14. On evaluating artificial intelligence systems: Competitions and benchmarks

Anthony G. Cohn, School of Computing, University of Leeds

This chapter discusses some approaches and methods used by the artificial intelligence (AI) community to measure and evaluate AI systems. It looks at the evolution of competitions, giving special attention to the Turing Test and the Winograd Schema Challenge. It also looks at the fascination of researchers for testing AI through games such as chess and Go. Several tests for measuring intelligence proposed for AI systems are examined, as well as the role of benchmark datasets in evaluating AI systems. The chapter ends with a discussion of the benefits and limitations of four approaches: custom dataset, benchmarks, competition and qualitative evaluation.

Introduction

This chapter discusses approaches and methods used by the artificial intelligence (AI) community to measure and evaluate AI systems. Ever since researchers started building AI systems, they have wanted to evaluate them. Some have sought to measure these systems against human benchmarks (such as playing human experts at chess or other games). Others have measured them against other AI systems. Some have done both.

Finding good benchmarks for evaluating systems, and conducting tests, is harder than it might seem. This is especially true because good methods apparently exist to evaluate human intelligence via standard tests and examinations.

Two major challenges for AI systems revolve around their "brittleness" and narrow scope (see also Chapter 2). AI systems that perform well have generally either been hand-engineered for particular problems, or trained on data relating to a particular task in a particular domain. This brittleness makes it hard to get good generalisation. Worse, AI systems are generally narrow in scope. They are able to tackle a limited set of problems and have limited knowledge of the world in general. Therefore, they will usually not even know when a problem lies beyond their competence.

Part of that challenge relates to the absence of common sense knowledge. One of the earliest challenges issued to the community was the need for AI systems to have "common sense" (McCarthy, 1959[1]). However, this remains one of the hardest aspects of human intelligence to build into an AI system.

That said, there are numerous examples of AI systems performing at, or even beyond human level in narrow, specialised domains and tasks. AlphaFold, for example, recently tackled protein structure prediction. It has been claimed that the latest version has solved a 50-year-old grand challenge problem in biology (Jumper et al., 2021[2]).

Tests for measuring intelligence

This section discusses tests for measuring intelligence proposed for AI systems. It then examines some of the competitions created to compare AI systems before looking at some benchmark datasets.

The Turing Test and other inducement prizes

There have been many tests proposed to evaluate AI systems. The Turing Test is probably the most famous (Turing, 1950[3]; Shieber, 2016[4]).[1] There have been various Turing Test competitions, including for the annual Loebner Prize.[2] The sometimes entertaining results have helped promulgate ideas about AI to the general public. However, the entrants have arguably not demonstrated any real important progress in AI.

Turing himself never proposed the test as a serious way of measuring AI systems or of measuring progress (Shieber, 2016[4]). The idea has been called "misguided and inappropriate" (Shieber, 1994[5]; Hayes and Ford, 1995[6]). Instead, Shieber (2016[4]) argues for new "inducement prize" contests – "award programs established to induce people to solve a problem of importance by directly rewarding the solver."

Inducement prizes have been around for centuries. Perhaps the most famous historical example is the Longitude Prize offered by the UK government in 1714. More recently, the IBM Watson AI XPRIZE[3] "challenges teams to demonstrate how humans can work with AI to tackle global challenges". The winner, scheduled to be announced in June 2021, was to win USD 5 million.

Five principles for inducement prizes

Shieber (2016[4]) suggests prizes should adhere to five principles (Table 14.1). Annual "bake-offs" favoured by some funding agencies between rival teams funded in a research programme tend to reduce diversity and do not reward ambitious risky approaches. As such, they have driven incremental progress towards specific targets. However, Shieber (2016[4]) argues that neither the "bake-offs" nor Turing Test prizes such as the Loebner Prize meet all five proposed principles.

Table 14.1. Five principles for inducement prizes

Principle	Rationale/description
Occasionality of occurrence	This ensures that awards are only given when warranted rather than automatically through an annual prize.
Flexibility of award	This aims to apply the spirit, rather than the letter, of the rules to determine a winner.
Transparency of result	The public should be able to inspect the results for both transparency and replicability.
Absoluteness of criteria	The award should satisfy absolute rather than relative criteria. In other words, it is not enough to be the best entrant in a competition to win an award.
Reasonableness of goal	This sets a constraint on the nature of the target(s); they should be beyond the current state of the art, but not impossibly so.

When the test involves a quantitative metric, it should have "headroom" (Shieber, 2016[4]). In other words, the level set should be a real indicator of a system that performs at human level. A speech recognition system, for example, may perform within a few percentage points of human-level performance. However, such a system may still be unable to understand speech in arbitrary contexts.

The question of "headroom" is also relevant to systems devised to address the Winograd Schema Challenge (WSC). The WSC aimed to ensure that statistics derived from mining large corpora could not be used to build successful systems rather than ensuring that the systems really understood the questions. At least for the current test set, statistics have been far more successful than anticipated (Kocijan et al., 2020[7]).

The Turing Test fails the reasonableness test since, in its full form, it is clearly too challenging for the current state of the art (Shieber, 2016[4]). He concludes a suitable form of a general AI inducement prize meeting all his principles is not possible in the short term.

The role of competitions in measuring artificial intelligence systems

This section discusses the main competitions that have come to play an increasingly important role in the AI community (Cohn, Dechter and and Lakemeyer, 2011[8]). Competitions of various kinds exist in nearly every subfield of AI. They range from theorem proving and SAT solvers… to trading agents, computational models of argumentation (Gaggl et al., 2018[9]), poker and other games… to object detection and recognition, among many others.[4]

Robot competitions are some of the few competitions that cover multiple aspects of AI and aim to evaluate integrated systems. Creating a competition that focuses on one specific aspect might seem to be a more achievable way of measuring progress in AI. However, in many cases, the sub-problem tackled is often called "AI complete" (Mallery, 1988[10]). In other words, the sub-problem is at least as hard as the AI problem in general. A solution to the "sub-problem", then, could be used to solve any AI challenge.

Box 14.1. The Winograd Schema Challenge

Levesque (2011[11]) first proposed an alternative to the Turing Test, which he called the Winograd Schema Challenge. A follow-up paper (Levesque, Davis and Morgenstern, 2012[12]) elaborates on the idea. The name derives from a pair of sentences given by Terry Winograd, a well-known early AI researcher:

1. The city councilmen refused the demonstrators a permit because they advocated violence.
2. The city councilmen refused the demonstrators a permit because they feared violence.

The two sentences only differ in one word, but the noun phrase referred to by the pronoun "they" changes. Winograd chose this sentence as a test for machine translation systems since, when translated to a gendered language such as French, "they" would be rendered as "elles" in the first sentence and "ils" in the second.

A key aspect of choice in such a pair of sentences is the requirement for general knowledge, or common sense knowledge, to identify the correct referent for the pronoun. It should not be possible to use selectional restrictions to solve the problem. In the sentence pair, "The men ate the burgers because they were [hungry/delicious]", for example, the AI only needs to know that men can eat things but burgers can't, and only food (not people) can be delicious. Similarly, using statistics of occurrence should not help determine the referent. For example, in the pair of sentences, "The motorcycle overtook the push-bike because it was going so fast/slow", the appropriate referent can be found through Google search statistics.

Competitions serve several purposes

Drive progress

Competitions drive progress in a community. The very announcement of a competition usually stimulates many of those working in the particular field to enter. Researchers enter both to evaluate their methods and techniques but also to gain kudos and potential career enhancement.

Measure the state of the art relatively objectively

Competitions provide a way of measuring the state of the art in a relatively objective way. The test is set externally rather than by the system's authors of what is now within the scope of the best systems and methods worldwide.

Build on failures

The failures of the systems in one competition generally drive improvements. This allows achieving a better performance in the following incarnation, thus setting short-term goals for improvements.

Build community

Competitions help the community come together and share expertise. While the events themselves are usually intensively competitive, there is usually a requirement for entrants to make their code available after the event. Workshops are organised to share what went well and what did not.

Criteria for successful competitions

A successful competition must set tasks that are neither too easy nor too hard but rather at the "cutting edge" so the best systems can at least partially succeed. As a field progresses, and successive instances of a competition are held in subsequent years, the tasks are typically made harder and more challenging.

This typically adds further realism so the challenges are closer to a real task that might be useful for humankind; the datasets might be larger; the robot environment less contrived, or stricter time limits applied.

In some cases, the initial competition was just too hard for the current state of the art. For example, a competition around the WSC at IJCAI-16 (Davis, Morgenstern and Ortiz, 2017[13]) contained two tasks. No system performed well enough in the first task (a simple pronoun resolution) to advance to the second round of solving a WSC.

Subsequently, neural language models such as BERT have achieved surprisingly good performance (at least to the challenge setters) (Devlin et al., 2019[14]). The best of these models achieved around 90% accuracy on the WSC273 dataset. Kocijan et al. (2020[7]) conclude that "new tests, more probing than the Winograd Schema Challenge, but still easy to administer and evaluate" are required.

Will progress via competitions that become incrementally harder every year eventually lead to human-level general intelligence and performance? Arguably, this is the way that humans learn, via the graded examinations in schools. However, whether this is the best approach to achieving AI is an open question.

Disadvantages of competitions

Competitions have undoubtedly led to much progress in AI, but they also have disadvantages. A new entrant to the field can sometimes generate new and innovative approaches. However, competitions more often drive incremental improvements to systems and methods.

The need for elaboration tolerance

Competitions can also blinker the community into solving problems that are easily assessable via a competition rather than pursuing fundamental and long-term research. Good competition design can address this issue. Ideally, the systems that researchers enter into competitions should have what has been called "elaboration tolerance" (McCarthy, 1959[1]).

A system has elaboration tolerance if it does not require substantive modification if the challenge itself was not substantively modified. More specifically, if a new version of the challenge is based on the original but differs in certain ways, the representation a system uses to solve the challenge only needs to be adapted or modified proportionately to the degree of the changes to the challenge.

Elaboration tolerance is a requisite of potential solutions to the challenge problems listed on the "Commonsense Reasoning" page.[5] Sloman (n.d.[15]) has suggested that elaboration tolerance is related to the ability of a system to "scale out" (rather than "scale up").

Whereas McCarthy's notion is purely about the *representation* a system uses, Sloman is concerned about the whole system. Sloman asks whether the system can be "combined with new mechanisms to perform a variety of tasks".

Hypothetically, a vision system could be used to label images in arbitrary sized corpora but cannot be used in a myriad of other contexts. These contexts might include producing descriptions of scenes or helping a robot in its activities as being able to scale up but not out. Both of these ideas are related to what has been called "brittleness" – the notion that a system can solve a particular class of problems well but fails on related but perhaps only subtly different problems.

Early artificial intelligence research with the games of chess and Go

AI has had a long fascination with game playing (Box 14.2). Indeed, some of the earliest AI researchers worked on building systems to play games, such as the simple tic-tac-toe/noughts and crosses, and checkers/draughts. Chess was a long-time challenge for AI, but in 1997 IBM's Deep Blue beat the reigning

world champion, Gary Kasparov. A key part of the success was the sheer brute computational force of Deep Blue.[6]

Box 14.2. Why has AI had such a fascination with game playing?

Since the earliest days of AI, for many reasons, researchers have been trying to develop programs to play games at a human level. Playing games well is often taken as a sign of intelligence, in particular more "intellectual" games such as chess. Developing an AI game player can thus be taken as a mark of progress towards machine intelligence.

Many skills required to play games well are also required in real life. These include the ability to plan a sequence of actions to achieve some goal. It also includes how to reason under uncertainty (what moves the other player will make or a selection of cards in a game of poker). There is also the use of probability to optimise decision making and machine learning to improve the system's performance, among other aspects.

Games have the advantage that framework can be easily implemented (i.e. the basic rules of the game rather than the best strategy), and can be simulated in a virtual world. A machine can learn by playing against itself (e.g. the AlphaZero system learning Go).

Of course, many aspects of real life such as language and vision are not present in game playing, at least as normally tackled in the AI game-playing literature. Real life is also much less structured than game playing, which usually has strict rules about turn taking and what moves are legal. Games are usually a "closed world" – everything about the game (except how to play well) can usually be completely described in a short tutorial, including all the possible objects and moves.

The game of Go, which is considerably more complex than chess, was considered to be a long way from being played well by computers. However, in 2016, a program named AlphaGo from DeepMind, beat the 9-dan player, Lee Sedol, 4-1 in a five game match. The match was organised as a formal competition, with a prize of US$1M for the winner. AlphaGo involved much human-coded knowledge, but a subsequent version, AlphaZero, was able to learn with only the rules of the game, just by playing itself many times.

Togelius (2016[16]) has gathered advice on running competitions for the AI game-playing community. Much of this advice is pragmatic. For example, any software developed to support the competition should be platform-agnostic. Other advice mirrors Shieber's principles, particularly that everything should be open-source.

Video games and virtual worlds

Togelius (2016[17]) has also argued that video games make an excellent test bed and benchmark for AI (as well as providing technology for new kinds of video games). He acknowledges that robots, operating in the real world, are perhaps the most obvious test bed for AI. However, he also notes several disadvantages of this approach. These include expense, speed (i.e. experiments typically take non-trivial amounts of time to perform, making learning difficult) and physical complexity.

Thus, games in their virtual worlds seem attractive, and in particular video games rather than board games. Moreover, designing AI systems to solve particular games does not necessarily imply anything about their ability to solve games in general. Therefore, Togelius records this as a motivation for designing the General Video Game Playing Competition (GVGAI) (Box 14.3). In this competition, entrant programs have to play ten new games known only to the competition organisers. To support the GVGAI, the General Video Game AI Framework has been developed to specify new games (Perez-Liebana et al., 2019[18]).

Specifying a standard framework for problems is a recurrent theme across competitions. For example, all the many problems in the Thousands of Problems for Theorem Provers (TPTP) benchmark (see below) are specified in the same way. Meanwhile, the Planning Domain Definition Language[7] has been used in the automated planning community and competitions for many years now. Over the years, it has had several extensions to increase its expressivity.

Many games that have been tackled are about perfect information: the only unknown is what actions the other player(s) might take. But other games mirror real life better in that only partial information is available and actions may be random. For example, the American Contract Bridge League has organised an annual competition since 1996. It has successively stronger players but still not at human champion level. Great progress has been made in poker in recent years, culminating in the *Libratus* system for which Sandholm and Brown (2018[19]) were also awarded the prestigious Minsky Medal.[8] Subsequently, Brown and Sandholm (2019[20]) designed a system called "Pluribus" that beat multiple human players simultaneously.

In response to the worldwide popularity of the game *Angry Birds*, an annual competition, AIBIRDS, has been held since 2012 (Renz et al., 2019[21]). While still played in an artificial environment, the game has many attractions from the point of view of measuring AI. First, there is incomplete knowledge of the physical parameters of the game. Second, there is an infinite set of actions available to an agent at any time. Third, it combines planning, image interpretation, reasoning, hypothesis formation and learning.

Interestingly, Renz et al. (2019[21]) point out that, "learning-based approaches have been largely unsuccessful. Despite all the successes of deep learning in the past few years, no deep-learning based agent has yet entered the semi-final round of [the] competition." Each year, humans also compete against the machines, and have dominated so far.

Box 14.3. General Video AI Framework

The core of the framework is a Video Game Description Language (VGDL). This provides a way of describing 2D video games concisely in a few dozen lines of plain text and can model both single and multi-player games. The originator listed desirable features of such a language. It should be

> *clear, human-readable and unambiguous. Its vocabulary should be highly expressive from the beginning, yet still extensible to novel types of games. Finally, its representation structure should be easy to parse and facilitate automatically generated games, in such a way that default settings and sanity checks enable most random game description to be actually playable (Schaul, 2013[22]).*

Games descriptions have two main parts. Level descriptions specify the 2D layout of the screen using different symbols. The game description proper specifies what the symbols in the level description mean in terms of the VGDL ontology (e.g. monsters or goals). Other parts of the game description define a possibly hierarchical set of objects with reference to an ontology, the set of interactions that happen when objects collide (e.g. swords kill monsters) and a set of termination criteria that define how/when the game ends.

The role of benchmarks in evaluating artificial intelligence systems

Benchmarks have long played an important role in AI. A benchmark is a dataset which is proposed as, or has become, a dataset by which different solution techniques are evaluated.[9] Benchmarks provide a way of comparing different solution methods on a single dataset or set of datasets. Sometimes they have emerged from research, initially used by a single group before becoming more widely used. Sometimes they have been explicitly created to stimulate research or evaluate competing methods.

In competitions, such as those discussed in the previous section, a test dataset is usually required. It then serves as a benchmark for the competition itself and often for subsequent research. However, it must be made openly available. Sometimes the dataset is reserved to be reused in subsequent competitions as it can be hard to create) as in the WSC. Often, there is an explicit call for benchmarks (e.g. in the International Competition on Computational Models of Argumentation (Gaggl et al., 2018[9]).

Some of the major benchmarks are listed here.[10] The TPTP (Box 14.4) has been an ongoing library of benchmark problems. Many are no longer seriously challenging, but it is continuously being extended with new problems (Sutcliffe, 2017[23]). In natural language understanding, General Language Understanding Evaluation (GLUE) (gluebench.com) is a "collection of NLU tasks including question answering, sentiment analysis and textual entailment, and an associated online platform for model evaluation, comparison and analysis." One task is a text inference variant of the WSC dataset. Performance of systems trained jointly on all the tasks in GLUE perform better than those trained on each task separately (Wang et al., 2019[24]).

Box 14.4. The TPTP Automated Theorem Proving Library

The TPTP library contains literally thousands of problems for evaluating and testing automated theorem provers (ATPs). AI has addressed the challenge of building ATPs since its earliest days [e.g. the geometry theorem proving system in Gelernter (Gelernter, 1959[25])]. The current library ranges from simple test problems for people to use to test their programs to open problems; some test problems used to be challenging for ATPs but are now trivial. The 53 domains range from Mathematics (relation, Boolean, Kleene algebras, set theory, topology, etc.) to Biology, Medicine, Social Sciences Philosophy, Puzzles and various aspects of Computer Science and AI (knowledge representation, planning, common sense reasoning, etc.). There is a standard representation for all problems in the library, making it easy for an ATP to read in a problem.

The benchmarks have all been more or less hand-curated/generated, but this can be expensive and difficult. Moreover, annotations on benchmarks are not always correct. Northcutt et al. (2021[26]), for example, estimated a 3.4% error rate across ten datasets surveyed. They thus recommend that special test sets should be carefully constructed to be error-free.

There is also work in the automatic generation of benchmarks and in particular, test sets. For example, a second competition is associated with AIBIRDS: the level generation competition (Stephenson et al., 2019[27]). Generating new levels or competitions is challenging, even for humans. This second level competition aims to stimulate and evaluate automated level generators. They note, "Submitted generators were required to deal with many physical reasoning constraints caused by the realistic nature of the game's environment, in addition to ensuring that the created levels were fun, challenging, and solvable."

In their discussion of the 2017 competition, Stephenson et al. (2019[27]) raise several interesting questions about the evaluation criteria for judging entries. At present, humans judge entries on a) "how fun and enjoyable the level is to play"; b) "how creative the level design is"; and c) "how well balanced the level of difficulty is". The definition of "fun" was left to the judges to avoid biasing them, and was used to determine the overall winner; b) and c) determine secondary prize-winners. The definition of similarity was also at the discretion of the judges.

No consideration has been given to automated judging, suggesting this task is not amenable to AI techniques. There has been some research on determining how to evaluate "fun" or creativity [e.g. (Boden, 1996[28]; Sweetser and Wyeth, 2005[29])], but this problem is still unsolved. The need for judges here contrasts with other competitions. Some are designed to avoid the need for human judges, such as the WSC (Levesque, Davis and Morgenstern, 2012[12]).

Conclusion

Doing research to win competitions risks being too driven by the goals of the competition. Entrants may seek to make enough incremental progress to do well in the next iteration rather than being truly innovative and devoting time to solve the fundamental problems remaining for machines to reach human-level AI.

Nonetheless, competitions have been valuable. Similarly, the availability of the many benchmark datasets allow for comparisons between different approaches. Table 14.2 summarises the main ways in which AI systems can be evaluated, along with the main advantages and disadvantages.

AI has always been most successful in narrow, usually technical domains, isolated from common sense and world knowledge. Finding good ways to measure common sense, and to build it into AI systems remains one of the greatest challenges for AI.

While many tests evaluate single aspects of intelligence, a challenge is to develop a comprehensive test for intelligence that is not susceptible to "game playing" by entrants. Typically, as soon as criteria are published which a system will be evaluated against, developers and researchers may try to optimise their systems for these specific criteria, rather than trying to create a more general system.

McCarthy's notion of elaboration tolerance is relevant in the search for a more general system. It tries to test whether a system has solved a specific problem, or instead has enough knowledge and reasoning/computational ability to solver a broad class of problems. Multiple conflicting evaluation criteria would be one strategy to address this issue: it would be impossible to optimise against all criteria simultaneously. Of course, having hold-out, unseen tests also helps.

This is part of a more general issue that AI systems do not know what they don't know. A typical AI system to, say, diagnose liver pathologies from data, only "knows" about (some) pathologies and aspects of the training data. It has no idea about more general knowledge of the world, even perhaps relatively related areas such as pathologies in a different part of the body. Even a young child has some appreciation of the limitations of its knowledge, and can confidently say "I don't know" in answer to some input.

Over the history of AI, many technologies have been held out as the solution for machine intelligence – from rule-based systems and early neural nets to Bayesian Networks and Deep Neural Nets. Many have demonstrated successes. However, they have all, so far, been shown to have problems. Not one offers a single solution to enable general machine intelligence.

Machine intelligence may well require an amalgam of different techniques and approaches. Recent initiatives towards "neuro-symbolic" systems, for example, combine neural net technologies with symbolic reasoning. Each has its advantages. Some have high-level reasoning and ability to explain reasoning for symbolic systems. Others can learn from large amounts of data for neural systems. The challenge is how to build integrated systems that successfully combine multiple technologies.

Table 14.2. A comparison of different approaches to evaluating AI systems

Evaluation method	Advantages	Disadvantages
Custom dataset	Can be designed specifically to test hypotheses. Potential for it to become a benchmark in the future.	Datasets can be hard to produce/acquire. No comparator results available.
Benchmark	Previous comparator results available. No effort required to produce it.	May not display all the advantages of the new method. Suitable benchmarks not always available. Tends to encourage research towards improving benchmark performance rather than AI in general.
Competition	Evaluation datasets/problem formulation specifically designed for problem. Encourages multiple entrants. Often there is a workshop associated where competitors discuss their results and the consequences for future progress.	Effort required to organise. Can be challenging to pitch competition at suitable level with respect to state of the art – it must be enough of a challenge but not too much. May require sponsorship to run or to encourage entrants.
Qualitative evaluation (i.e. human inspection of system)	Can be tailored to specific dataset and problem. Not rigorous but useful for understanding strengths and weaknesses of an approach.	Requires human effort and can suffer from selection bias.

Note: The different rows are often used in combination: for example, a competition might involve a custom dataset, as well as benchmarks and perhaps a qualitative evaluation.

References

Anderson, M. (11 May 2017), "Twenty years on from Deep Blue vs Kasparov: How a chess match started the big data revolution", The Conversation blog, https://theconversation.com/twenty-years-on-from-deep-blue-vs-kasparov-how-a-chess-match-started-the-big-data-revolution-76882. [32]

Boden, M. (1996), "Creativity", in Boden, M. (ed.), *Artificial Intelllgence (Handbook of Perception and Cognition)*, Academic Press, Cambridge, MA. [28]

Brown, N. and T. Sandholm (2019), "Superhuman AI for multiplayer poker", *Science*, Vol. 365/6456, pp. 885-890, http://dx.doi.org/10.1126/science.aay2400. [20]

Brown, N. and T. Sandholm (2018), "Superhuman AI for heads-up no-limit poker: Libratus beats top professionals", *Science*, Vol. 359/6374, pp. 418-424, http://dx.doi.org/10.1126/science.aao1733. [19]

Cohn, A., R. Dechter and G. and Lakemeyer (2011), "Editorial: The competition section: A new paper category", *Artificial Intelligence*, Vol. 175/9-10, p. iii, https://doi.org/10.1016/S0004-3702(11)00060-9. [8]

Commonsense Reasoning (n.d.), "Problem Page", webpage, http://commonsensereasoning.org/problem_page.html (accessed on 1 January 2021). [33]

Davis, E., L. Morgenstern and C. Ortiz (2017), "The first Winograd Schema Challenge at IJCAI-16", *AI Magazine*, Vol. 38/3, pp. 97-98, https://doi.org/10.1609/aimag.v38i4.2734. [13]

Devlin, J. et al. (2019), "BERT: Pre-training of deep bidirectional transformers for language understanding", *arXiv*, http://dx.doi.org/arXiv:1810.04805v2. [14]

Fox, M. and D. Long (2002), *PDDL+: Modelling Continuous Time-dependent Effects*. [31]

Gaggl, S. et al. (2018), "Summary report of the second international competition on computational models of argumentation", *AI Magazine*, Vol. 39/4, p. 73, https://doi.org/10.1609/aimag.v39i4.2781. [9]

Gelernter, H. (1959), "Realization of a geometry theorem proving machine", *IFIP Congress*, pp. 273-281. [25]

Hall, M. (3 April 2019), "What makes a good benchmark dataset?", Views and News about Geoscience and Technology blog, https://agilescientific.com/blog/2019/4/3/what-makes-a-good-benchmark-dataset. [30]

Hayes, P. and K. Ford (1995), "Turing test considered harmful", in *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, Morgan Kaufmann Publishers, San Francisco. [6]

Jumper, J. et al. (2021), "Highly accurate protein structure prediction with AlphaFold", *Nature*, Vol. 596/7873, pp. 583-589, http://dx.doi.org/10.1038/s41586-021-03819-2. [2]

Kocijan, V. et al. (2020), "A review of Winograd Schema Challenge datasets and approaches", *arXiv*, http://dx.doi.org/abs/2004.13831. [7]

Levesque, H. (2011), "The Winograd Schema Challenge", *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning 2011*. [11]

Levesque, H., E. Davis and L. Morgenstern (2012), *The Winograd Schema Challenge*, AAAI Press, Palo Alto, CA. [12]

Mallery, J. (1988), "Thinking about foreign policy: Finding an appropriate role for artificially intelligent computers", in *The 1988 Annual Meeting of the International Studies Association, St. Louis, MO*. [10]

McCarthy, J. (1959), "Programs with common sense", in *Proceedings of Teddington Conference on the Mechanization of Thought Processes, 1959*, Stanford University, Stanford, CA, http://jmc.stanford.edu/articles/mcc59/mcc59.pdf. [1]

Northcutt, C., A. Athalye and J. Mueller (2021), "Pervasive label errors in test sets destabilize machine learning benchmarks", *arXiv preprint*, http://dx.doi.org/arXiv:2103.14749. [26]

Perez-Liebana, D. et al. (2019), *General Video Game AI*, Morgan Kaufman, San Francisco. [18]

Renz, J. et al. (2019), "AI meets Angry Birds", *Nature Machine Intelligence*, Vol. 1/328, https://doi.org/10.1038/s42256-019-0072-x. [21]

Schaul, T. (2013), "A video game description language for model-based or interactive learning", *2013 IEEE Conference on Computational Intelligence in Games (CIG)*, pp. 1-8, http://dx.doi.org/10.1109/CIG.2013.6633610. [22]

Shieber, S. (2016), "Principles for designing an AI competition, or why the Turing test fails as an inducement prize", *AI Magazine*, Vol. 37/1, pp. 91-96, https://doi.org/10.1609/aimag.v37i1.2646. [4]

Shieber, S. (1994), "Lessons from a restricted Turing test", *Communications of the ACM*, Vol. 37/6, pp. 70-78, http://dx.doi.org/10.1145/175208.175217. [5]

Sloman, A. (n.d.), “John McCarthy – Some Reminiscences”, webpage, https://www.cs.bham.ac.uk/research/projects/cogaff/sloman-jmc-aisb.html (accessed on 1 January 2021). [15]

Stephenson, M. et al. (2019), “The 2017 AIBIRDS level generation competition”, *IEEE Transactions on Games*, Vol. 11/3, pp. 275-284, http://dx.doi.org/10.1109/TG.2018.2854896. [27]

Sutcliffe, G. (2017), “The TPTP problem library and associated infrastructure: from CNF to TH0, TPTP v6. 4.0”, *Journal of Automated Reasoning*, Vol. 59, pp. 583-402, http://dx.doi.org/doi.org/10.1007/s10817-017-9407-7. [23]

Sweetser, P. and P. Wyeth (2005), “GameFlow: A model for evaluating player enjoyment in games”, *Computers in Entertainment*, Vol. 3/3, p. 3, https://doi.org/10.1145/1077246.1077253. [29]

Togelius, J. (2016), “AI researchers, video games are tour friends!”, in Merelo, J. et al. (eds.), *Computational Intelligence.International Joint Conference, IJCCI 2015 Lisbon, Portugal, 12-14 November 2015, Revised Selected Papers*, https://doi.org/10.1007/978-3-319-48506-5_1. [17]

Togelius, J. (2016), “How to run a successful game-based AI competition”, in *IEEE Transactions on Computational Intelligence and AI in Games*, http://dx.doi.org/10.1109/TCIAIG.2014.2365470. [16]

Turing, A. (1950), “Computing machinery and Intelligence”, *Mind*, Vol. LIX/236, pp. 433-460, https://doi.org/10.1093/mind/LIX.236.433. [3]

Wang, A. et al. (2019), “Glue: A multi-task benchmark and analysis platform for natural language understanding”, *arXiv*, http://dx.doi.org/arXiv:1804.07461. [24]

Notes

[1] Turing called the test the “imitation game” (now the title of a film about Turing’s life). He proposed it as a test of whether a machine could display intelligent behaviour indistinguishable to a human observer confronted with both the machine and another human; the observer could only communicate with the machine and the other human by typed natural language.

[2] The Loebner Prize, instituted in November 1991, is normally held annually. Competitors enter chatbots and human judges (since 2019 members of the public) judge which is the most human-like. It was instituted by Hugh Loebner and offers a prize of USD 100 000 for “the first program that judges cannot distinguish from a real human in a Turing test that includes deciphering and understanding text, visual and auditory input. Once this is achieved, the annual competition will end.”

[4] These are often associated with AI conferences, but sometimes independently organised. Kaggle hosts many competitions: www.kaggle.com/competitions, many with prizes. The platform www.aicrowd.com/ also hosts challenges to enable “data science experts and enthusiasts to collaboratively solve real-world problems”. Another platform hosting competitions is https://competitions.codalab.org/. All URLs accessed 24 March 2021.

[5] As an example of elaboration tolerance, consider one of the problems on Commonsense Reasoning (n.d.[33]): “A gardener who has valuable plants with long delicate stems protects them against the wind by staking them; that is, by plunging a stake into the ground near them and attaching the plants to the stake with string”. Variants of the problem suggested included: “What would happen: If the stake is only placed upright on the ground, not stuck into the ground? If the string were attached only to the plant, not to the stake? To the stake, but not to the plant? If the plant is growing out of rock? Or in water? If, instead of string, you use a rubber band? Or a wire twist-tie? Or a light chain? Or a metal ring? Or a cobweb? If instead of tying the ends of the string, you twist them together? Or glue them?...”

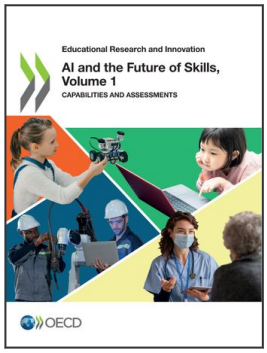
[6] Some have speculated that a bug in the code contributed to Deep Blue’s success. In this view, the bug provoked a random move for the computer. Kasparov apparently attributed this move to a “deeper strategy” and subsequently played too cautiously (Anderson, 11 May 2017[32]).

[7] Fox and Long (2002[31]) write: “The adoption of a common formalism for describing planning domains fosters far greater reuse of research and allows more direct comparison of systems and approaches, and therefore supports faster progress in the field. A common formalism is a compromise between expressive power (in which development is strongly driven by potential applications) and the progress of basic research (which encourages development from well-understood foundations). The role of a common formalism as a communication medium for exchange demands that it is provided with a clear semantics.”

[8] The team behind AlphaGo won the inaugural Minsky Medal. The Marvin Minsky Medal is awarded by IJCAI for “Outstanding Achievements in AI”.

[9] Hall (Hall, 3 April 2019[30]) proposes characteristics of a good machine learning benchmarks such as being openly available, well documented, labelled and with an accompanying demonstration.

[10] There are many others including those included in repositories such as www.kaggle.com/ https://paperswithcode.com/datasets.



From:

AI and the Future of Skills, Volume 1

Capabilities and Assessments

Access the complete publication at:
https://doi.org/10.1787/5ee71f34-en

Please cite this chapter as:

Cohn, Anthony G. (2021), "On evaluating artificial intelligence systems: Competitions and benchmarks", in OECD, *AI and the Future of Skills, Volume 1: Capabilities and Assessments*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/d755c6d6-en