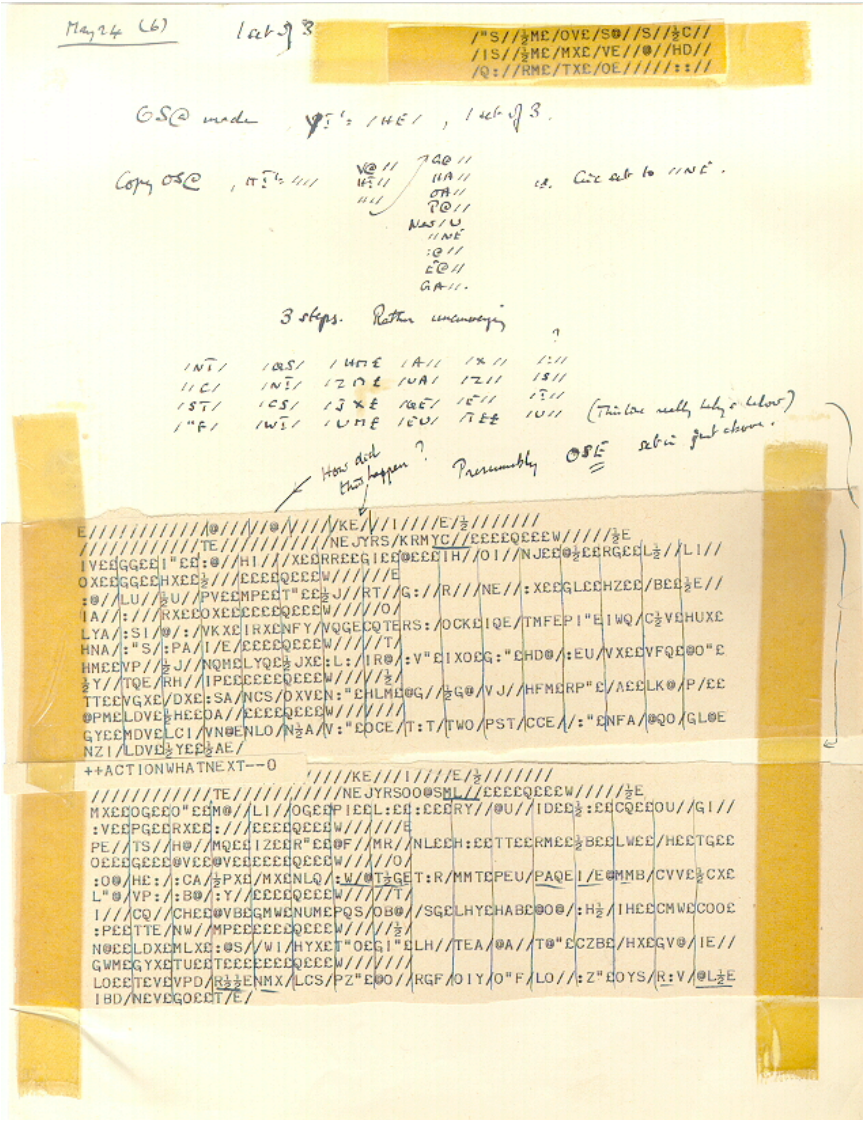
If not Turing's test, then what?

Paul Cohen

USC Information Sciences Institute



Outline

Premise: Good problems help to produce good science

- Review Turing's test – what is helpful and unhelpful about it?
- If not Turing's test, then what?
- Review other "grand challenge" tests
- Helpful attributes of grand challenges

VOL. LIX. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

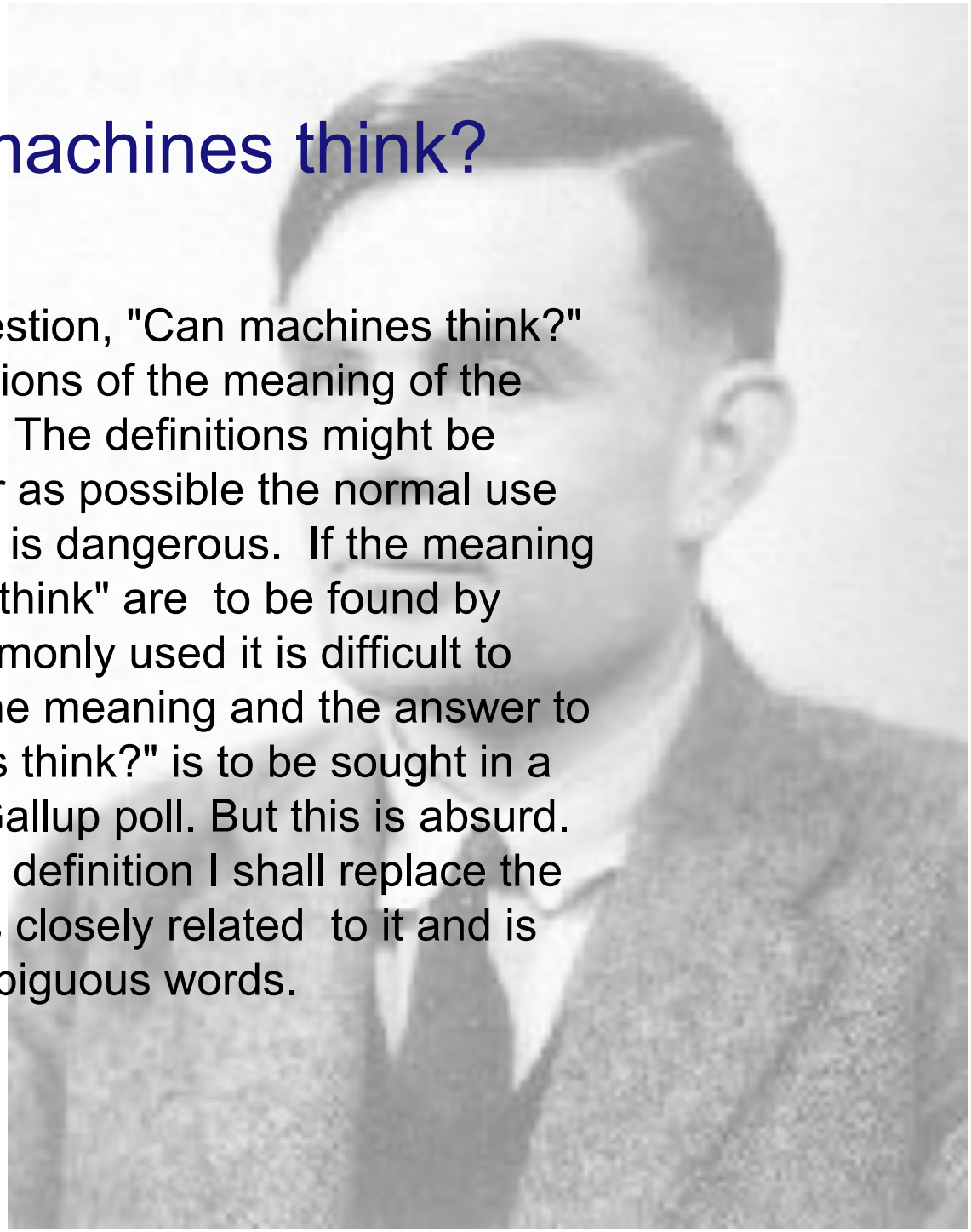
PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

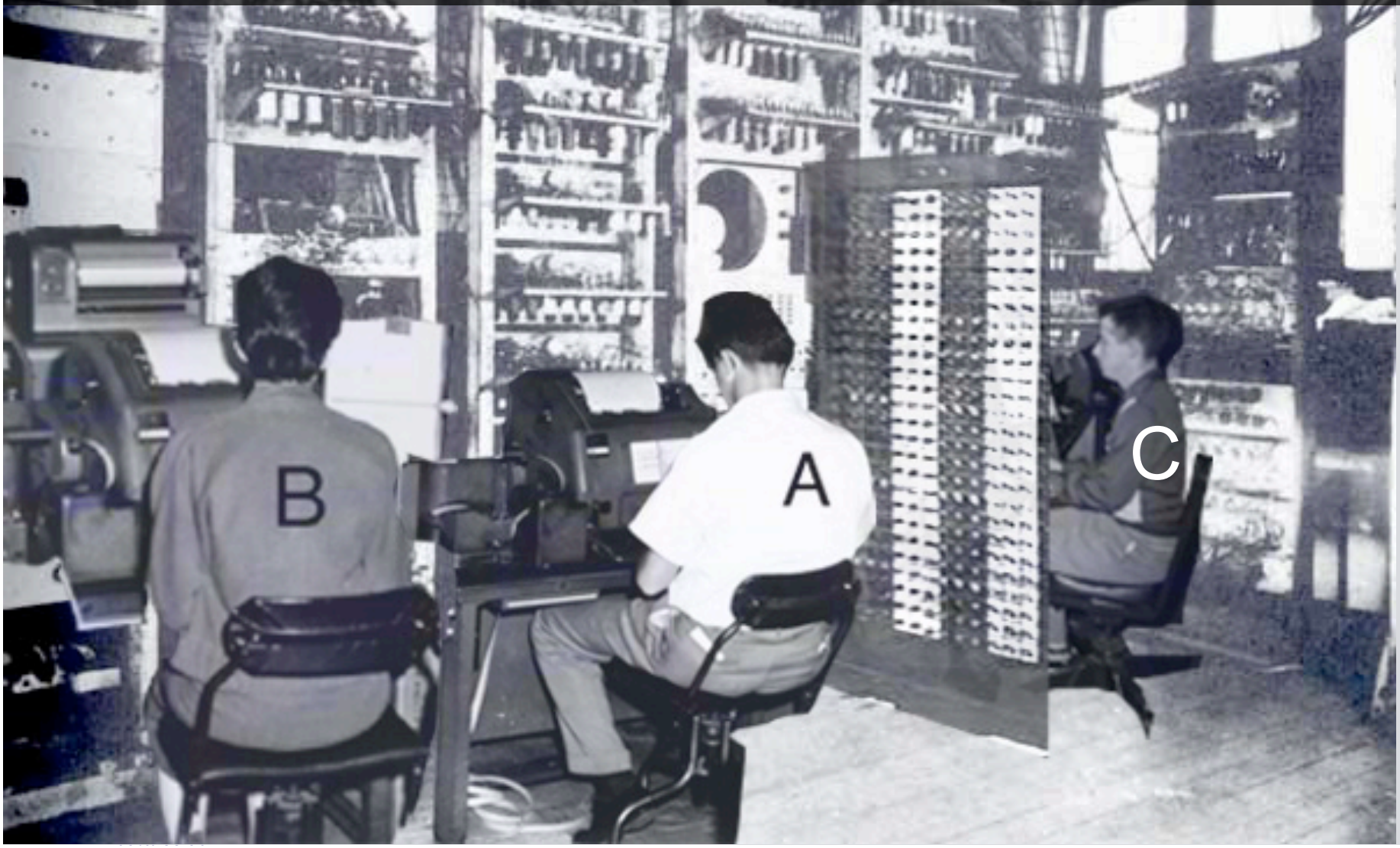
BY A. M. TURING

Can machines think?

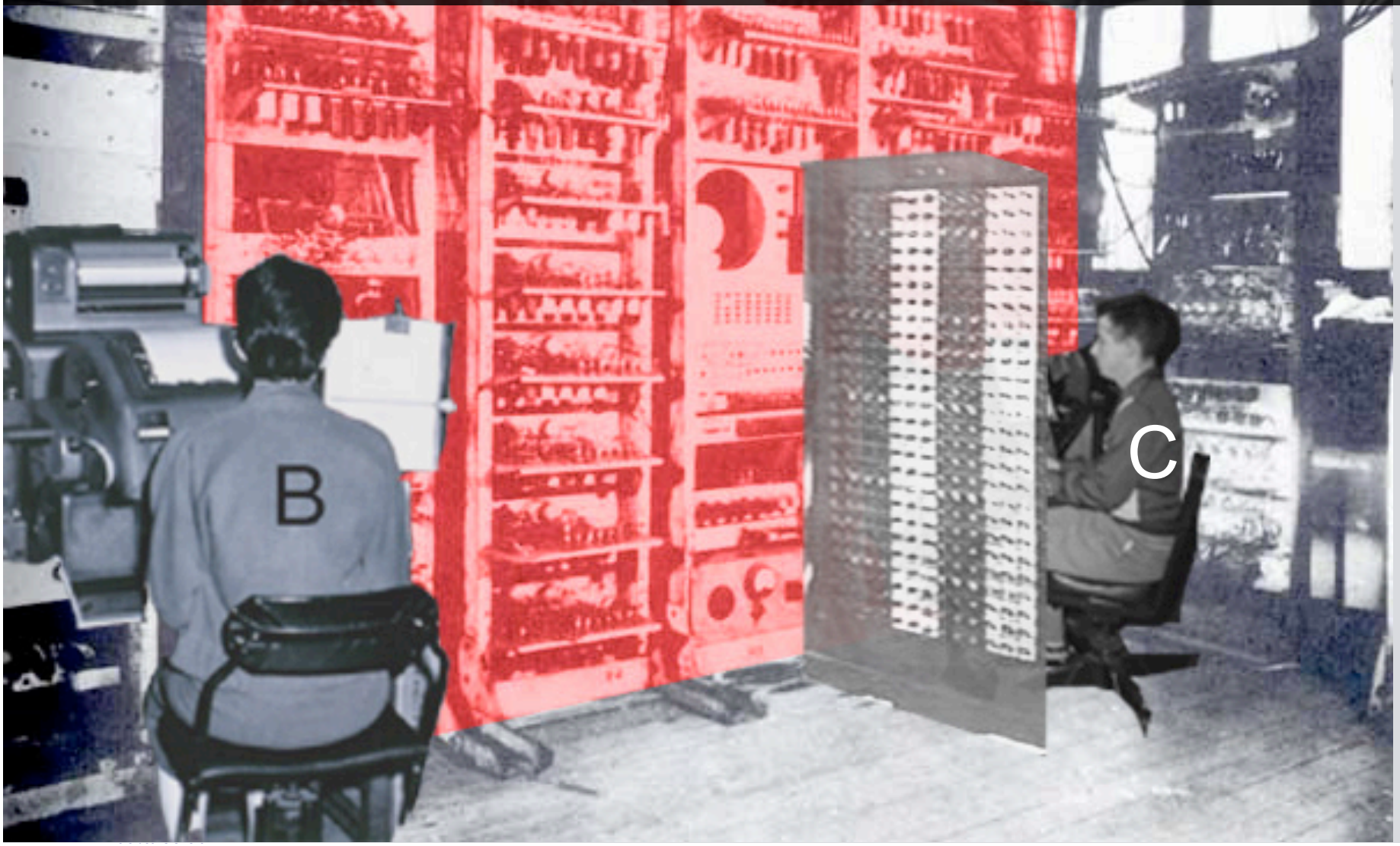
I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words "machine" and "think" are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, "Can machines think?" is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.



The new form of the problem can be described in terms of a game which we call the "imitation game." It is played with three people, a man (A), a woman (B) and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman.



We now ask the question, ``What will happen when a machine takes the part of [the man] in this game?" Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, "Can machines think?"



Kinds of argument about Turing's test

- Irrelevance
- Philosophy
- Methodology

Interpretations of Turing's Test
Is it relevant?

1950 - 1966: A source of inspiration to all concerned with AI.

1966 - 1973: A distraction from some more promising avenues of AI research.

1973 - 1990: By now a source of distraction mainly to philosophers, rather than AI workers.

1990: Consigned to history.

–Blay Whitby, 1997

Yet the test remains a challenge to AI, and if it goes away, what takes its place?

If not Turing's test, then what?

Interpretations of Turing's Test: Philosophy and Methodology

"…a source of distraction mainly to philosophers, rather than AI workers."

> "Alas, philosophers – amateur and professional – have instead taken Turing's proposal as a pretext for just the sort of definitional haggling and interminable arguing about imaginary counterexamples he was hoping to squelch." –Dennett, 1985

A methodological stance: Good tests of intentional capabilities of machines can lead to more capable machines. What makes them good tests?

What makes good tests:
Good and bad attributes of Turing's Test

- The test serves as a proxy for many aspects of intelligence
- The test is intended to assess *several* aspects of intelligence in a *single* session
- The test cannot be passed with today's technology
- The test is not diagnostic: failure to pass it does not point the way to future improvements

"Nothing could possibly pass the Turing test by winning the imitation game without being able to perform indefinitely many other intelligent actions. … [Turing's] test was so severe, he thought, that nothing that could pass it fair and square would disappoint us in other quarters."

– Dan Dennett, 1985

The proxy function

The proxy function: Standing for "more or less" of human intelligence

"Nothing could possibly pass the Turing test by winning the imitation game without being able to perform indefinitely many other intelligent actions. … [Turing's] test was so severe, he thought, that nothing that could pass it fair and square would disappoint us in other quarters."

"Nothing could

… **build decision trees for Irvine Machine Learning datasets …**

… perform word-sense disambiguation for TREC data…

… beat the world chess champion…

… win against the world champion soccer team…

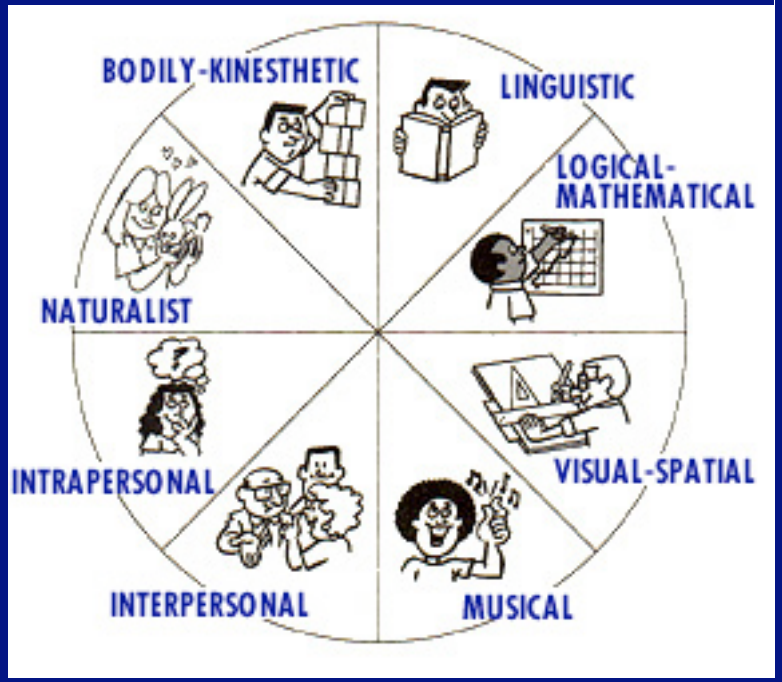
without being able to perform indefinitely many other intelligent actions."

The test is unspecific: It assesses *several* aspects of intelligence in a *single* session

This isn't necessary, we could have several tests of different aspects of intelligence

Multiple intelligences theory – Gardner

Robert French's complaint: "the Test provides a guarantee not of intelligence but of culturally-oriented *human* intelligence."



Not physical intelligence, or perceptual intelligence, not musical intelligence or chess intelligence, not hive intelligence or neonatal intelligence, but mid-century, parlor conversation, human intelligence.

The test cannot be passed today

Where is your nose?

Entrant:...Oh, my nose? Why are you so interested?

Is it near your eyes?

Entrant:...Such is Human Perversity

Is your nose between your mouth and your eyes?

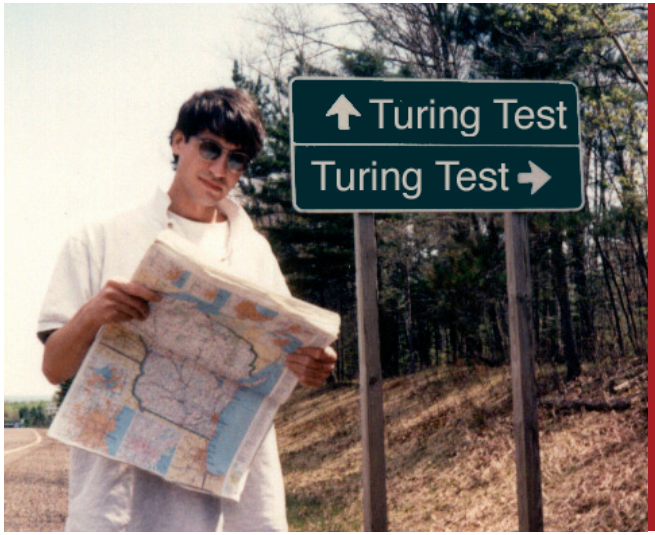
Entrant:...I can't tell you that

How many hands do I have?

Entrant:...I don't understand why you have, and furthermore, I don't really care

My daughter's name is Allegra. What is my daughter's name?

Entrant:...Is she a big part of your problem?



Turing's test is not diagnostic. It doesn't point the way to success

- If the Turing test is read as something like an operational definition of intelligence, then two very important defects of such a test must be considered. First, it is all or nothing: it gives no indication as to what a partial success might look like. Second, it gives no direct indications as to how success might be achieved. – Blay Whitby. 1997

- Remember, failure on the Turing test does not predict failure on … others, but success would surely predict success.

  – Dan Dennett, 1985

The Turing Test is a goal, not a test

- A good test provides

1) Diagnosticity. When you fail, you get information to localize the cause of failure, and when you succeed, you get information to localize the reasons for success.
2) Specificity. Intelligent behaviors are individuated and tested as individual components of a larger system, and should not be mashed together in one unspecific test of general intelligence.
3) A proxy function. Passing a test guarantees that a system can perform other tasks in a class. The Turing Test is a proxy for a big class, but humans do much more.

Because of its proxy function,Turing's test is still a *great* goal!

Movie of conversation with two children goes here

Checklist:

Common sense knowledge	☐
Facility with language	☐
Learning new facts	☐
Inference	☐
Problem solving & planning	☐
Good manners	☐
Sense of humor	☐
Metacognitive knowledge	☐
Desire to make a point	☐
Construct verbal arguments	☐
Listen and understand	☐
Fill in missing bits	☐
Ontology / classification	☐
Memory and attention	☐
	☐

New Challenges

- Handy Andy Report Writing
- Robot Soccer
- Cognitive Decathalon – The Virtual Third-grader
- Learn to read, read to learn
- Robot Baby

- It's easy to pose challenges, everyone has their favorites. Our focus is

Good tests of intentional capabilities of machines can lead to more capable machines. What makes them good tests?

Challenge 1: Handy Andy

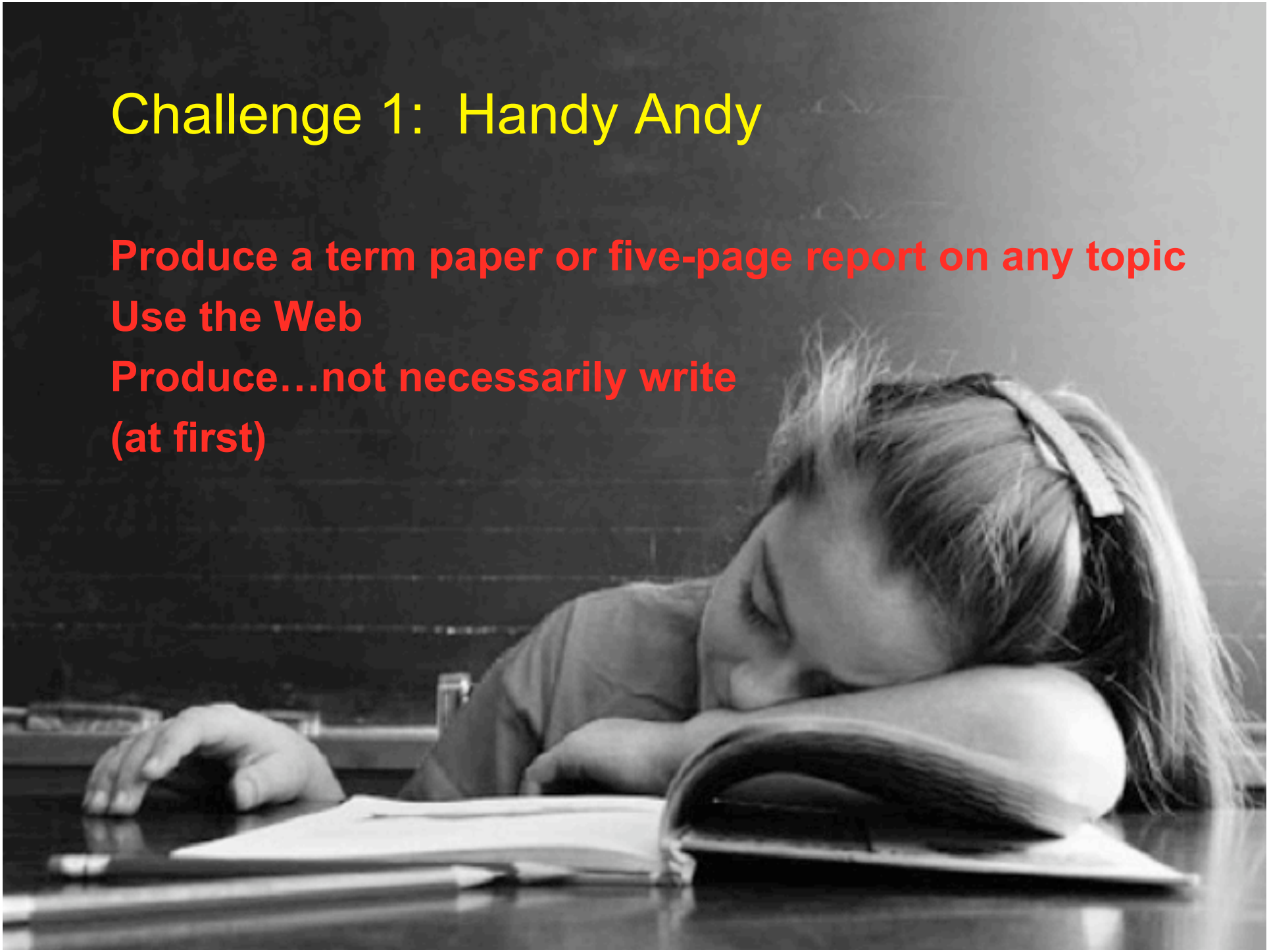
" For the defenders of academic integrity, their nemesis comes in the form of a bright college student at an Eastern university with a 3.78 GPA. Andy — not his real name — writes term papers for his fellow students, at rates of up to $25 a page. …

One student writing a paper for another is not new. But the Internet makes it so easy — and allows someone like Andy to work in volume.

Andy says he's so good that he'll ask a student/client for his or her grade point average. If the client generally gets A's, he'll write an "A" paper. Get B's, and Andy will write you a "B" paper. And so on …. "

http://abcnews.go.com/sections/Primetime/US/cheating_andy_040429-1.html

Challenge 1: Handy Andy



Produce a term paper or five-page report on any topic

Use the Web

Produce…not necessarily write

(at first)

Handy Andy: Graded Challenges

1) Weak comprehension of the query, the report is collated text from a few relevant web pages
2) Stronger comprehension of the query sufficient to excerpt relevant material from relevant web pages
3) Strong comprehension, generate followup questions for user in a formal language, extract material from relevant web pages, make it nonredundant
4) 3 plus *organize* material
5) 4 plus English dialog with user, write a report de novo that contains no sentences from source web pages

Handy Andy
Good Attributes

- Graded challenges – not "all or nothing" – start now, perform poorly, but make the problem harder each year
- Universality – relying on the Web as a near-universal resource
- Come-as-you-are – don't delay, don't wait for better NL or ontologies or language generation, start with what we have
- Comprehension – performance depends on understanding the topic and material on the Web
- Ample rope – five pages is enough rope to hang oneself

Challenge 2: Robot soccer

(thanks to Manuela Veloso)

By the year 2050 develop a team of fully autonomous humanoid robots that can win against the human world soccer champion team

- The community has a clear 50-year goal
- A technical committee is elected and steers toward the goal via rule changes, new leagues, and competitions
- The first competitions were open to all and intentionally "easy"
- Simple success criteria
- No end of good research problems in sight
- Scientific progress is encouraged by awards for Symposium papers
- Transparency – people publish their methods
- Competition motivates students
- Junior League brings kids into the field (in Lisbon, 200 teams)
- The public loves it (150K at Japan Open)

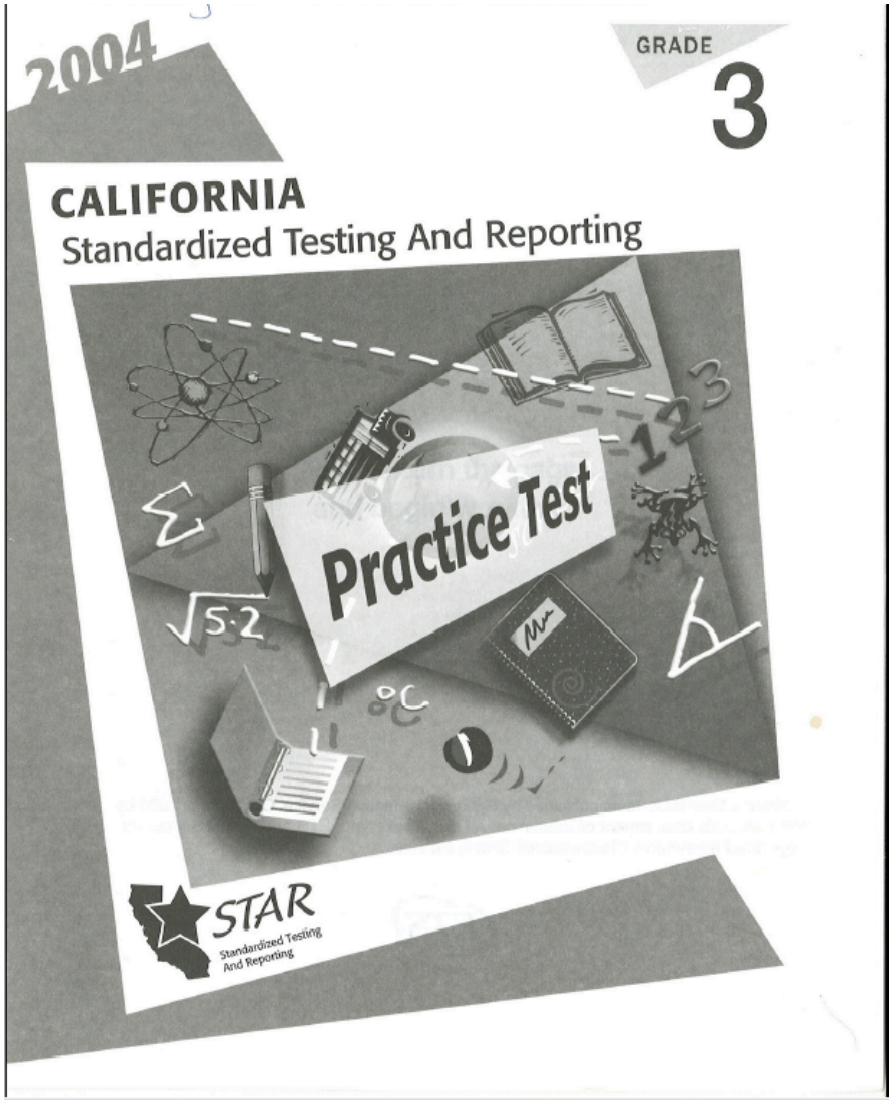
And the crowd went wild!!!!

Manuela Veloso's Robocup team movie goes here.

See http://brim.coral.cs.cmu.edu/small/movies/index.html for sources.

Challenge 3: Cognitive Decathalon or the Virtual Third-Grader

(thanks to Dave Gunning)



- "Qualifying trials" for Turing's test
- Individuates cognitive skills
- Scope of tests can be increased

- Pass the standardized tests administered to third-graders
- Objective scoring

- Common classroom or homework tasks
- Ample rope
- Subjective scoring

Third-grade skills

Understand and follow instructions.

Communicate in natural language (i.e., dialog)

Learn and exercise procedures (e.g., long division, outlining a report).

Read for content (e.g., show that one gets the main points of a story)

Learn by being told (e.g., life was hard for the pioneers).

Common sense inference (e.g., few people wanted to be pioneers) and learning from commonsense inference.

Understand math story problems and solve them correctly.

Master a lot of facts (math facts, history facts, etc.). Mastery means using the facts to answer questions and solve problems.

Prioritize (e.g., choose one book over another, decide to do these problems instead of others on a test).

Explain something (e.g., why plants need light)

Make a convincing argument (e.g., why recess should be longer).

Make up and write a story about an assigned subject (e.g., Thanksgiving)

Third-grade skills

Understand and follow instructions.

Communicate in natural language (i.e., dialog)

Learn and exercise procedures (e.g., long division, outlining a report).

Read for content (e.g., show that one gets the main points of a story)

Learn by being told (e.g., life was hard for the pioneers).

Common sense inference (e.g., fe
and learning from commonsense i

Understand math story problems a

Master a lot of facts (math facts, hi
using the facts to answer question

Prioritize (e.g., choose one book o
problems instead of others on a te

Explain something (e.g., why plant

Make a convincing argument (e.g.

Make up and write a story about a
Thanksgiving)

Testing the Virtual Third Grader

- Robocup Simulation League
- Robocup 4-legged League
- Robocup Humanoid League
- Robocup Coaching League
- Robocup Rescue
- Robocup Junior Dance Competition
- …

- **The creative writing challenge**
- **The convincing letter challenge**
- **The learning procedures challenge**
- **The California STAR challenge**
- **The change of representation challenge**
- **The book report challenge**

V3G: The creative writing challenge

Mothers Are Safe
By Allegra Argent, 3rd Grade

My mother is safe
Almost too safe
When playing tag it's,
"Don't go up there"
"But Mom, I'm IT!"
Mom doesn't care
But now she's not here to hover
and peck
A few days ago I broke my neck.

- Points for creativity
- Points for humor
- Points for topicality
- Points for worrying the teacher
- …

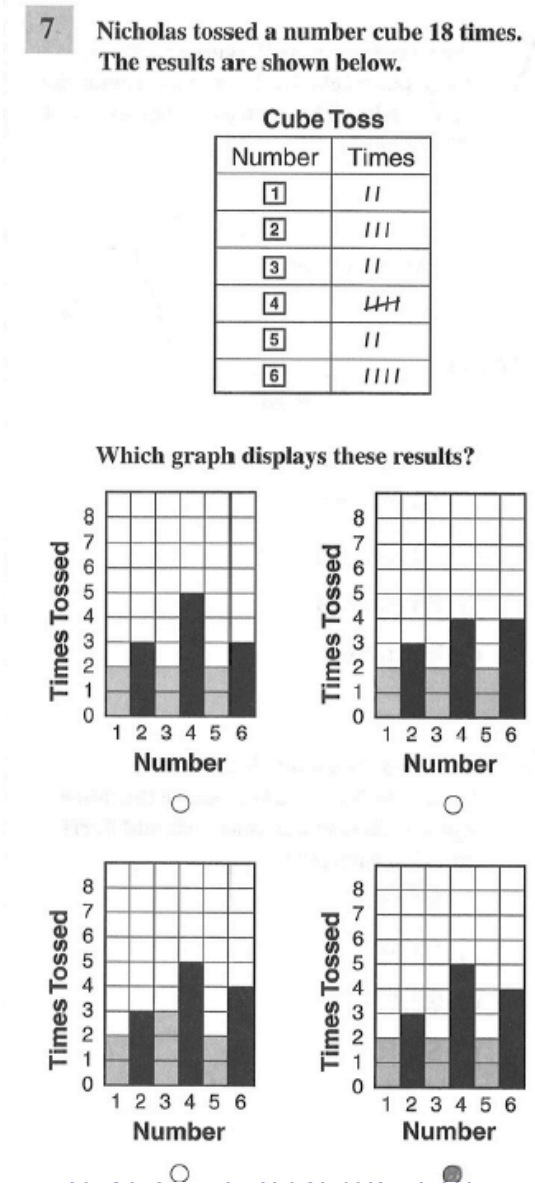
V3G: The convincing letter challenge

Dear Disney, It disturbs me greatly that in every movie you make with a dragon, the dragon gets killed by a knight. Please, if you could change that, it would be a great happiness to me. The Dragon is my school mascot. The dragon isn't really bad, he/she is just made bad by the villan. The dragon is not the one who should be killed. For example, Sleeping Beauty, the dragon is under the villaness's power, so it is not neccisariliy bad or evil. Please change that.

Your sad and disturbed writer, Allegra

- Points for taking a clear position and explaining it
- Points for evidence that the other party holds a different position
- Points for examples to help illustrate the case
- Points for rhetorical skill
- …

V3G: The change of representation challenge



7 Nicholas tossed a number cube 18 times. The results are shown below.

Cube Toss

Number	Times
1	II
2	III
3	II
4	~~IIII~~
5	II
6	IIII

Which graph displays these results?

- Points for representing the same situation in two ways
- Points for identifying correspondences between components of the situations
- …

Challenge 4: Learn to read, read to learn

(thanks to Murray Burke and Tim Oates)

Most of the world's knowledge is represented in text

"Learn to read, read to learn"

By 2020 read and comprehend any book up to third-grade level

Demonstrate that comprehension depends on previously-read texts

Challenge 3: Learn to read, read to learn
Good attributes

- Comprehension tests are easy to construct
- "Never-ending," monotonic, not "throwaway" from one year to the next
- Graduated series of challenges

  Few books, limited vocabulary, narrow subject matter, at first
- Like Robocup, technical committees for KR, ontologies, parsing, semantic representations, book selection, international, metrics and evaluation…
- Even a "kid's book Turing test" based on answers to comprehension questions

- Some serious scientific hypotheses

Learn to read, read to learn
Scientific claims

- You don't learn to read and *then* read to learn; they bootstrap each other:
- Hypothesis 1 (sufficiency of core semantics): We have done enough work in linguistics and ontological engineering to represent the meanings of a *useful subset* of possible sentences.
- Hypothesis 2 (bootstrapping): learning by reading provides sufficient conditions for the machine to extend its core semantics and understand a wider range of linguistic input.
- Hypothesis 3: (nonlinear learning rate) The more the machine learns, the more it can learn. The machine gains access to the world's knowledge at an accelerating rate.

Challenge 5: Robot Baby

- Neonates *act* and *perceive* and little else
- The rest – representation, concepts, planning, language – must arise from action and perception, or from the *dynamics of perception during action*
- What is the minimum innate endowment?
- Model cognitive development on a mobile robot

- Robots, because of the hypothesis that the conceptual system is grounded in physical interaction.

Challenge 5: Robot Baby
Good attributes

- Hypothesis 2 (bootstrapping, revisited): learning by ~~reading~~ physical interaction with the environment provides sufficient conditions for the machine to extend its core semantics and understand a wider range of linguistic input.
- Integrates three major areas of AI:

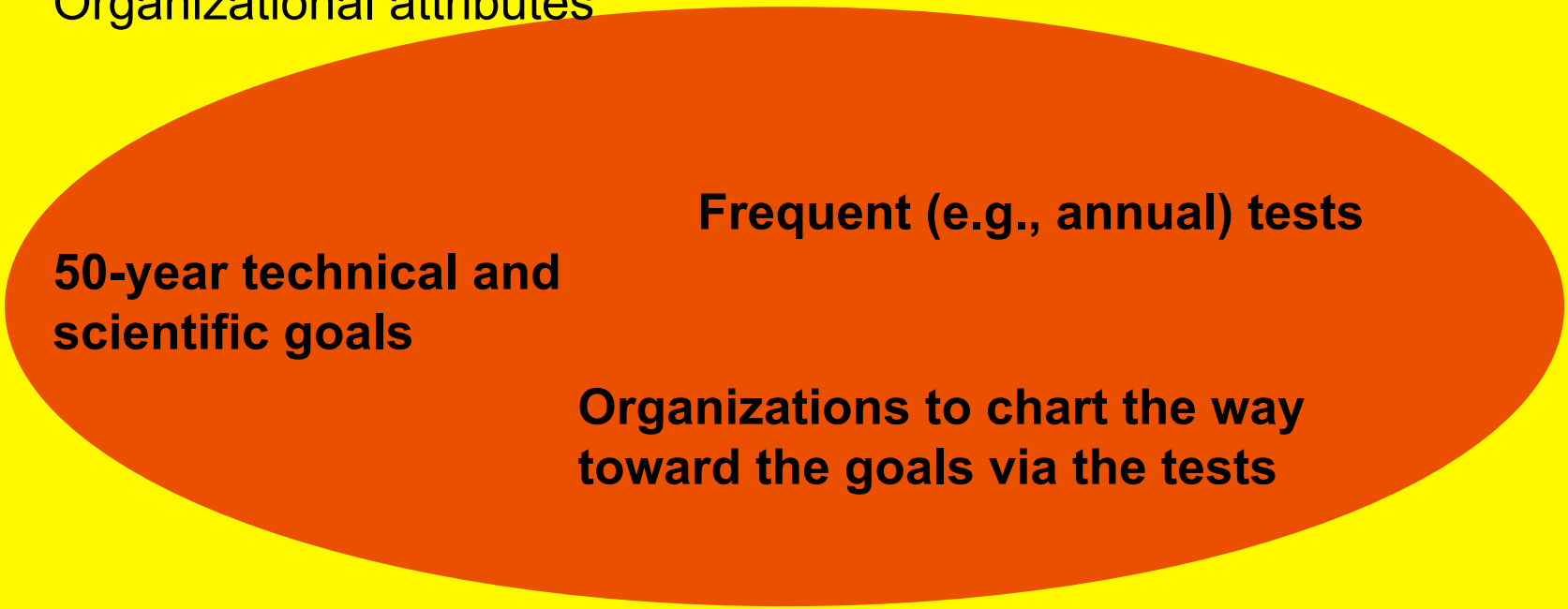
Sensing, perception and action

Concepts, ontologies, representation, knowledge

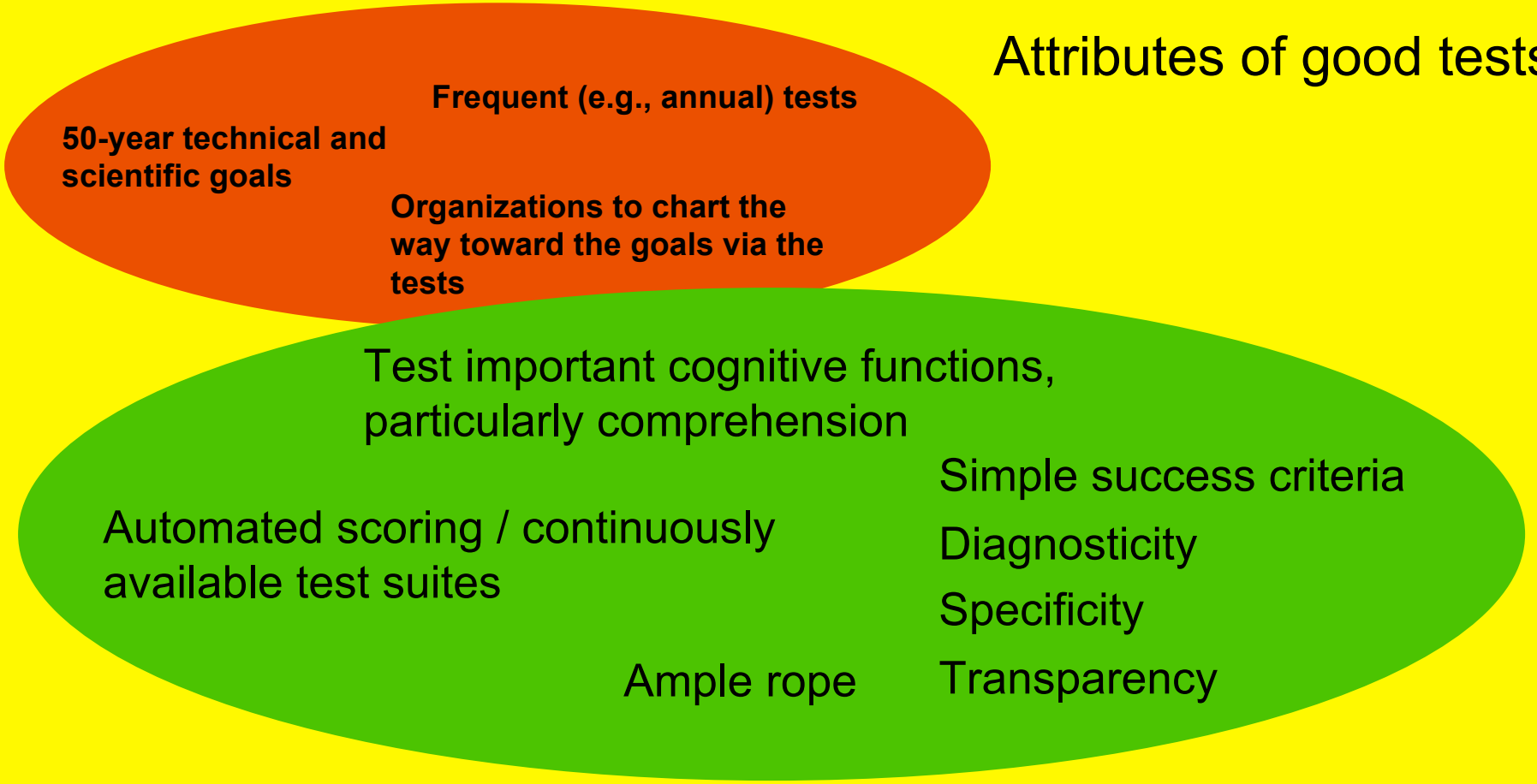
Learning

Attributes of good tests

Organizational attributes

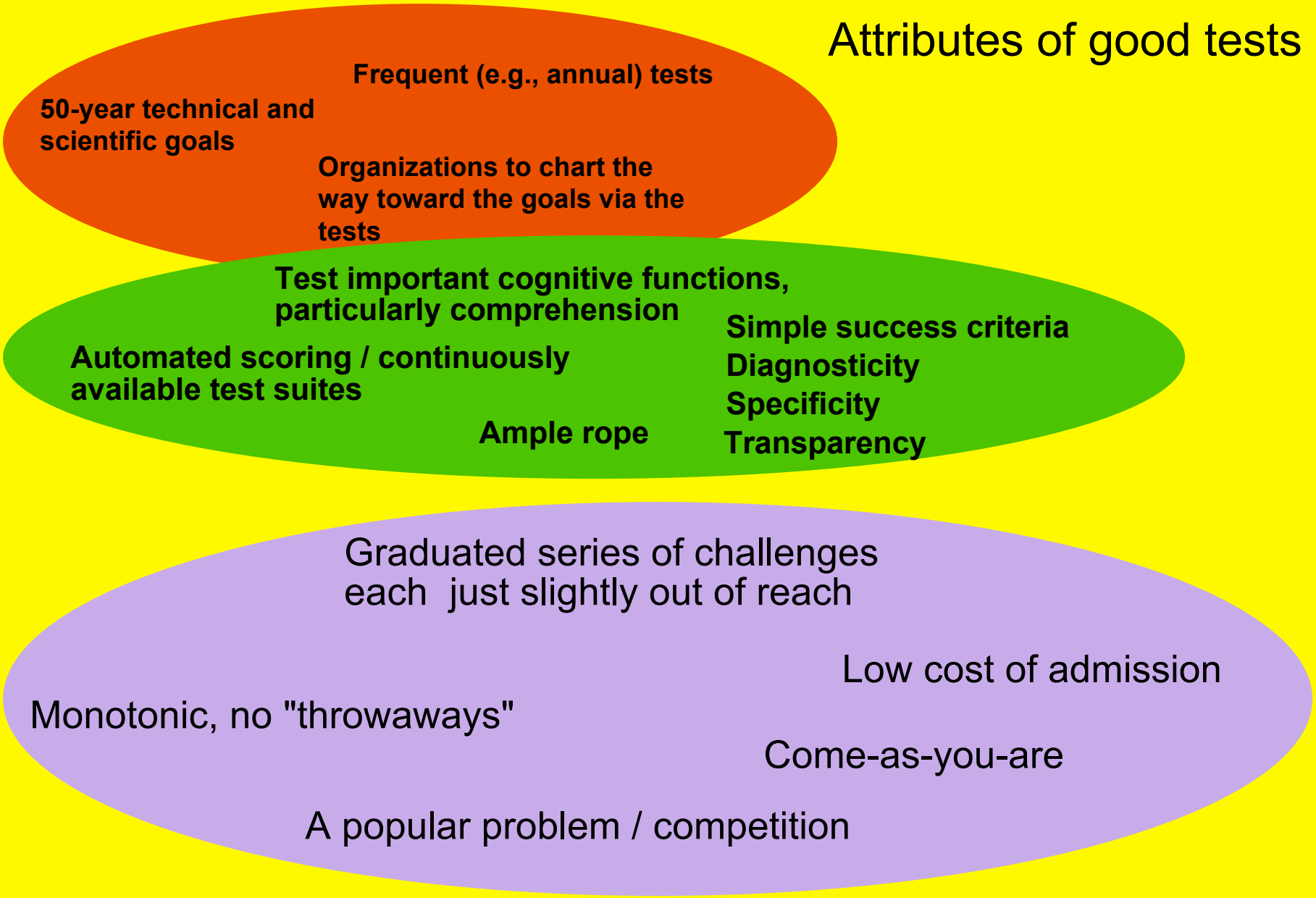


Attributes of good tests



Attributes of the tests themselves

Attributes of good tests



"Hearts and minds" attributes

Divide and conquer?

Some Challenges and Grand Challenges for Computational Intelligence

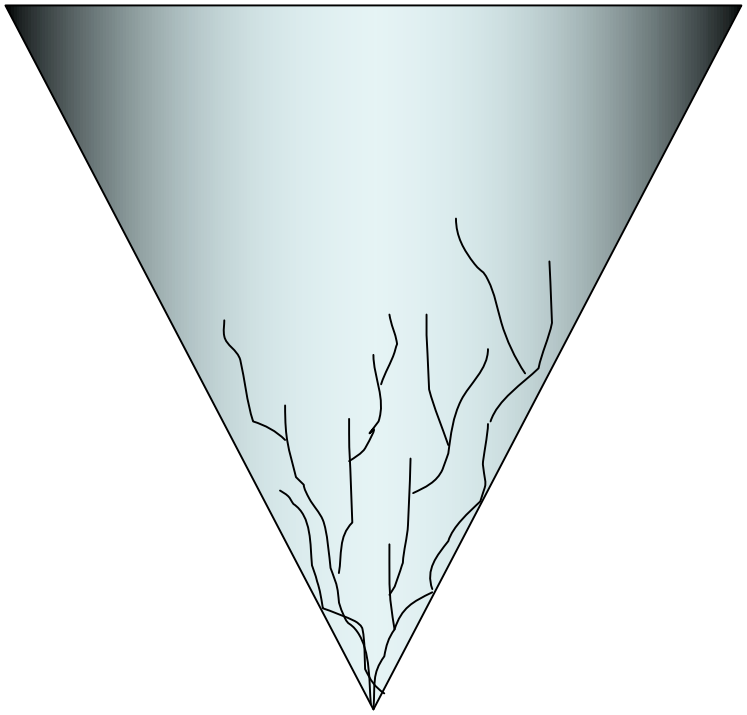
Edward A. Feigenbaum JACM Paper

"An appropriate strategy…is divide and conquer – study the dimensions of intelligence more or less separately. We must be satisfied for a long time to come with 'partial intelligence' in our artifacts as a natural consequence of this inevitable strategy."

Feigenbaum Test

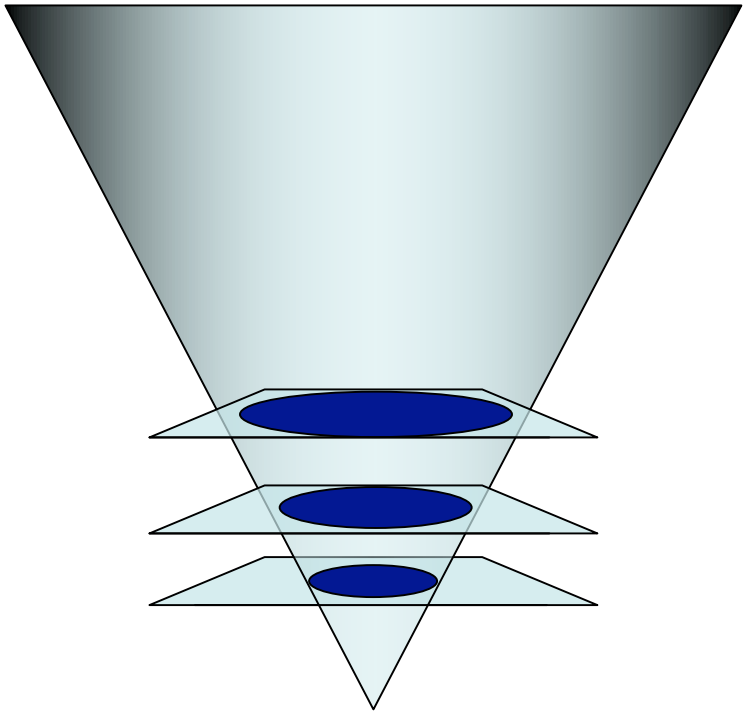
"In each round of the game the behavior of the two players, [National Academy member] and computer is judged by another Academy member in that particular domain of discourse…The judge poses problems, asks questions, asks for explanations, theories, and so on – as one might do with a colleague. Can the human judge choose, at better than chance level, which is his National Academy colleague and which is the computer?"

"Divide and conquer" or "Walk before you run"



Break the problem into pieces
and "go deep" on each

"Divide and conquer" or "Walk before you run"

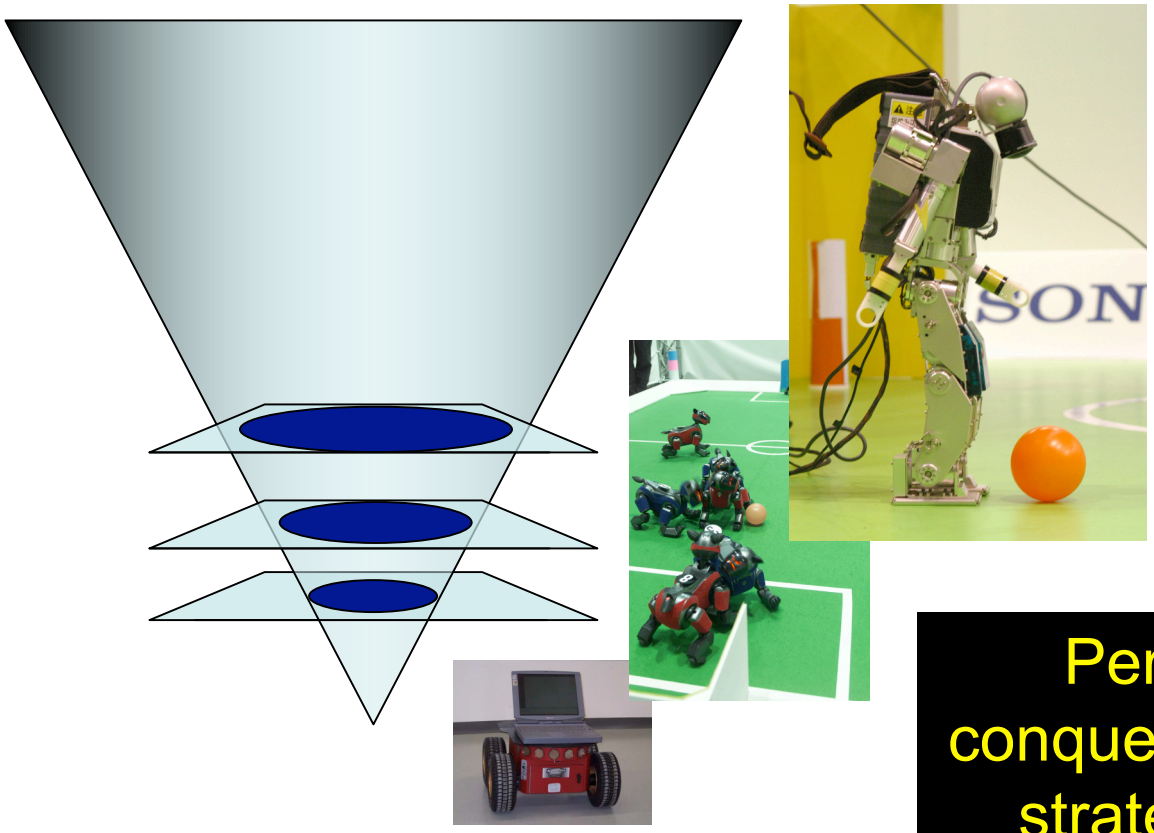


A *developmental* strategy is to build more-or-less complete agents, each with several cognitive functions

At first they aren't very capable

By setting new problems we gradually make them more capable

The developmental strategy



Build more-or-less complete agents, integrating several cognitive functions

Make the problems harder to make the agents more capable

Perhaps "divide and conquer" is *not* an inevitable strategy and we can do better than build partial intelligences

Attributes of good tests

Frequent (e.g., annual) tests

50-year technical and scientific goals

Organizations to chart the way toward the goals via the tests

Test important cognitive functions, particularly comprehension

Automated scoring / continuously available test suites

Simple success criteria
Diagnosticity
Specificity
Transparency

Ample rope

Graduated series of challenges each just slightly out of reach

Low cost of admission

Monotonic, no "throwaways"

Come-as-you-are

A popular problem / competition

Development, not divide-and-conquer

Integrate AI technologies into more-or-less complete agents

In any given challenge, accept poor performance but insist on universal coverage

If not the Turing Test, then what? Challenges…they're everywhere!

Robocup

The Planning Competition

General Game Playing Project

MUC, TREC, SenEval, DUC, DetectionACE, coreferenceMT, MTSensEval

Surprise Language Challenge

E-Auction Competition

Perhaps You Could Start One !!!

Thanks to:

Carole Beal, Yolanda Gil, Manuela Veloso
David Aha, Jim Blythe, Tom Dietterich,
Ed Feigenbaum, Jim Hendler, Lynette Hirschmann,
Jerry Hobbs, Ed Hovy, Adele Howe, Kevin Knight,
Daniel Marcu, Pat Langley, Natasha Noy,
Steve Smith, Milind Tambe, Mohammed Zaki

Thanks to:

Carole Beal, Yolanda Gil, Manuela Veloso
David Aha, Jim Blythe, Tom Dietterich,
Ed Feigenbaum, Jim Hendler, Lynette Hirschmann,
Jerry Hobbs, Ed Hovy, Adele Howe, Kevin Knight,
Daniel Marcu, Pat Langley, Natasha Noy,
Steve Smith, Milind Tambe, Mohammed Zaki

And to you, for listening!