

Adapting the Turing Test for Declaring Document Analysis Problems Solved

Daniel Lopresti

Computer Science & Engineering
Lehigh University
Bethlehem, PA, USA

George Nagy

Electrical, Computer, and
Systems Engineering
Rensselaer Polytechnic Institute
Troy, NY, USA

** Daniel Lopresti acknowledges support from a DARPA IPTO grant administered by Raytheon BBN Technologies.*



When is a Problem Solved?*

We define our open problems as automating a task: this is quite different from math, physics, theoretical CS, etc.

Some ways of measuring success:

- Relative accuracy of new algorithm vs. previous methods.
- Relative accuracy of algorithm vs. human "ground truth."
- Distinguishability of algorithm from human result.
- Current degree of community interest (publishability).
- Economic considerations (net payoff for using method).

* Building on our ICDAR 2011 paper: "When is a Problem Solved?," D. Lopresti and G. Nagy, *Proceedings of the Eleventh International Conference on Document Analysis and Recognition (ICDAR 2011)*, September 2011, Beijing, China, pp. 32-36.

Viewpoint #2*

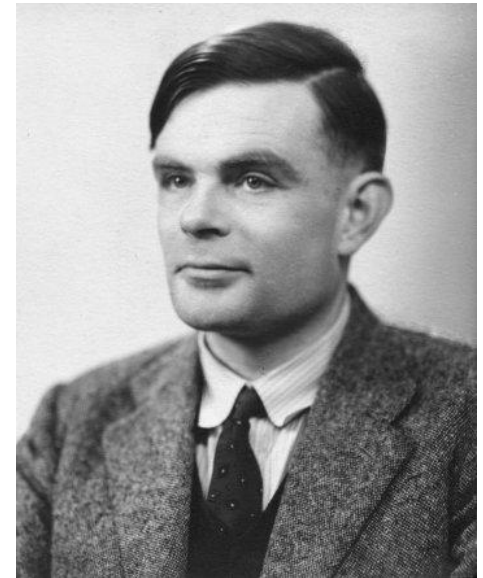
The Turing Test:

"A problem is solved if there is a method which has been widely publicized and documented and freely available to the community which generates output for a given input that a human judge cannot reliably distinguish from the output of a human expert."

* Building on our ICDAR 2011 paper: "When is a Problem Solved?," D. Lopresti and G. Nagy, *Proceedings of the Eleventh International Conference on Document Analysis and Recognition (ICDAR 2011)*, September 2011, Beijing, China, pp. 32-36.

Alan Turing

Alan Turing, (23 June 1912 - 7 June 1954), was an English mathematician, logician, cryptanalyst, and computer scientist. He was highly influential in the development of computer science, providing a formalisation of the concepts of "algorithm" and "computation" with the Turing machine, which played a significant role in the creation of the modern computer. Turing is widely considered to be the father of computer science and artificial intelligence.



* From http://en.wikipedia.org/wiki/Alan_Turing

The Turing Test

VOL. LIX. No. 236.]

[October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

By A. M. TURING

1. *The Imitation Game.*

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. He knows them by labels X and Y, and at the end of the game he says either 'X is A and Y is B' or 'X is B and Y is A'. The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair?
Now suppose X is actually A, then A must answer. It is A's

28

433

1. *The Imitation Game.*

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

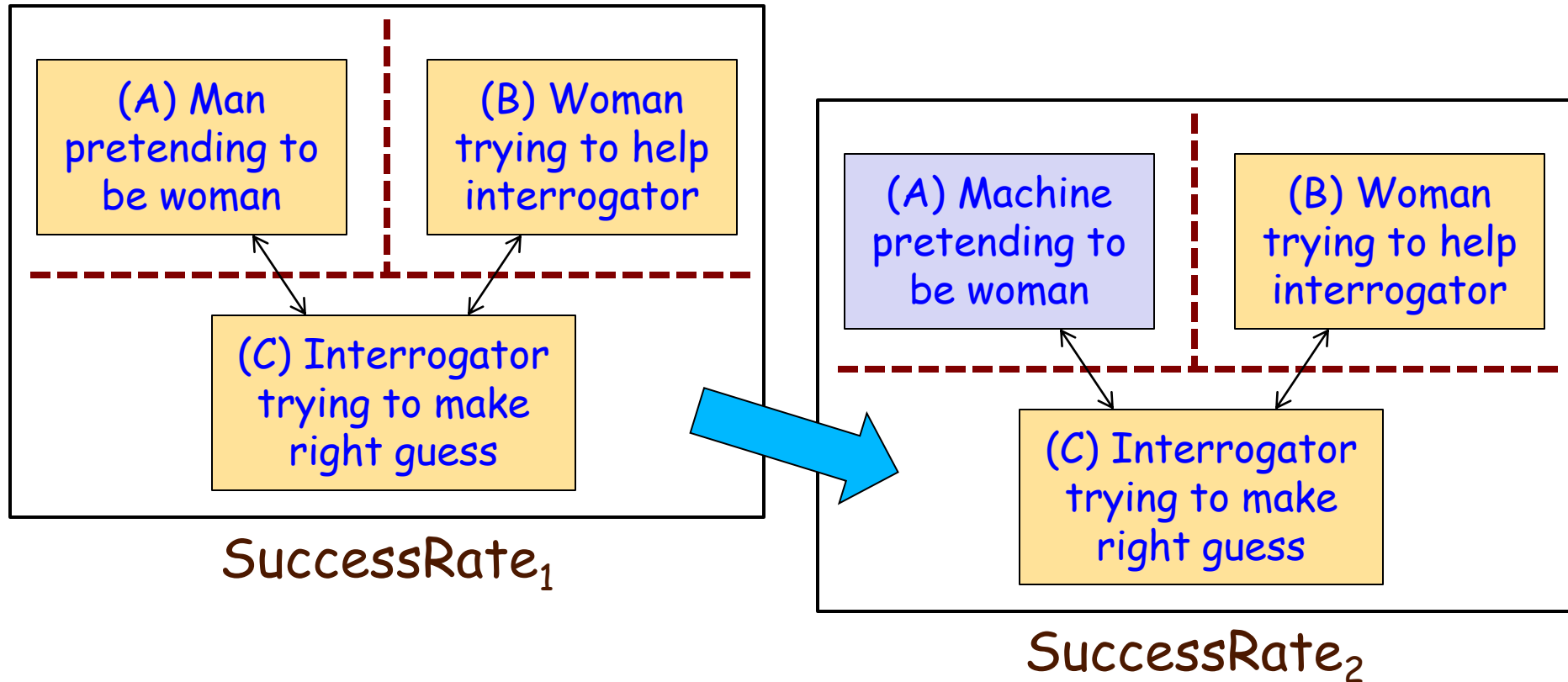
The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator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

We now ask the question, 'What will happen when a machine takes the part of A 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?'

A. M. Turing, "Computing Machinery and Intelligence," *Mind*, vol. 59, no. 236, October 1950, pp. 433-460.



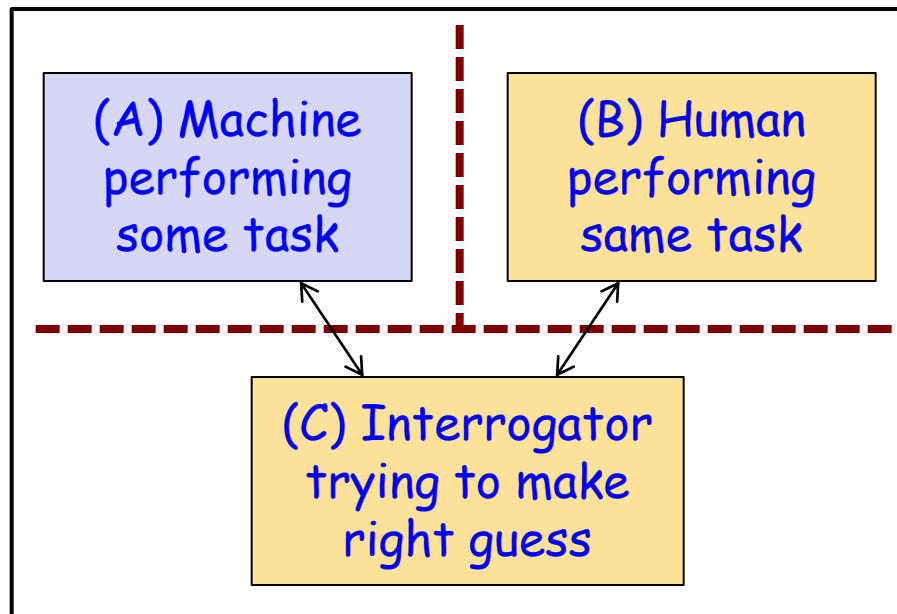
The Turing Test



Is $\text{SuccessRate}_2 \approx \text{SuccessRate}_1$?

The Turing Test

The Turing Test is an elegantly simple idea, so it should be simple to implement, right?



Is SuccessRate no better than random chance ?

- Note this differs from Turing's original formulation.
- When considering a real implementation, other, more serious complications arise.

Long Bet*

the rules OF LONG BETS bets & predictions ON THE RECORD make a PREDICTION about LONG BETS FAQ & ANSWERS

THE ARENA FOR ACCOUNTABLE PREDICTIONS

A LONG BET

BET 1 DURATION 27 years (02002-02029)


"By 2029 no computer - or "machine intelligence" - will have passed the Turing Test." [DETAILED TERMS »](#)

PREDICTOR
Mitchell Kapor

CHALLENGER
Ray Kurzweil

STAKES \$20,000
will go to *The Electronic Frontier Foundation* if Kapor wins,
or *The Kurzweil Foundation* if Kurzweil wins.

Voting has been temporarily disabled.

DISCUSS & SHARE
Add your voice to a conversation with the bettors: [Join the discussion »](#)
Bookmark this bet, and share it with friends: [ADD THIS](#) 

Kapor's Argument
The essence of the Turing Test revolves around whether a computer can successfully impersonate a human. The test is to be put into practice under a set of detailed

Kurzweil's Argument
The Significance of the Turing Test. The implicit, and in my view brilliant, insight in Turing's eponymous test is the ability of written human language to represent

"By 2029 no computer - or 'machine intelligence' - will have passed the Turing Test."

PREDICTOR:
Mitchell Kapor
CHALLENGER:
Ray Kurzweil
STAKES: \$20,000

* M. Kapor and R. Kurzweil, "A Long Bet: By 2029 no computer - or 'machine intelligence' - will have passed the Turing Test," <http://longbets.org/1/>.

Long Bet Rules

Turing was nonspecific about how to administer his Test, but concreteness is needed when \$20,000 is at stake.

- Each of three Turing Test judges is to conduct an online interview ("chat") with each of four human players as well as the machine for two hours.
- At the end of these interviews, the judges indicate whether or not each candidate is human and also rank them from "least human" to "most human."
- The machine is said to pass the Turing Test if it fools two or more judges and if its median rank is equal to or greater than at least two of the human players.

Adapting the Turing Test

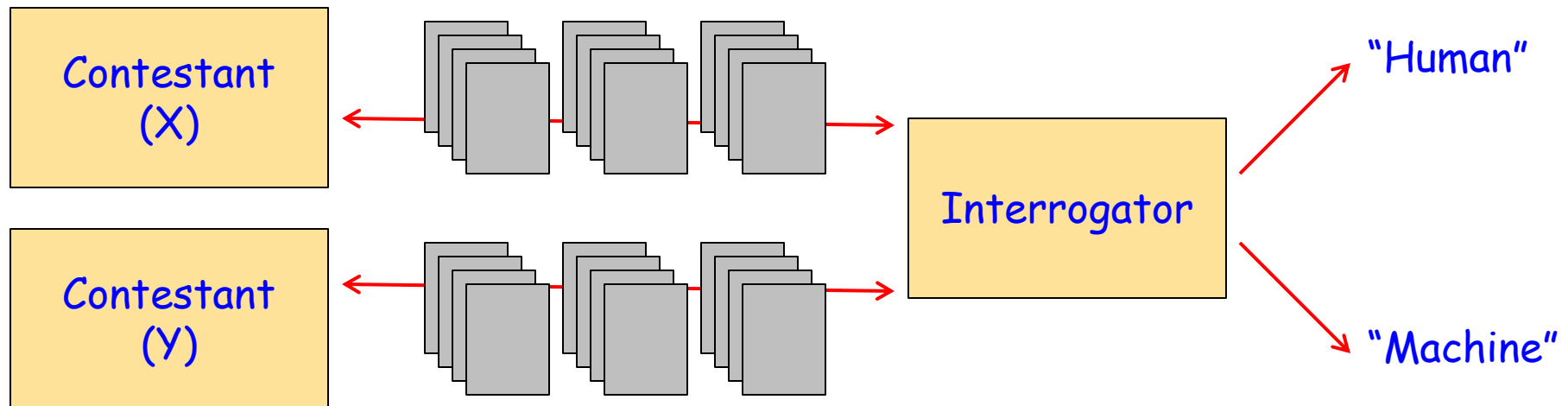
The Long Bet is a one-time event with a significant amount of prize money involved. As a result, it makes sense to employ a heavy-weight protocol for the test.

How can the Turing Test be applied in document analysis?

- What are the essential qualities to preserve?
- What can be dispensed with, or at least simplified?
- When implemented, how would the test “look”?
- When might such a test be appropriate?

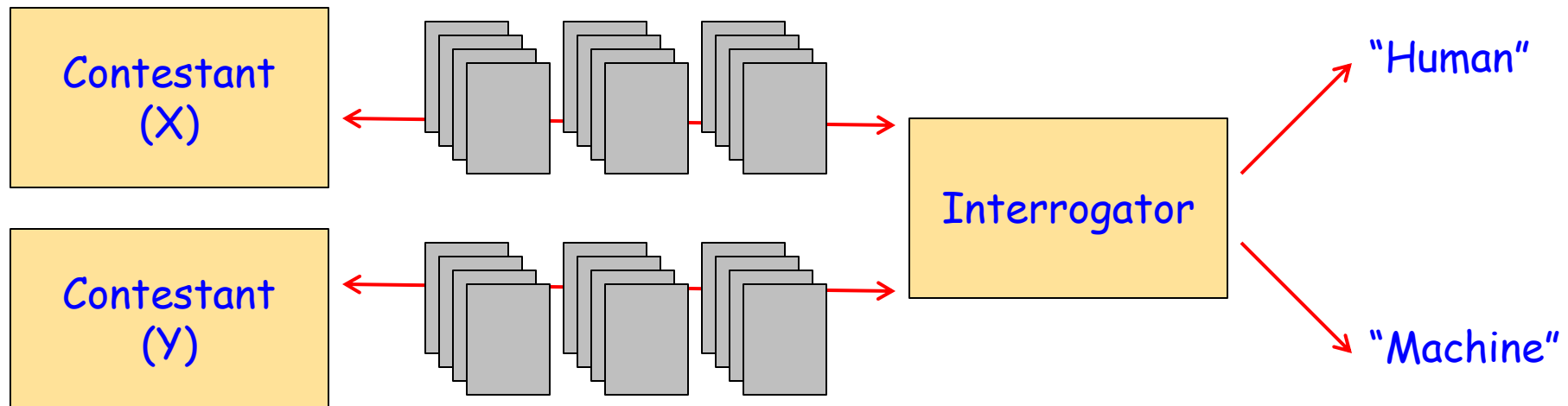
Properties to Preserve #1

Human judgment is applied to determine a simple machine/human distinction and nothing more complex than this. Automated evaluation (i.e., a computation to determine how "similar" a machine output is to some predefined human "ground truth") is ruled out.



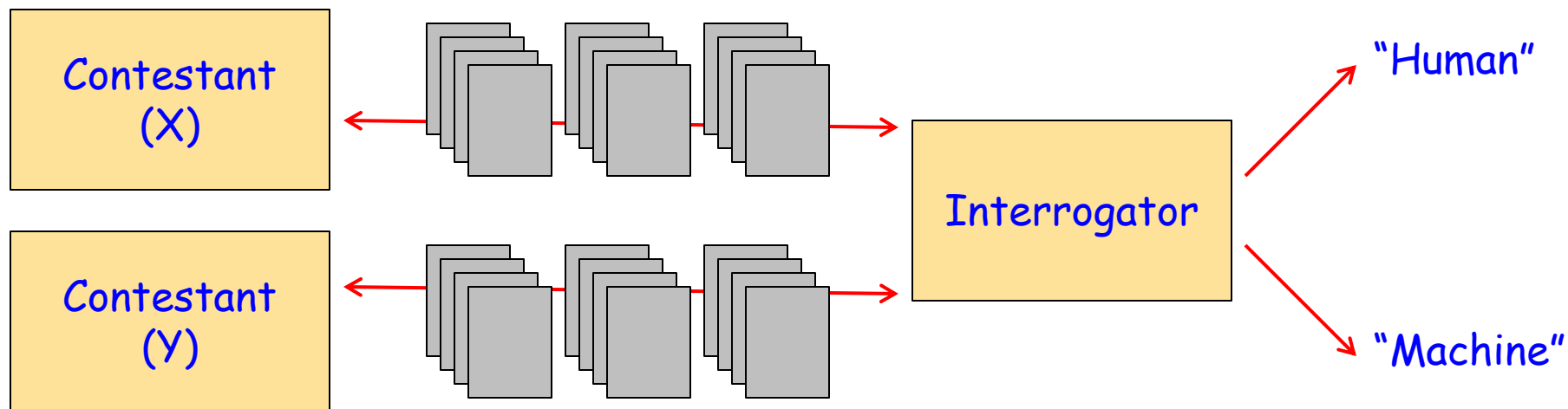
Properties to Preserve #2

A judge may ask any number of questions before making a determination. A "question" here is a challenge that requires a response from the player. For document analysis applications, this will normally consist of a page image to be processed in some way.



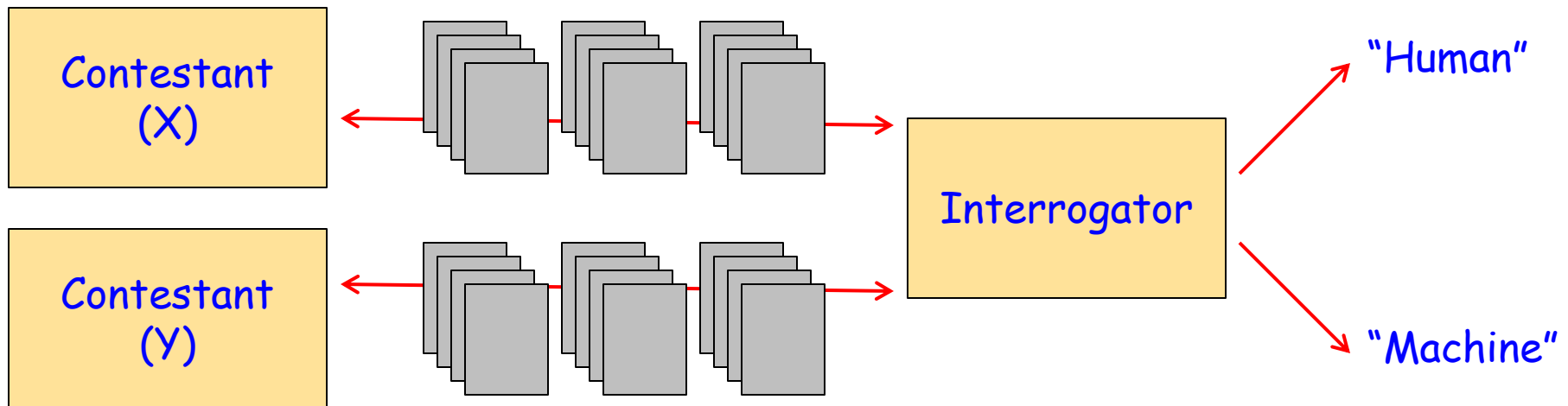
Properties to Preserve #3

The judge decides which questions to use, and is free to conduct the questioning of the players without constraint on the choice, sequence, and number of questions.



Properties to Preserve #4

A series of such evaluations, with anyone being allowed to volunteer to serve as judge or as the human player, is conducted before declaring a problem "solved" (if/when the success rates of the best-performing judges are statistically no better than random).



Properties to Adapt

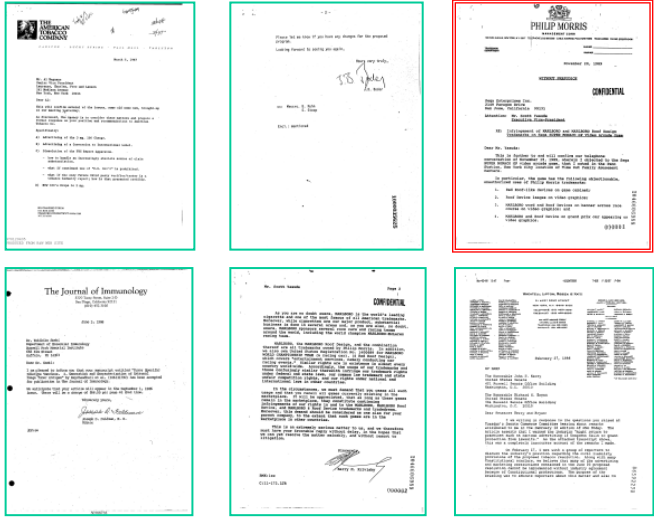
Some aspects of Turing's original Test must be updated:

- The judge and players do not interact via a natural language question-and-answer process. Instead, they employ a graphical user interface which supports the upload of image files and visual inspection of results.
- The domain of discourse is no longer open-ended. Note that this replaces Turing's original question "Can machines think?" with our "Is this problem solved?"

GUI from Judge's Perspective

Task is: Logo Detection **Current Challenge is #12**

Pre-defined Challenge Library

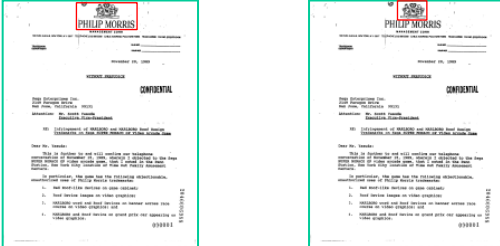


Create New Challenge

File name Upload

Submit to Player A Submit to Player B

Responses



Determination: A human, B machine A machine, B human

New Properties

The manner in which we implement the test makes use of global interconnectedness that Turing could not envision.

- A modified Turing Test could run on an open platform such as the DAE server.*
- Anyone - in particular, members of the research community - would be permitted to volunteer at any point in time to serve as the judge or the human player to test a preregistered algorithm on some specific task.
- The need to pair a judge with a human player can be addressed through crowdsourcing (e.g., using micro-payments to recruit subjects like Mechanical Turk).

* Later in this session: "The Non-geek's guide to the DAE Platform," B. Lamiroy and D. Lopresti.

Other Considerations

How can we eliminate out-of-scope querying by a judge?

How can we prevent the human player from signaling to the judge in a way that is impossible for a program?

- By openly publishing traces of all tests conducted on an algorithm, other researchers can be encouraged to follow along and render their own opinions.
- In this way, the behaviors of judges and players will themselves be subject to scrutiny.
- Ultimately, the community will determine which tests were conducted fairly and hence are used in computing the statistics that answer the question at hand.

Suitability of the Approach

What about tasks that are natural for machines but very tedious for humans?

- Clearly it makes no sense to ask human players to try to perform the same search functions over billions of documents that google does so well (Turing pondered similar issues regarding the machine being "too good").
- We could "dumb down" the algorithm drastically by, say, running it on very slow hardware, but this is pointless.
- This suggests only that some tasks are not suited for evaluating this way, not that the basic idea of a Turing-like Test is flawed.

Role of Learning

Turing ends his paper with a prediction that within about 50 years, machines will be able to pass his Test and that some notion of machine learning will play a key role.

- In the Test itself, Turing did not envision one player seeing the interactions with the other. Learning might be possible based on what a single player sees. But it is more interesting if players can observe each other.
- In some cases, the human may learn from the machine!
- Learning/adaptation distinguishes humans from machines. While the machine may consistently lose at first, we would be pleased to see it one day catch up.

Conclusions

We suggest that a Turing-like Test may be the right mechanism for declaring a DIA problem has been solved:

- Note that community opinions play a major role.
- Many technical details remain to be worked out.
- Seems likely that an open platform such as the DAE server may prove to be a key component.
- This idea is intended to be provocative - please venture forth with your own thoughts and suggestions!