DRAFT of June 19, 2015 at 12:43 Noon.
Do not cite, quote, or redistribute.

PRINCIPLES FOR DESIGNING AN AI COMPETITION, OR WHY THE TURING TEST FAILS AS AN INDUCEMENT PRIZE

STUART M. SHIEBER

Introduction

There has been a spate recently of calls for replacements for the Turing Test. Gary Marcus in *The New Yorker* asks "What Comes After the Turing Test?" and wants "to update a sixty-four-year-old test for the modern era" (Marcus, 2014). Moshe Vardi in his *Communications of the ACM* article "Would Turing Have Passed the Turing Test?" opines that "It's time to consider the Imitation Game as just a game" (Vardi, 2014). The popular media recommends that we "Forget the Turing Test" and replace it with a "better way to measure intelligence" (Locke, 2014). Behind the chorus of requests is an understanding that the Test has served the field of artificial intelligence poorly as a challenge problem to guide research.

This shouldn't be surprising: The Test wasn't proposed by Turing to serve that purpose. Turing's (1950) *Mind* paper in which he defines what we now call the Turing Test concludes with a short discussion of research strategy towards machine intelligence. What he says is this:

> We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision. . . . I do not know what the right answer is, but I think [various] approaches should be tried. (Turing, 1950, page 460)

What he does not say is that we should be running Turing Tests.

Perhaps Turing saw that the Test is not at all suitable for this purpose, as I will argue in more detail here. But that didn't stop some with an entrepreneurial spirit in staging Turing-Test–inspired competitions. Several, including myself (Shieber, 1994) and Hayes and Ford (1995), argued such stunts to be misguided and inappropriate. The problem with misapplication of the Turing Test in this way has been exacerbated by the publicity around a purported case of a chatbot in June 2014 becoming "the first machine to pass the Turing test" (Anonymous, 2014), when of course no such feat took place (Shieber, 2014a). (It is no coincidence that all of the articles cited in the first paragraph came out in June 2014.)

Date: June 19, 2015 at 12:43 Noon.

It is, frankly, sad to see the Turing Test besmirched by its inappropriate application as a challenge problem for AI. But at least this set of events has had the salutary effect of focusing the AI research community on the understanding that if the Turing Test isn't a good challenge problem for guiding research toward new breakthroughs, we should attend to devising more appropriate problems to serve that role. These calls to replace the pastime of Turing-Test-like competitions are really pleas for a new "inducement prize contest".

Inducement prize contests are award programs established to induce people to solve a problem of importance by directly rewarding the solver, and the idea has a long history in other research fields – navigation, aviation, and autonomous vehicles, for instance. If we are to establish an inducement prize contest for artificial intelligence, it behooves us to learn from the experience of the previous centuries of such contests to design our contest in a way that is likely to have the intended effect. In this paper, I adduce five principles that an inducement prize contest for AI should possess: occasionality of occurrence, flexibility of award, transparency of result, absoluteness of criteria, and reasonableness of goal. Any proposal for an alternative competition, moving "beyond the Turing Test" in the language of the January 2015 Association for the Advancement of Artificial Intelligence workshop,[1] ought to be evaluated according to these principles.

The Turing Test itself fails the reasonableness principle, and its implementations to date in various competitions have failed the absoluteness, occasionality, flexibility, and transparency principles, a clean sweep of inappropriateness for an AI inducement prize contest. Creative thinking will be needed to generate a contest design satisfying these principles.

Inducement prize contests

There is a long history of inducement prizes in a broad range of areas, including: navigation (the 1714 Longitude Prize), chemistry (the 1783 French Academy of Sciences prize for soda ash production), automotive transportation (the 1895 Great Chicago Auto Race), aviation (numerous early 20th century prizes culminating in the 1919 Orteig Prize for nonstop trans-Atlantic flight; the 1959 Kremer Prize for human-powered flight), space exploration (the 1996 Ansari X Prize for reusable manned spacecraft), and autonomous vehicles (the 2004 DARPA Grand Challenge). Inducement prizes are typically offered on the not unreasonable assumption that they provide a highly financially leveraged method for achieving progress in the award area. Estimates of the leverage have ranged up to a factor of 50 (Schroeder, 2004).

[1] http://www.aaai.org/Workshops/ws15workshops.php#ws06

There have been two types of competitions related to AI,[2] though neither type serves well as an inducement prize contest:

(1) There have been regularly scheduled enactments of (or at least inspired by) the Turing Test. The most well known is the Loebner Prize Competition, held annually, though other similar competitions have been held, such as the June 2014 Royal-Society-sponsored competition in London, whose organizers erroneously claimed that entrant "Eugene Goostman" had passed the Turing Test (Shieber, 2014a). Although Hugh Loebner billed his eponymous prize as a curative for the astonishing claim that "People had been discussing AI, but nobody was doing anything about it" (Lindquist, 1991), his competition is not set up to provide appropriate incentives and has not engendered any progress in the area so far as I can tell (Shieber, 1994).

(2) Research funders, especially US government funders like DARPA, NSF, and NIST, have funded regular (typically annual) "bakeoffs" among funded research groups working on particular applications – speech recognition, message understanding, question-answering, and so forth. These competitions have been spectacularly successful at generating consistent incremental progress on the measured objectives, speech recognition error rate reduction, for instance. Such competitions are evidently effective at generating improvements on concrete engineering tasks over time. They have, however, had the perverse effect of reducing the diversity of approaches pursued and generally increasing risk aversion among research projects.

Principles

An inducement prize contest for AI has the potential to promote research on hard AI problems without the frailties of these previous competitions. We, the AI community, would like a competition to promote creativity, reward risk, and curtail incrementalism. This requires careful attention to the principles underlying the competition, and it behooves us to attend to history. We should look to previous successful inducement prize contests in other research fields in choosing a task and competition structure that obeys the principles that made those competitions successful. These principles include the following:

- The competition should be *occasional*, occurring only when plausible entrants exist.

[2] The XPrize Foundation, in cooperation with TED, announced on March 20, 2014 the intention to establish the "AI XPrize presented by TED", described as "a modern-day Turing test to be awarded to the first A.I. to walk or roll out on stage and present a TED Talk so compelling that it commands a standing ovation from you, the audience." (XPRIZE Foundation, 2014) The competition has yet to be finalized however.

- The awarding process should be *flexible*, so awards follow the spirit of the competition rather than the letter of the rules.
- The results should be *transparent*, so that any award is given only for systems that are open and replicable in all aspects.
- The criteria for success should be based on *absolute* milestones, not relative progress.
- The milestones should be *reasonable*, that is, not so far beyond current capability that their achievement is inconceivable in any reasonable time.

The first three of these principles concern the *rules* of the contest, while the final two concern the *task* being posed. I discuss them seriatim, dispensing quickly with the rule-oriented principles to concentrate on the more substantive and crucial task-related ones.

Occasionality.

> *The competition should be* occasional, *occurring only when plausible entrants exist.*

The frequency of testing entrants should be determined by the availability of plausible entrants, not by an artificially mandated schedule. Once one stipulates that a competition must be run, say, every year, one is stuck with the prospect of awarding a winner whether any qualitative progress has been made or not, essentially forcing a quantitative incremental notion of progress that leads to the problems of incrementalism noted above.

Successful inducement prize contests are structured so that actual tests of entrants occur only when an entrant has demonstrated a plausible chance of accomplishing the qualitative criterion. The current Kremer Prize (the 1988 Kremer International Marathon Competition) stipulates that it is run only when an entrant officially applies to make an attempt under observation by the committee. Even then, any successful attempt must be ratified by the committee based on extensive documentation provided by the entrant. Presumably to eliminate frivolous entries, entrants are subject to a nominal fee of £100, as well as the costs to the committee of observing the attempt (The Royal Aeronautical Society, 1988).

This principle is closely connected to the task-related principle of absoluteness, discussed further below.

Flexibility.

> *The awarding process should be* flexible, *so awards follow the spirit of the competition rather than the letter of the rules.*

The goal of an inducement prize contest is to generate real qualitative progress. Any statement of evaluative criteria is a means to that end, not the end in itself. It is therefore useful to include in the process flexibility in the criteria, to make sure that the spirit, and not the letter, of the law are followed. For instance, the DARPA Grand Challenge allowed for disqualifying entries "that cannot demonstrate intelligent autonomous behavior" (Schroeder, 2004, page 14). Such flexibility in determining when evaluation of an entrant is appropriate and successful allows useful wiggle room to drop frivolous attempts or gaming of the rules. For this reason, the 1714 Longitude Prize placed awarding of the prize in the hands of an illustrious committee chaired by Isaac Newton, Lucasian Professor of Mathematics. Similarly, the Kremer Prize places "interpretation of these Regulations and Conditions . . . with the Society's Council on the recommendation of the Organisers." (The Royal Aeronautical Society, 1988, page 6)

Transparency.

> *The results should be* transparent*, so that any award is given only for systems that are open and replicable in all aspects.*

The goal of establishing an inducement prize in AI is to expand knowledge for the public good. We therefore ought to require entrants (not to mention awardees) to make available sufficient information to allow replication of their awarded event: open-source code and any required data, open access to all documentation. It may even be useful for any award to await an independent party replicating and verifying the award. There should be no award for secret knowledge.

The downside of requiring openness is that potential participants may worry that their participation could poison the market for their technological breakthroughs, and therefore they would avoid participation. But to the extent that a potential participant believes that there is a large market for their satisfying the award criteria, there is no reason to motivate them with the award in the first place.

Absoluteness.

> *The criteria for success should be based on* absolute *milestones, not relative progress.*

Any competition should be based on absolute rather than relative criteria. The criterion for awarding the prize should be the satisfaction of specific milestones rather than mere improvement on some figure of merit. For example, the 1714 Longitude Act established three separate awards based on specific milestones:

> "That the first author or authors, discover or discoverers of any such method, his or their executors, administrators, or assigns, shall be entitled to, and have such reward as herein after is mentioned; that is to say, to a reward or sum of ten thousand pounds, if it determines the said longitude to one degree of a great circle, or sixty geographical miles; to fifteen thousand pounds if it determines the same to two thirds of that distance; and to twenty thousand pounds,

> if it determines the same to one half of that same distance." (British Parliament, 1714)

Aviation and aeronautical prizes specify milestones as well. The Orteig prize, first offered in 1919, specified a transatlantic crossing in a single airplane flight, achieved by Charles Lindbergh in 1927. The Ansari X Prize required a nongovernmental organization to perform two launches to 100 kilometers within two weeks of a reusable manned spacecraft, a requirement fulfilled by Burt Rutan's SpaceShipOne eight years after the prize's 1996 creation.

If a winner is awarded merely on the basis of having the best current performance on some quantitative metric, entrants will be motivated to incrementally outperform the previous best, leading to "hill climbing". This is exactly the behavior we see in funder bakeoffs. If the prevailing approach sits in some mode of the research search space with a local optimum, a strategy of trying qualitatively different approaches to find a region with a markedly better local optimum is unlikely to be rewarded with success the following year. Prospective entrants are thus given incentive to work on incremental quantitative progress, leading to reduced creativity and low risk. We see this phenomenon as well in the Loebner Competition; some two decades of events have used exactly the same techniques, essentially those of Weizenbaum's (1966) Eliza program. If, by contrast, a winner is awarded only upon hitting a milestone defined by a sufficiently large quantum of improvement, one that the organizers believe requires a qualitatively different approach to the problem, local optimization ceases to be a winning strategy, and examination of new approaches becomes more likely to be rewarded.

Reasonableness.

> *The milestones should be* reasonable*, that is, not so far beyond current capability that their achievement is inconceivable in any reasonable time.*

Although an absolute criterion requiring qualitative advancement provides incentive away from incrementalism, it runs the risk of driving off participation if the criterion is too difficult. We see this in the qualitative part of the Loebner Prize Competition. The competition rules specify that (in addition to awarding the annual prize to whichever computer entrant performs best on the quantitative score) a gold medal would be awarded and the competition discontinued if an entrant passes a multimodal extension of the Turing Test. The task is so far beyond current technology that it is safe to say that this prize has incentivized no one.

Instead, the award criterion should be beyond the state of the art, but not so far that its achievement is inconceivable in any reasonable time. Here again, successful inducement prizes are revealing. The first Kremer prize specified a human-powered flight over a figure eight course of half a mile. It did not specify a transatlantic flight, as the Orteig Prize for powered flight did. Such a milestone would have been unreasonable. Frankly, it is the difficulty of designing a criterion that walks the fine line between a qualitative improvement unamenable to hill climbing and a reasonable goal in the foreseeable future that makes designing an inducement prize contest so tricky. Yet without finding a Goldilocks-satisfying test (not too hard, not too easy, but just right), it is not worth running a competition. The notion of reasonableness is well captured by the X Prize Foundation's target of "audacious but achievable" (Anonymous, 2015).

The reasonableness requirement leads to a further consideration in choosing tasks where performance is measured on a quantitative scale. The task must have *headroom*. Consider again human-powered flight, measured against a metric of staying aloft over a prescribed course for a given distance. Before the invention of the airplane, human-powered flight distances would have been measured in feet, using technologies like jumping, poles, or springs. True human-powered flight – at the level of flying animals like birds and bats – is measured in distances that are, for all practical purposes, unlimited when compared to that human performance. The task of human-powered flight thus has plenty of headroom. We can set a milestone of 50 feet or half a mile, far less than the ultimate goal of full flight, and still expect to require qualitative progress on human-powered flight.

By comparison, consider the task of speech recognition as a test for intelligence. It has long been argued that speech recognition is an "AI-complete" task. Performance at human levels can require arbitrary knowledge and reasoning abilities. The apocryphal story about the sentence "It's hard to wreck a nice beach" makes an important point: The speech signal underdetermines the correct transcription. Arbitrary knowledge and reasoning – real intelligence – may be required in the most subtle cases. It might be argued, then, that we could use speech transcription error rate in an inducement prize contest to promote breakthroughs in AI. The problem is that the speech recognition task has very little headroom. Although human-level performance may require intelligence, near-human-level performance does not. The difference in error rate between human speech recognition and computer speech recognition may be only a few percentage points. Using error rate is thus a fragile compass for directing research.

Indeed, this requirement of reasonableness may be the hardest one to satisfy for challenges that incentivize research that leads to machine intelligence. Traditionally, incentive prize contests have aimed at breakthroughs in functionality, but intelligence short of human level is notoriously difficult to define in terms of functionality; it seems intrinsically intensional. Merely requiring a particular level of performance on a particular functionality falls afoul of what might be called "Montaigne's misconception". Michel de Montaigne in his arguing for the intelligence of animals notes the abilities of individual animals at various tasks:

> Take the swallows, when spring returns; we can see them ferreting through all the corners of our houses; from a thousand places they select one, finding it the most suitable place to make their nests: is that done without judgement or discernment? . . . Why does the spider make her web denser in one place and slacker in another, using this knot here and that knot there, if she cannot reflect, think, or reach conclusions?
>
> We are perfectly able to realize how superior they are to us in most of their works and how weak our artistic skills are when it comes to imitating them. Our works are coarser, and yet we are aware of the faculties we use to construct them: our souls use all their powers when doing so. Why do we not consider that the same applies to animals? Why do we attribute to some sort of slavish natural inclination works that surpass all that we can do by nature or by art? (de Montaigne, 1987 [1576], pages 19–20)

Of course, an isolated ability does not intelligence make. It is the generality of cognitive performance that we attribute intelligence to. Montaigne gives each type of animal credit for the cognitive performances of all others. Swallows build, but they do not weave. Spiders weave, but they do not play chess. People, our one uncontroversial standard of intelligent being, do all of these. Turing understood this point in devising his test. He remarked that the functionality on which his test is based, verbal behavior, is "suitable for introducing almost any one of the fields of human endeavour that we wish to include." (Turing, 1950, page 435)

Any task based on an individual functionality that does not allow extrapolation to a sufficiently broad range of additional functionalities is not adequate as a basis for an inducement prize contest for AI, however useful the functionality happens to be. (That is not to say that such a task might not be appropriate for an inducement prize contest for its own sake.) There is tremendous variety in the functionalities on which particular computer programs surpass people, many of which require and demonstrate intelligence in humans. Chess programs play at the level of the most elite human chess players, players who rely on highly trained intelligence to obtain their performance. Neural networks recognize faces at human levels and far surpassing human speeds. Computers can recognize spoken words under noise conditions that humans find baffling. But like Montaigne's animals, each program excels at only one kind of work. It is the generalizability of the Turing Test task that results in its testing not only a particular functionality, but the flexibility we take to indicate intelligence. Furthermore, the intensional character of intelligence, that the functionality be provided "in the right way", and not by mere memorization or brute computation, is also best tested by examining the flexibility of behavior of the subject under test.

It is a tall order to find a task that allows us to generalize from performance on a single functionality to performance on a broad range of functionalities while, *at the same time*, being not so far beyond current capability that its achievement is inconceivable in any reasonable time. It may well be that there are no appropriate prize tasks in the intersection of *audacious* and *achievable*.

Application of the principles

How do various proposals for tasks fare with respect to these principles? The three principles of flexibility, occasionality, and transparency are properties of the competition rules, not the competition task, so we can assume that an enlightened organizing body would establish them appropriately. But what of the task properties – absoluteness and reasonableness? For instance, would it be reasonable to use that most famous task for establishing intelligence in a machine, the Turing Test, as the basis for an inducement prize contest for AI?

The short answer is "no". I am a big fan of the Turing Test. I believe, and have argued in detail (Shieber, 2007), that it works exceptionally well as a conceptual sufficient condition for attributing intelligence to a machine, which was, after all, its original purpose. However, just because it works as a thought experiment addressing that philosophical question does not mean that it is appropriate as a concrete task for a research competition.

As an absolute criterion, the test as described by Turing is fine (though it has never been correctly put in place in any competition to date). But the Turing Test is far too difficult to serve as the basis of a competition. It fails the reasonableness principle. [3] Passing a full-blown Turing Test is so far beyond the state of the art that it is as silly to establish that criterion in an inducement prize competition as it is to establish transatlantic human-powered flight. It should go without saying that watered-down versions of the Turing Test based on purely relative performance among entrants is a non-starter.

The AI XPrize rules have not yet been established, but the sample criteria that Chris Anderson has proposed (XPRIZE Foundation, 2014) also fail our principles.

[3] As an aside, it is unnecessary, and therefore counterproductive, to propose tasks that are strict supersets of the Turing Test for a prize competition. For instance, tasks that extend the Turing Test by requiring non-textual inputs to be handled as well – audition or vision, say – or non-textual behaviors to be generated – robotic manipulations of objects, for instance – complicate the task, making it even less reasonable than the Turing Test itself already is.

The first part, presentation of a TED Talk on one of a set of one hundred predetermined topics can be satisfied by a "memorizing machine" (Shieber, 2014b) that has in its repertoire one hundred cached presentations. The second part, responding to some questions put to it on the topic of its presentation is tantamount to a Turing Test, and therefore fails the reasonableness criterion.[4]

What about special cases of the Turing Test, in which the form of the queries presented to the subject under test is more limited than open-ended natural-language communication, yet still requires knowledge and reasoning indicative of intelligence? The Winograd schema challenge (Levesque et al., 2012) is one such proposal. The test involves determining pronoun reference in sentences of the sort first proposed by Winograd (1972, page 33): "The city councilmen refused the demonstrators a permit because they feared violence." Determining whether the referent of *they* is *the city councilmen* or *the demonstrators* requires not only a grasp of the syntax and semantics of the sentence but an understanding of and reasoning about the bureaucratic roles of governmental bodies and social aims of activists. Presumably, human-level performance on Winograd schema queries requires human-level intelligence. The problem with the Winograd schema challenge may well be a lack of headroom. It might be the case that simple strategies could yield performance quite close to (but presumably not matching) human level. Such a state of affairs would make the Winograd schema challenge problematic as a guide for directing research towards machine intelligence.[5]

Are there better proposals? I hope so, though I fear there may not be any combination of task domain and award criterion that has the required properties. Intelligence may be a phenomenon about which we know sufficiently little that substantial but reasonable goals elude us for the moment. There is one plausible alternative however. We might wait on establishing an AI inducement prize contest until such time as the passing of the Turing Test itself seems audacious but achievable. That day might be quite some time.

Acknowledgements

I am indebted to Barbara Grosz and Todd Zickler for helpful discussions on the subject of this paper, as well as the participants in the AAAI "Beyond the Turing Test" workshop in January 2015 for their thoughtful comments.

[4] Anderson proposes that the system answer only one or two questions, which may seem like a simplification of the task. But to the extent that it is, it can be criticized on the same grounds as other topic- and time-limited Turing Tests (Shieber, 2014b).

[5] There are practical issues with the Winograd schema challenge as well. Generating appropriate challenge sentences is a specialized and labor-intensive process that may not provide the number of examples required for operating an incentive prize contest.

References

Anonymous. Computer simulating 13-year-old boy becomes first to pass Turing test. *The Guardian*, 9 June 2014. URL http://www.theguardian.com/technology/2014/jun/08/super-computer-simulates-13-year-old-boy-passes-turing-test.

Anonymous. The X-files. *The Economist*, 16 May 2015. URL http://www.economist.com/news/science-and-technology/21651164-want-new-invention-organise-competition-and-offer-prize-x-files.

British Parliament. An Act for Providing a Publick Reward for such Person or Persons as shall Discover the Longitude at Sea, 1714. URL http://cudl.lib.cam.ac.uk/view/MS-RGO-00014-00001/22.

Patrick Hayes and Kenneth Ford. Turing test considered harmful. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 972–977, 1995. URL http://www.ijcai.org/Past%20Proceedings/IJCAI-95-VOL%201/pdf/125.pdf.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. The Winograd schema challenge. In *Proceedings of the Italy 13th International Conference on Principles of Knowledge Representation and Reasoning*, pages 552–561, 2012. URL http://www.aaai.org/ocs/index.php/KR/KR12/paper/view/4492.

Christopher Lindquist. Quest for machines that think. Computerworld, 18 November 1991.

Susannah Locke. Forget the Turing test. this is a better way to measure artificial intelligence. *Vox*, 30 November 2014. URL https://www.vox.com/2014/11/30/7309879/turing-test.

Gary Marcus. What comes after the Turing Test? *The New Yorker*, 9 June 2014. URL http://www.newyorker.com/tech/elements/what-comes-after-the-turing-test.

Michel de Montaigne. *An Apology for Raymond Sebond*. Viking Penguin, New York, NY, 1987 [1576]. Translated and edited with an introduction and notes by M. A. Screech.

The Royal Aeronautical Society. Human powered flight: Regulations and conditions for the Kremer international marathon competition. Technical report, The Royal Aeronautical Society, London, England, August 1988. URL http://aerosociety.com/Assets/Docs/About_Us/HPAG/Rules/HP_Kremer_Marathon_Rules.pdf.

Alex Schroeder. The application and administration of inducement prizes in technology. Technical Report IP-11-2004, Independence Institute, Golden, Colorado, December 2004. URL http://i2i.org/articles/IP_11_2004.pdf.

Stuart M. Shieber. Lessons from a restricted Turing test. *Communications of the Association for Computing Machinery*, 37(6):70–78, 1994. doi: 10.1145/175208.175217.

Stuart M. Shieber. The Turing test as interactive proof. *Noûs*, 41(4):686–713, December 2007. doi: 10.1111/j.1468-0068.2007.00636.x.

Stuart M. Shieber. No, the Turing Test has not been passed. In *The Occasional Pamphlet*. 10 June 2014a. URL http://blogs.law.harvard.edu/pamphlet/2014/06/10/no-the-turing-test-has-not-been-passed/.

Stuart M. Shieber. There can be no Turing-Test-passing memorizing machines. *Philosophers' Imprint*, 14(16):1–13, June 2014b. URL http://hdl.handle.net/2027/spo.3521354.0014.016.

Alan M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, October 1950.

Moshe Y. Vardi. Would Turing have passed the Turing Test? *Communications of the ACM*, 57(9):5, 2014. doi: 10.1145/2643596.

Joseph Weizenbaum. Eliza—A computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery*, 9(1):36–45, 1966.

Terry Winograd. *Understanding Natural Language*. Academic Press, 1972.

XPRIZE Foundation. A.I. XPRIZE presented by TED, August 2014. URL http://www.xprize.org/ted.