In AAAI-15 Workshop on Beyond the Turing Test 2015.

General Video Game Playing as a Benchmark for Human-Competitive AI

Joel Lehman and Risto Miikkulainen
Dept. of Computer Science, The University of Texas at Austin
{joel,risto}@cs.utexas.edu

Abstract

Text-based conversational Turing test benchmarks for artificial intelligence have significant limitations: It is possible to do well by only faking intelligence, thereby subverting the test's intent. An ideal replacement test would facilitate and focus AI research, be easy to implement and automate, and ensure that human-competitive performance implied a powerful and general AI. The idea in this paper is that general video game playing is one such promising candidate for an improved human-level AI benchmark. To pass such a test, a computer program must succeed in efficiently completing an unannounced and diverse suite of video games, interacting with the game only through simulated versions of the same information streams available to a human player; the test is easy to automate but difficult to exploit, and can stress nearly all aspects of human intelligence through strategic sampling of the vast library of existing video games. In this way, general video game playing may provide the basis for a simple but effective benchmark competition for human-level AI.

Introduction

In Turing tests, computers attempt to convince humans that they are also human. The most popular versions of such tests limit interaction to text-based messages (Mauldin 1994), although there exist other generalizations (Hingston 2009). Two main motivations for implementing Turing test contests are (1) that they can provide a reliable benchmark for recognizing human-level AI and (2) that they can focus and encourage progress in artificial intelligence (AI) research.

Yet while Turing tests lead sometimes to impressive behavior, passing them has so far been disconnected from ground-breaking advances in AI. Furthermore, it appears doubtful that in at least the dominant chatbot form such tests have had much connection to mainstream AI research (Shieber 1994). Indeed, because the current capabilities of AI fall far short from creating truly intelligent general conversational agents, the winners of chatbot-based imitation tests tend to rely upon clever hard-coded conversational gambits. The main problem is that performance in human imitation does not seem to degrade smoothly or provide many tractable sub-tasks. As a result Turing test performance does not correlate with progress in AI research.

Beyond encouraging AI progress and being impervious to exploitation, an ideal test would also be fully automated and easy to apply. Such an ideal test would enable researchers to potentially train AI through repeated interactions with the test without exorbitant cost. However, human imitation tests inherently require interaction with real humans, which renders training an AI algorithm over thousands or millions of interactive trials impractical (which is otherwise a typical AI approach). Thus in total, these theoretical and practical concerns motivate exploring alternatives to imitation-based Turing tests. In particular, the next section motivates general video game playing as one such practical alternative.

Intelligence and General Video Game Playing

With edge-case exceptions, a video game is an interactive computer simulation designed to challenge a human player in varied dimensions of performance and intelligence. Reflecting diverse human interests, genres of video games are many and diffuse, stressing sweeping aspects of human intelligence. For example, completing text-based adventure games requires understanding language and context; progressing through puzzle games requires creative thinking, and building and simulating internal models; succeeding in trivia or educational games requires subject knowledge or the ability to learn new concepts; and completing modern 3D role-playing games requires complex sensory integration and navigating scripted social interactions. Note that while not traditionally considered games, computerized standardized tests (including intelligence tests themselves) could also be accommodated into this framework.

More fundamentally, completing an unforeseen video game requires learning the structure and purpose of the game, and recognizing visual or auditory signals indicative of progress. Importantly, when humans play video games, they must infer the privileged information that is nearly always provided to game-playing AI, such as an abstract representation of the game, heavily-engineered game state information, and forward models to simulate the future. Thus if an AI is similarly constrained (which could also include limiting reaction times to those of humans), completing a large and diverse set of video games indicates that it possesses a wide range of human-level cognitive abilities.

The pragmatic benefits of a general video game benchmark include that it is easily automated and that it can include games from the large library of existing video games. Emulators of many computers and video game systems exist, and they can simulate execution of video games from many different platforms on a single modern computer. Such emulators can be modified to feed video and auditory signals into an AI program, and accept the simulated input device signals that the AI generates in response, thereby allowing the AI to interact with the game with the same constraints as a human player. An example of such a setup is the Arcade Learning Environment (ALE) (Bellemare et al. 2013), which provides an emulator-based interface between AI algorithms and classic Atari 2600 video games.

An important question is whether passing such a test would require human-level AI. If the set of games on which an AI is tested is chosen strategically to be diverse and is randomly-selected from a large set, then brute-force or hand-coded solutions become inviable. That is, because games as a whole are expansively diverse and can require a panoply of intellectual abilities to attain success, it would be challenging (if not impossible) to derive simple exploitative strategies that exhaustively cover the entire gamut of video games.

Another important question is if a general game playing test would help catalyze AI research. Because the answer is implementation-dependent, this issue is addressed in the next section, which describes a possible implementation.

Proposed Competition Implementation

There are two existing general game playing AI competitions, the general game playing competition (GGP) (Genesereth, Love, and Pell 2005), and the general video game competition (GVGP) (Levine et al. 2013). However, GGP is focused on abstract logical games; and both GGP and GVGP provide AI agents with an exact formal description of the game, enabling them to forward-simulate the game perfectly (which in general is impossible for a human player).

In contrast to these existing contests, this paper proposes a generalization of the ALE environment to arbitrary computer and console emulators. In the proposed contest, AI agents would be limited to observing incoming video and auditory signals of video games, and would act only through simulated versions of the same input devices available to human players. A diverse and unannounced suite of games would be chosen to stress a maximal range of intelligent behavior. One intuitive method to score the competition would be to average the rankings of agents' scores over the suite of games. Similarly, an intuitive measure of human-level performance is for an AI to achieve an averaged ranking better than half of human players evaluated on the same games.

An important consideration is how to design the competition so that it helps not only evaluate if an AI is of human-level competence but also encourages progress in AI research. One issue that emerged with chatbot imitation contests is that performance at imitating humans at high-level tasks may degrade too abruptly to offer any gradient for progress. In other words, it may be in effect only a binary test – either you have all the complex components of intelligence necessary to imitate a human, or you fail dramatically.

A similar problem may arise if current AI techniques compete on the most taxing version of the proposed video game benchmark. Instead, it may be more productive to incrementally scale up difficulty as AI capabilities mature, and to offer diverse competitions over games of different genres and complexities. For example, perhaps a suite of games designed for young children might offer a reasonable first target. There is evidence that at least simple Atari games may be tractable with current AI technology from only raw pixel values (Hausknecht et al. 2013). Additionally, a selection of sub-competitions where AIs are given different crutches (similar to the existing GGP and GVGP competitions) may help to engage a wider range of AI communities. For example, to fit current reinforcement learning algorithms to the benchmark, the AI can be provided the game score directly to bypass the need to infer the measure of progress.

However, computer chess provides a cautionary tale about pursuing such a sub-dividing approach to an extreme; it was originally believed that human-level chess would require human-level AI, but a brute-force solution was developed with little relation to general AI. Thus over iterations of the competition, the design of sub-competitions should be rebalanced between generality and tractability. This consideration highlights a important rule of thumb unheeded by some chatbot-based Turing tests: The test's original intent (to measure human-level intelligence) should be valued over rigidity in how it is applied.

Conclusion

With careful consideration it may be possible to create a simple, flexible, and effective general video game benchmark competition that can both validate human-level AI and encourage progress in AI research.

References

Bellemare, M. G.; Naddaf, Y.; Veness, J.; and Bowling, M. 2013. The arcade learning environment: an evaluation platform for general agents. *Journal of Artificial Intelligence Research* 47(1):253–279.

Genesereth, M.; Love, N.; and Pell, B. 2005. General game playing: Overview of the AAAI competition. *AI magazine* 26(2):62.

Hausknecht, M.; Lehman, J.; Miikkulainen, R.; and Stone, P. 2013. A neuroevolution approach to general atari game playing. In *IEEE Transactions on Computational Intelligence and AI in Games*.

Hingston, P. 2009. A turing test for computer game bots. *Computational Intelligence and AI in Games, IEEE Transactions on* 1(3):169–186.

Levine, J.; Congdon, C. B.; Ebner, M.; Kendall, G.; Lucas, S. M.; Miikkulainen, R.; Schaul, T.; and Thompson, T. 2013. General video game playing. *Dagstuhl Follow-Ups* 6.

Mauldin, M. L. 1994. Chatterbots, tinymuds, and the turing test: Entering the loebner prize competition. In *AAAI*, volume 94, 16–21.

Shieber, S. M. 1994. Lessons from a restricted turing test. *Communications of the ACM* 37(6):70–78.