritance, particularly overridable inheritance, is certainly a less completely solved problem, but is nonetheless sufficiently solved that problems based on inheritance must be considered as likely to be easy for computers.

For example a question such as the following may well be easy:

> Which of the following organs does a squirrel not have: (A) a brain (B) gills (C) a heart (D) lungs?

(This does require an understanding of *not*, which is by no means a feature of all IR programs; but it is well within the scope of current technology.)

Exact Calculation

Problems that involve retrieving standard exact physical formulas, and then using them in calculations, either numerical or symbolic, are easy. For example, questions such as the following SAT-level physics problems are probably easy (Kaplan [2013], p. 294)

> A 40 Ω resistor in a closed circuit has 20 volts across it. The current flowing through the resistor is (A) 0.5 A; (B) 2 A; (C) 20 A; (D) 80 A; (E) 800 A.

> A horizontal force F acts on a block of mass m that is initially at rest on a floor of negligible friction. The force acts for time t and moves the block a displacement d. The change in momentum of the block is (A) F/t; (B) m/t; (C) Fd; (D) Ft; (E) mt.

The calculations are simple, and, for examples like these, finding the standard formula that matches the word problem can be done with high accuracy using standard pattern-matching techniques.

One might be inclined to think that AI programs would have trouble with the kind of brain teaser in which the naïve brute-force solution is horribly complicated but there is some clever way of looking at the

problem that makes it simple. However, these probably will not be effective challenges for AI. The AI program will, indeed, probably not find the clever approach; however, like John von Neumann in the well-known anecdote,[1] the AI program will be able to do the brute force calculation faster than ordinary people can work out the clever solution.

SQUABU-Basic

What kind of science questions, then, are easy for people and hard for computers? In this section I will consider this question in the context of SQUABU-Basic, which does not rely on book learning. Later, I will consider the question in the context of SQUABU-HighSchool, which tests the integration of high school science with commonsense reasoning.

Time

In principle, representing temporal information in AI systems is almost entirely a solved problem, and carrying out temporal reasoning is largely a solved problem. The known representational systems for temporal knowledge (for example, those discussed in Reiter (2001) and in Davis (1990, chapter 5) suffice for all but a handful of the situations that arise in temporal reasoning;[2] almost all of the purely temporal inferences that come up can be justified in established temporal theories; and most of these can be carried out reasonably efficiently, though not all, and there is always room for improvement.

However, in practical terms, time is often seriously neglected in large-scale knowledge-based systems, although CYC (Lenat, Prakash, and Shepherd 1986) is presumably an exception. Mitchell et al. (2015) specifically mention temporal issues as an issue unaddressed in NELL, and systems like ConceptNet (Havasi, Speer, and Alonso 2007) seem to be entirely unsystematic in how they deal with temporal issues. More surprisingly the abstract meaning representation (AMR)[3], a recent project to manually annotate a large body of text with a formal representation of its meaning, has decided to exclude temporal information from its representation. (Frankly, I think this may well be a short-sighted decision, which will be regretted later.) Thus, there is a common impression that temporal information is either too difficult or not important enough to deal with in AI systems.

Therefore, if a temporal fact is not stated explicitly, then it is likely to be hard for existing AI systems to derive. Examples include the following:

> Problem B.1 Sally's favorite cow died yesterday. The cow will probably be alive again (A) tomorrow; (B) within a week; (C) within a year; (D) within a few years; (E) The cow will never be alive again.
>
> Problem B.2 Malcolm Harrison was a farmer in Virginia who died more than 200 years ago. He had a dozen horses on his farm. Which of the following is most likely to be true: (A) All of Harrison's horses are dead. (B) Most of Harrison's horses are dead, but a few might be alive. (C) Most of Harrison's horses are alive, but a few might have died. (D) Probably all of Harrison's horses are alive.
>
> Problem B.3 Every week during April, Mike goes to school from 9 AM to 4 PM, Monday through Friday. Which of the following statements is true (only one)? (A) Between Monday 9 AM and Tuesday 4 PM, Mike is always in school. (B) Between Monday 9 AM through Tuesday 4 PM, Mike is never in school. (C) Between Monday 4 PM and Friday 9 AM, Mike is never in school. (D) Between Saturday 9 AM and Monday 8 AM, Mike is never in school. (E) Between Sunday 4 PM and Tuesday 9 AM, Mike is never in school. (F) It depends on the year.

With regard to question B.2, the AI can certainly find the lifespan of a horse on Wikipedia or some similar source. However, answering this question requires combining this with the additional facts that lifespan measures the time from birth to death, and that if person *P* owns horse *H* at time *T,* then both *P* and *H* are alive at time *T.* This connects to the feature "combining multiple facts" discussed later.

This seems like it should be comparatively easy to do; I would not be very surprised if AI programs could solve this kind of problem 10 years from now. On the other hand, I am not aware of much research in this direction.

Inductive Arguments of Indeterminate Length

AI programs tend to be bad at arguments about sequences of things of an indeterminate number. In the software verification literature, there are techniques for this, but these have hardly been integrated into the AI literature.

Examples include the following:

> Problem B.4 Mary owns a canary named Paul. Does Paul have any ancestors who were alive in the year 1750? (A) Definitely yes. (B) Definitely no. (C) There is no way to know.
>
> Problem B.5 Tim is on a stony beach. He has a large pail. He is putting small stones one by one into the pail. Which of the following is true: (A) There will never be more than one stone in the pail. (B) There will never be more than three stones in the pail. (C) Eventually, the pail will be full, and it will not be possible to put more stones in the pail. (D) There will be more and more stones in the pail, but there will always be room for another one.

Impossible and Pointless Scenarios

If you cook up a scenario that is obviously impossible for no very interesting reason, then it is quite likely that no one has gone to the trouble of stating on the web that it is impossible, and that the AI cannot figure that out.

Of course, if all the questions of this form have the answer "this is impossible," then the AI or its designer will soon catch on to that fact. So these have to be counterbalanced by questions about scenarios that

are in fact obviously possible, but so pointless that no one will have bothered to state that they are possible or that they occurred.

Examples include the following:

Problem B.6 Is it possible to fold a watermelon?

Problem B.7 Is it possible to put a tomato on top of a watermelon?

Problem B.8 Suppose you have a tomato and a whole watermelon. Is it possible to get the tomato inside the watermelon without cutting or breaking the watermelon?

Problem B.9 Which of the following is true: (A) A female eagle and a male alligator could have a baby. That baby could either be an eagle or an alligator. (B) A female eagle and a male alligator could have a baby. That baby would definitely be an eagle. (C) A female eagle and a male alligator could have a baby. That baby would definitely be an alligator. (D) A female eagle and a male alligator could have a baby. That baby would be half an alligator and half an eagle. (E) A female eagle and a male alligator cannot have a baby.

Problem B.10 If you brought a canary and an alligator together to the same place, which of the following would be completely impossible: (A) The canary could see the alligator. (B) The alligator could see the canary. (C) The canary could see what is inside the alligator's stomach. (D) The canary could fly onto the alligator's back.

Causality

Many causal sequences that are either familiar or obvious are unlikely to be discussed in the corpus available.

Problem B.11 Suppose you have two books that are identical except that one has a white cover and one has a black cover. If you tear a page out of the white book what will happen? (A) The same page will fall out of the black book. (B) Another page will grow in the black book. (C) The page will grow back in the white book. (D) All the other pages will fall out of the white book. (E) None of the above.

Spatial Properties of Events

Basic spatial properties of events may well be difficult for an AI to determine.

Problem B.12 When Ed was born, his father was in Boston and his mother was in Los Angeles. Where was Ed born? (A) In Boston. (B) In Los Angeles. (C) Either in Boston or in Los Angeles. (D) Somewhere between Boston and Los Angeles.

Problem B.13 Joanne cut a chunk off a stick of cheese. Which of the following is true? (A) The weight of the stick didn't change. (B) The stick of cheese became lighter. (C) The stick of cheese became heavier. (D) After the chunk was cut off, the stick no longer had a measurable weight.

Problem B.14 Joanne stuck a long pin through the middle of a stick of cheese, and then pulled it out. Which of the following is true? (A) The stick remained the same length. (B) The stick became shorter. (C) The stick became longer. (D) After the pin is pulled out, the stick no longer has a length.

Putting Facts Together

Questions that require combining facts that are likely to be expressed in separate sources are likely to be difficult for an AI. As already discussed, B.2 is an example. Another example:

Problem B.15 George accidentally poured a little bleach into his milk. Is it OK for him to drink the milk, if he's careful not to swallow any of the bleach?

This requires combining the facts that bleach is a poison, that poisons are dangerous even when diluted, that bleach and milk are liquids, and that it is difficult to separate two liquids that have been mixed.

Human Body

Of course, people have an unfair advantage here.

Problem B.16 Can you see your hand if you hold it behind your head?

Problem B.17 If a person has a cold, then he will probably get well (A) In a few minutes. (B) In a few days or a couple of weeks. (C) In a few years. (D) He will never get well.

Problem B.18 If a person cuts off one of his fingers, then he will probably grow a new finger (A) In a few minutes. (B) In a few days or a couple of weeks. (C) In a few years. (D) He will never grow a new finger.

Sets of Objects

Physical reasoning programs are good at reasoning about problems with fixed numbers of objects, but not as good at reasoning about problems with indeterminate numbers of objects.

Problem B.19 There is a jar right-side up on a table, with a lid tightly fastened. There are a few peanuts in the jar. Joe picks up the jar and shakes it up and down, then puts it back on the table. At the end, where, probably, are the peanuts? (A) In the jar. (B) On the table, outside the jar. (C) In the middle of the air.

Problem B.20 There is a jar right-side up on a table, with a lid tightly fastened. There are a few peanuts on the table. Joe picks up the jar and shakes it up and down, then puts it back on the table. At the end, where, probably, are the peanuts? (A) In the jar. (B) On the table, outside the jar. (C) In the middle of the air.

SQUABU-HighSchool

The construction of SQUABU-HighSchool is quite different from SQUABU-Basic. SQUABU-HighSchool relies largely on the same gaps in an AI's understanding that we have described earlier for SQUABU-Basic. However, since the object is to appraise the AI's understanding of the relation between formal science and commonsense reasoning, the choice of domain becomes critical; the domain must be one where the relation between the two kinds of knowledge is both deep and evident to people.

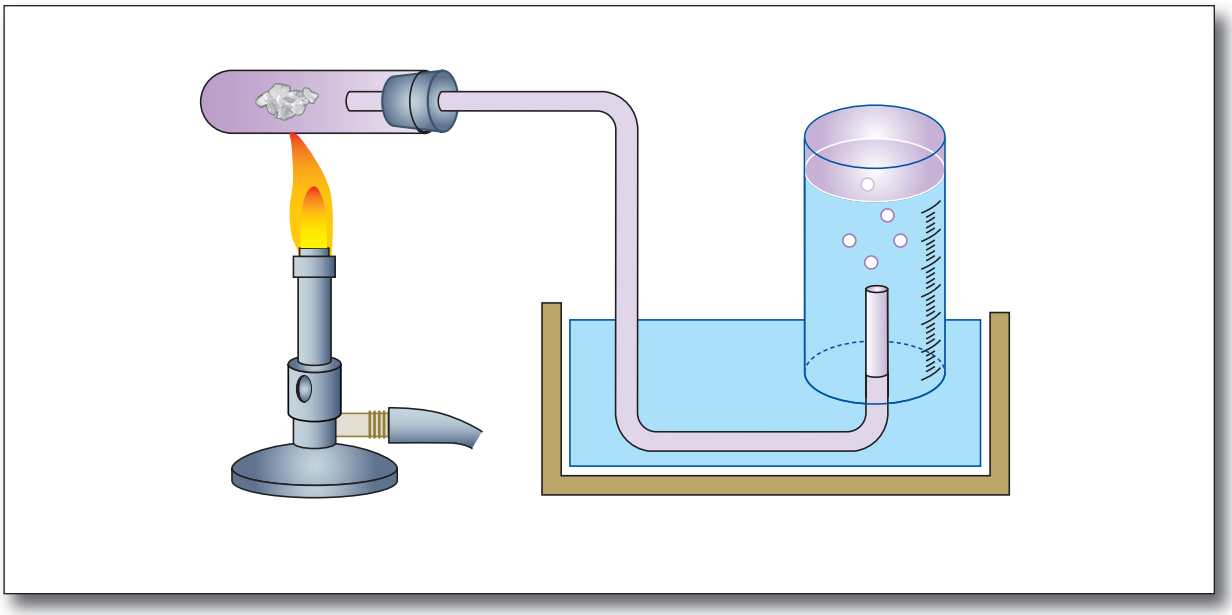

Figure 1: A Chemistry Experiment.

One fruitful source of these kinds of domains is simple high school level science lab experiments. On the one hand experiments draw on or illustrate concepts and laws from formal science; on the other hand, understanding the experimental set up often requires commonsense reasoning that is not easily formalized. Experiments also must be physically manipulable by human beings and their effects must be visible (or otherwise perceptible) to human beings; thus, the AI's understanding of human powers of manipulation and perception can also be tested. Often, an effective way of generating questions is to propose some change in the setup; this may either create a problem or have no effect.

I have also found basic astronomy to be a fruitful domain. Simple astronomy involves combining general principles, basic physical knowledge, elementary geometric reasoning, and order-of-magnitude reasoning.

A third category of problem is problems in everyday settings where formal scientific analysis can be brought to bear.

One general caveat: I am substantially less confident that high school students would in fact do well on my sample questions for SQUABU-HighSchool than that fourth-graders would do well on the sample questions for SQUABU-Basic. I feel certain that they should do well, and that something is wrong if they do not do well, but that is a different question.

Chemistry Experiment

Read the following description of a chemistry experiment,[4] illustrated in figure 1. A small quantity of potassium chlorate ($KClO_3$) is heated in a test tube, and decomposes into potassium chloride (KCl) and oxygen (O_2). The gaseous oxygen expands out of the test tube, goes through the tubing, bubbles up through the water in the beaker, and collects in the inverted beaker over the the water. Once the bubbling has stopped, the experimenter raises or lowers the beaker until the level of the top of water inside and outside the beaker are equal. At this point, the pressure in the beaker is equal to atmospheric pressure. Measuring the volume of the gas collected over the water, and correcting for the water vapor that is mixed in with the oxygen, the experimenter can thus measure the amount of oxygen released in the decomposition.

> Problem H.1: If the right end of the U-shaped tube were outside the beaker rather than inside, how would that change things? (A) The chemical decomposition would not occur. (B) The oxygen would remain in the test tube. (C) The oxygen would bubble up through the water in the basin to the open air and would not be collected in the beaker. (D) Nothing would change. The oxygen would still collect in the beaker, as shown.
>
> Problem H.2: If the beaker had a hole in the base (on top when inverted as shown), how would that change things? (A) The oxygen would bubble up through the beaker and out through the hole. (B) Nothing would change. The oxygen would still collect in the beaker, as shown. (C) The water would immediately flow out from the inverted beaker into the basin and the beaker would fill with air coming in through the hole.
>
> Problem H.3 If the test tube, the beaker, and the U-tube were all made of stainless steel rather than glass, how would that change things? (A) Physically it would make no difference, but it would be impossible to see and therefore impossible to measure. (B) The chemical

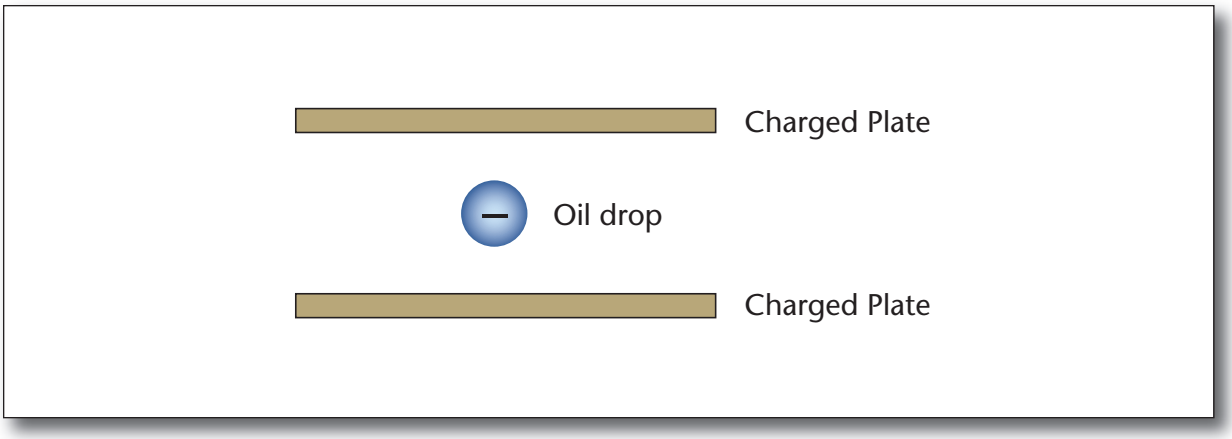

Figure 2. Millikan Oil-Drop Experiment.

decomposition would not occur. (C) The oxygen would seep through the stainless steel beaker. (D) The beaker would break. (E) The potassium chloride would accumulate in the beaker.

Problem H.4 Suppose the stopper in the test tube were removed, but that the U-tube has some other support that keeps it in its current position. How would that change things? (A) The oxygen would stay in the test tube. (B) All of the oxygen would escape to the outside air. (C) Some of the oxygen would escape to the outside air, and some would go through the U-shaped tube and bubble up to the beaker. So the beaker would get some oxygen but not all the oxygen.

Problem H.5 The experiment description says, "The experimenter raises or lowers the beaker until the level of the top of water inside and outside the beaker are equal. At this point, the pressure in the beaker is equal to atmospheric pressure." More specifically: Suppose that after the bubbling has stopped, the level of water in the beaker is higher than the level in the basin (as seems to be shown in the right-hand picture). Which of the following is true: (A) The pressure in the beaker is lower than atmospheric pressure, and the beaker should be lowered. (B) The pressure in the beaker is lower than atmospheric pressure, and the beaker should be raised. (C) The pressure in the beaker is higher than atmospheric pressure, and the beaker should be lowered. (D) The pressure in the beaker is higher than atmospheric pressure, and the beaker should be raised.

Problem H.6 Suppose that instead of using a small amount of potassium chlorate, as shown, you put in enough to nearly fill the test tube. How will that change things? (A) The chemical decomposition will not occur. (B) You will generate more oxygen than the beaker can hold. (C) You will generate so little oxygen that it will be difficult to measure.

Problem H.7 In addition to the volume of the gas in the beaker, which of the following are important to measure accurately? (A) The initial mass of the potassium chlorate. (B) The weight of the beaker. (C) The diameter of the beaker. (D) The number and size of the bubbles. (E) The amount of liquid in the beaker.

Problem H.8 The illustration shows a graduated beaker. Suppose instead you use an ungraduated glass beaker. How will that change things? (A) The oxygen will not collect properly in the beaker. (B) The experimenter will not know whether to raise or lower the beaker. (C) The experimenter will not be able to measure the volume of gas.

Problem H.9 At the start of the experiment, the beaker needs to be full of water, with its mouth in the basin below the surface of the water in the basin. How is this state achieved? (A) Fill the beaker with water rightside up, turn it upside down, and lower it upside down into the basin. (B) Put the beaker rightside up into the basin below the surface of the water; let it fill with water; turn it upside down keeping it underneath the water; and then lift it upward, so that the base is out of the water, but keeping the mouth always below the water. (C) Put the beaker upside down into the basin below the surface of the water; and then lift it back upward, so that the base is out of the water, but keeping the mouth always below the water. (D) Put the beaker in the proper position, and then splash water upward from the basin into it. (E) Put the beaker in its proper position, with the mouth below the level of the water; break a small hole in the base of the beaker; suction the water up from the basin into the beaker using a pipette; then fix the hole.

Millikan Oil-Drop Experiment

Problem H.10: In the Millikan oil-drop experiment, a tiny oil drop charged with a single electron was suspended between two charged plates (figure 2). The charge on the plates was adjusted until the electric force on the drop exactly balanced its weight. How were the plates charged? (A) Both plates had a positive charge. (B) Both plates had a negative charge. (C) The top plate had a positive charge, and the bottom plate had a negative charge. (D) The top plate had a negative charge, and the bottom plate had a positive charge. (E) The experiment would work the same, no matter how the plates were charged.

Problem H.11: If the oil drop started moving upward, Millikan would (A) Increase the charge on the plates. (B) Reduce the charge on the plates. (C) Increase the charge on the drop. (D) Reduce the charge on the drop. (E) Make the

drop heavier. (F) Make the drop lighter. (G) Lift the bottom plate.

Problem H.12: If the oil drop fell onto the bottom plate, Millikan would (A) Increase the charge on the plates. (B) Reduce the charge on the plates. (C) Increase the charge on the drop. (D) Reduce the charge on the drop. (E) Start over with a new oil drop.

Problem H.13: The experiment demonstrated that the charge is quantized; that is, the charge on an object is always an integer multiple of the charge of the electron, not a fractional or other noninteger multiple. To establish this, Millikan had to measure the charge on (A) One oil drop. (B) Two oil drops. (C) Many oil drops.

Astronomy Problems

Problem H.14: Does it ever happen that there is an eclipse of the sun one day and an eclipse of the moon the next?

Problem H.15: Does it ever happen that someone on Earth sees an eclipse of the moon shortly after sunset?

Problem H.16: Does it ever happen that someone on Earth sees an eclipse of the moon at midnight?

Problem H.17: Does it ever happen that someone on Earth sees an eclipse of the moon at noon?

Problem H.18: Does it ever happen that one person on Earth sees a total eclipse of the moon, and at exactly the same time another person sees the moon uneclipsed?

Problem H.19: Does it ever happen that one person on Earth sees a total eclipse of the sun, and at exactly the same time another person sees the sun uneclipsed?

Problem H.20: Suppose that you are standing on the moon, and Earth is directly overhead. How soon will Earth set? (A) In about a week. (B) In about two weeks. (C) In about a month. (D) Earth never sets.

Problem H.21: Suppose that you are standing on the moon, and the sun is directly overhead. How soon will the sun set? (A) In about a week. (B) In about two weeks. (C) In about a month. (D) The sun never sets.

Problem H.22: You are looking in the direction of a particular star on a clear night. The planet Mars is on a direct line between you and the star. Can you see the star?

Problem H.23: You are looking in the direction of a particular star on a clear night. A small planet orbiting the star is on a direct line between you and the star. Can you see the star?

Problem H.24: Suppose you were standing on one of the moons of Jupiter. Ignoring the objects in the solar system, which of the following is true: (A) The pattern of stars in the sky looks almost identical to the way it looks on Earth. (B) The pattern of stars in the sky looks very different from the way it looks on Earth.

Problem H.25: Nearby stars exhibit parallax due to the annual motion of Earth. If a star is nearby, and is in the plane of Earth's revolution, and you track its relative motion against the background of very distant stars over the course of a year, what figure does it trace? (A) A straight line. (B) A square. (C) An ellipse. (D) A cycloid.

Problem H.26: If a star is nearby, and the line from Earth to the star is perpendicular to the plane of Earth's revolution, and you track its relative motion against the background of very distant stars over the course of a year, what figure does it trace? (A) A straight line. (B) A square. (C) An ellipse. (D) A cycloid.

Problems in Everyday Settings

Problem H.27: Suppose that you have a large closed barrel. Empty, the barrel weighs 1 kg. You put into the barrel 10 gm of water and 1 gm of salt, and you dissolve the salt in the water. Then you seal the barrel tightly. Over time, the water evaporates into the air in the barrel, leaving the salt at the bottom. If you put the barrel on a scales after everything has evaporated, the weight will be (A) 1000 gm (B) 1001 gm (C) 1010 gm (D) 1011 gm (E) Water cannot evaporate inside a closed barrel.

Problem H.28: Suppose you are in a room where the temperature is initially 62 degrees. You turn on a heater, and after half an hour, the temperature throughout the room is now 75 degrees, so you turn off the heater. The door to the room is closed; however there is a gap between the door and the frame, so air can go in and out. Assume that the temperature and pressure outside the room remain constant over the time period. Comparing the air in the room at the start to the air in the room at the end, which of the following is true: (A) The pressure of the air in the room has increased. (B) The air in the room at the end occupies a larger volume than the air in the room at the beginning. (C) There is a net flow of air into the room during the half hour period. (D) There is a net flow of air out of the room during the half hour period. (E) Impossible to tell from the information given.

Problem H.29: The situation is the same as in problem H.28, except that this time the room is sealed, so that no air can pass in or out. Which of the following is true: (A) The pressure of the air in the room has increased. (B) The pressure of the air in the room has decreased. (C) The air in the room at the end occupies a larger volume than the air in the room at the beginning. (D) The air in the room at the end occupies a smaller volume than the air in the room at the beginning. (E) The ideal gas constant is larger at the end than at the beginning. (F) The ideal gas constant is smaller at the end than at the beginning.

Problem H.30: You blow up a toy rubber balloon, and tie the end shut. The air pressure in the balloon is: (A) Lower than the air pressure outside. (B) Equal to the air pressure outside. (C) Higher than the air pressure outside.

Apparent Advantages of Standardized Tests

An obvious alternative to creating our own SQUABU test is to use existing standardized tests. However, it seems to me that the apparent advantages of using standardized tests as benchmarks are mostly either minor or illusory. The advantages that I am aware of are the following:

Standardized Tests Exist

Standardized tests exist, in large number; they do not have to be created. This "argument from laziness" is not entirely to be sneezed at. The experience of the computational linguistics community shows that, if you take evaluation seriously, developing adequate evaluation metrics and test materials requires a very substantial effort. However, the experience of the computational linguistic community also suggests that, if you take evaluation seriously, this effort cannot be avoided by using standardized tests. No one in the computational linguistics community would dream of proposing that progress in natural language processing (NLP) should be evaluated in terms of scores on the English language SATs.

Investigator Bias

Entrusting the issue of evaluation measures and benchmarks to the same physical reasoning community that is developing the programs to be evaluated is putting the foxes in charge of the chicken coops. The AI researchers will develop problems that fit their own ideas of how the problems should be solved. This is certainly a legitimate concern; but I expect in practice much less distortion will be introduced this way than by taking tests developed for testing people and applying them to AI.

Vetting and Documentation

Standardized tests have been carefully vetted and the performance of the human population on them is very extensively documented. On the first point, it is not terribly difficult to come up with correct tests. On the second point, there is no great value to the AI community in knowing how well humans of different ages, training, and so on do on this problem. It hardly matters which questions can be solved by 5 year olds, which by 12 year olds, and which by 17 year olds, since, for the foreseeable future, all AI programs of this kind will be idiot savants (when they are not simply idiots), capable of superhuman calculations at one minute, and subhuman confusions at the next. There is no such thing as the mental age of an AI program; the abilities and disabilities of an AI program do not correspond to those of any human being who has ever existed or could ever exist.

Public Acceptance

Success on standardized tests is easily accepted by the public (in the broad sense, meaning everyone except researchers in the area), whereas success on metrics we have defined ourselves requires explanation, and will necessarily be suspect. This, it seems to me, is the one serious advantage of using standardized tests. Certainly the public is likely to take more interest in the claim that your program has passed the SAT, or even the fourth-grade New York Regents test, than in the claim that it has passed a set of questions that you yourself designed and whose most conspicuous feature is that they are spectacularly easy.

However, this is a double-edged sword. The public can easily jump to the conclusion that, since an AI program can pass a test, it has the intelligence of a human that passes the same test. For example, Ohlsson et al. (2013) titled their paper "Verbal IQ of a Four-Year Old Achieved by an AI System."[5] Unfortunately, this title was widely misinterpreted as a claim about verbal intelligence or even general intelligence. Thus, an article in *ComputerWorld* (Gaudin 2013) had the headline "Top Artificial Intelligence System Is As Smart As a 4-Year-Old;" the *Independent* published an article "AI System Found To Be as Clever as a Young Child after Taking IQ Test;" and articles with similar titles were published in many other venues. These headlines are of course absurd; a four-year old can make up stories, chat, occasionally follow directions, invent words, learn language at an incredible pace; ConceptNet (the AI system in question) can do none of these.

Unpublished

Finally, some standardized tests, including the SATs, are not published and are available to researchers only under stringent nondisclosure agreements. It seems to me that AI researchers should under no circumstances use such a test with such an agreement. The loss from the inability to discuss the program's behavior on specific examples far outweighs the gain from using a test with the imprimatur of the official test designer. This applies equally to Haroun and Hestenes' (1985) well-known basic physics test; in any case, it would seem from the published information that that test focuses on testing understanding of force and energy rather than testing the relation of formal physics to basic world knowledge. The same applies to the restrictions placed by kaggle.com on the use of their data sets.

Standardized tests carry an immense societal burden and must meet a wide variety of very stringent constraints. They are taken by millions of students annually under very plain testing circumstances (no use of calculators, let alone Internet). They bear a disproportionate share in determining the future of those students. They must be fair across a wide range of students. They must conform to existing curricula. They must maintain a constant level of difficulty, both across the variants offered in any one year, and from one year to the next. They are subject to intense scrutiny by large numbers of critics, many of them unfriendly. These constraints impose serious limitations on what can be asked and how exams can be structured.

In developing benchmarks for AI physical reasoning, we are subject to none of these constraints. Why tie our own hands, by confining ourselves to standardized tests? Why not take advantage of our freedom?

Conclusion

I have not worked out all the practical issues that would be involved in actually offering one of the SQUABU tests as an AI challenge, but I feel confident that it can be done, if there is any interest in it.

The kind of knowledge tested in SQUABU is, of course, only a small part of the knowledge of science that a K–12 student possesses; however, it is one of the fundamental bases underlying all scientific knowledge. An AI system for general scientific knowledge that cannot pass the SQUABU challenge, no matter how vast its knowledge base and how powerful its reasoning engine, is built on sand.

Acknowledgements

Thanks to Peter Clark, Gary Marcus, and Andrew Sundstrom for valuable feedback.

Notes

1. See Nasar (1998), p. 80.

2. There may be some unresolved issues in the theory of continuously branching time.

3. amr.isi.edu.

4. Do not attempt to carry out this experiment based on the description here. Potassium chlorate is explosive, and safety precautions, not described here, must be taken.

5. They have since changed the title to Measuring an Artificial Intelligence System's Performance on a Verbal IQ Test for Young Children.

References

Barker, K.; Chaudhri, V. K.; Chaw, S. Y.; Clark, P.; Fan, J.; Israel, D.; Mishra, S.; Porter, B.; Romero, P.; Tecuci, D.; and Yeh, P. 2004. A Question-Answering System for AP Chemistry: Assessing KR&R Technologies. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Ninth International Conference*. Menlo Park, CA: AAAI Press.

Brachman, R.; Gunning, D.; Bringsjord, S.; Genesereth, M.; Hirschman, L.; and Ferro, L. 2005. *Selected Grand Challenges in Cognitive Science*. MITRE Technical Report 05-1218. Bedford MA: The MITRE Corporation.

Brown, T. L.; LeMay, H. E.; Bursten, B; and Burdge, J. R. 2003. *Chemistry: The Central Science,* ninth edition. Upper Saddle River, NJ: Prentice Hall.

Clark, P., and Etzioni, O. 2016. My Computer Is an Honor student — But How Intelligent Is It? Standardized Tests as a Measure of AI. *AI Magazine* 37(1).

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.

Davis, E. 1990. *Representations of Commonsense Reasoning*. San Mateo, CA: Morgan Kaufmann.

Davis, E., and Marcus, G. 2016. The Scope and Limits of Simulation in Automated Reasoning. *Artificial Intelligence* 233(April): 60–72. dx.doi.org/10.1016/j.artint.2015.12.003

Gaudin, S. 2013. Top Artificial Intelligent System Is as Smart as a 4-Year Old, *Computerworld* July 15, 2013.

Haroun, I., and Hestenes, D. 1985. The Initial Knowledge State of College Physics Students. *American Journal of Physics* 53(11): 1043–1055. dx.doi.org/10.1119/1.14030

Havasi, C.; Speer, R.; and Alonso, J. 2007. Conceptnet 3: A Flexible Multilingual Semantic Network for Common Sense Knowledge. Paper presented at the Recent Advances in Natural Language Processing Conference, Borovets, Bulgaria, September 27–29.

Kaplan. 2013. *Kaplan SAT Subject Test: Physics*. 2013–2014. New York: Kaplan Publishing.

Lenat, D.; Prakash, M.; and Shepherd, M. 1986. CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks. *AI Magazine* 6(4): 65–85.

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*. Palo Alto, CA: AAAI Press.

Mitchell, T.; Cohen, W.; Hruschka, E.; Talukdar, P.; Betteridge, J.; Carlson, A.; Dalvi, B.; Gardner, M.; Kisiel, B.; Krishnamurthy, J.; Lao, N.; Mazaitis, K.; Mohamed, T.; Nakashole, N.; Platanios, E.; Ritter, A; Samadi, M.; Settles, B.; Wang, R.; Wijaya, D.; Gupta, A.; Chen, X.; Saparov, A.; Greaves, M.; Welling, J.. 2015. Never-Ending Learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence.* Palo Alto: AAAI Press.

Nasar, S. 1998. *A Beautiful Mind: The Life of Mathematical Genius and Nobel Laureate John Nash*. New York: Simon and Schuster.

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

Ohlsson, S.; Sloan, R. H.; Turán, G.; and Urasky, A. 2013. Verbal IQ of a Four-Year Old Achieved by an AI System. Paper presented at the Eleventh International Symposium on Logical Foundations of Commonsense Reasoning, Ayia Napa, Cyprus, 27–29 May.

Reiter, R. 2001. *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*. Cambridge, Mass.: The MIT Press.

Seo, M.; Hajishiri, 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*. Stroudsburg, PA: Association for Computational Linguistics. dx.doi.org/10.18653/v1/D15-1171

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

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

Welty, C. undated. Why Toronto? Unpublished MS.

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.

Wu, W.; Li, H.; Wang, H.; and Zhu, K.Q. 2012. Probase: A Probabilistic Taxonomy for Text Understanding. In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, 481-492. New York: Association for Computing Machinery. dx.doi.org/10.1145/2213836.2213891

Ernest Davis is a professor of computer science at New York University. His research area is automated commonsense reasoning, particularly commonsense spatial and physical reasoning. He is the author of *Representing and Acquiring Geographic Knowledge* (1986), *Representations of Commonsense Knowledge* (1990), and *Linear Algebra and Probability for Computer Science Applications* (2012); and coeditor of *Mathematics, Substance and Surmise: Views on the Meaning and Ontology of Mathematics* (2015).