Editorial Introduction to the Special Articles in the Spring Issue

Beyond the Turing Test

Gary Marcus, Francesca Rossi, Manuela Veloso

■ *The articles in this special issue of* AI Magazine *include those that propose specific tests and those that look at the challenges inherent in building robust, valid, and reliable tests for advancing the state of the art in AI.*

Alan Turing's renowned test on intelligence, commonly known as the Turing test, is an inescapable signpost in AI. To people outside the field, the test — which hinges on the ability of machines to fool people into thinking that they (the machines) are people — is practically synonymous with the quest to create machine intelligence. Within the field, the test is widely recognized as a pioneering landmark, but also is now seen as a distraction, designed over half a century ago, and too crude to really measure intelligence. Intelligence is, after all, a multidimensional variable, and no one test could possibly ever be definitive truly to measure it. Moreover, the original test, at least in its standard implementations, has turned out to be highly gameable, arguably an exercise in deception rather than a true measure of anything especially correlated with intelligence. The much ballyhooed 2015 Turing test winner Eugene Goostman, for instance, pretends to be a thirteen-year-old foreigner and proceeds mainly by ducking questions and returning canned one-liners; it cannot see, it cannot think, and it is certainly a long way from genuine artificial general intelligence.

Our hope is to see a new suite of tests, part of what we have dubbed the Turing Championships, each designed in some way to move the field forward, toward previously unconquered territory. Most of the articles in this special issue stem from our first workshop toward creating such an event, held during the AAAI Conference on Artificial Intelligence in January 2015 in Austin, Texas.

The articles in this special issue can be broadly divided into those that propose specific tests, and those that look at the challenges inherent in building robust, valid, and reliable tests for advancing the state of the art in artificial intelligence.

In the article My Computer is an Honor Student — But How Intelligent Is It? Standardized Tests as a Measure of AI, Peter Clark and Oren Etzioni argue that standardized tests developed for children offer one starting point for testing machine intelligence.

Ernest Davis in his article How to Write Science Questions That Are Easy for People and Hard for Computers, proposes an alternative test called SQUABU (science questions appraising basic understanding) that aims to asks questions that are easy for people but hard for computers.

In Toward a Comprehension Challenge, Using Crowdsourcing as a Tool, Praveen Paritosh and Gary Marcus propose a crowdsourced comprehension challenge, in which machines will be asked to answer open-ended questions about movies, YouTube videos, podcasts, stories, and podcasts.

The article The Social-Emotional Turing Challenge, by William Jarrold and Peter Z. Yeh, considers the importance of social-emotional intelligence and proposes a methodology for designing tests that assess the ability of machines to infer things like motivations and desires (often referred to in the psychological literature as theory of mind.)

In Artificial Intelligence to Win the Nobel Prize and Beyond: Creating the Engine for Scientific Discovery, Hiroaki Kitano urges the field to build AI systems that can make significant, even Nobel-worthy, scientific discoveries.

In Planning, Executing, and Evaluating the Winograd Schema Challenge, Leora Morgenstern, Ernest Davis, and Charles L. Ortiz, Jr., describe the Winograd Schema Challenge, a test of commonsense reasoning that is set in a linguistic context.

In the article Why We Need a Physically Embodied Turing Test and What It Might Look Like, Charles L. Ortiz, Jr., argues for tests, such as a construction challenge (build something given a bag of parts), that focus on four aspects of intelligence: language, perception, reasoning, and action.

Measuring Machine Intelligence Through Visual Question Answering, by C. Lawrence Zitnik, Aishwarya Agrawal, Stanislaw Antol, Margaret Mitchell, Dhruv Batra, and Devi Parikh, argues for using visual question answering as an essential part of a multimodal challenge to measure intelligence.

Tomaso Poggio and Ethan Meyers in Turing++ Questions: A Test for the Science of (Human) Intelligence, which also focuses on visual questions, propose to develop tests where competitors must not only match human behavior but also do so in a way that is consistent with human physiology, in this way aiming to use a successor to the Turing test to bridge between the fields of neuroscience, psychology, and artificial intelligence.

The article I-athlon: Toward a Multidimensional Turing Test, by Sam Adams, Guruduth Banavar, and Murray Campbell, proposes a methodology for designing a test that consists of a series of varied events, in order to test several dimensions of intelligence. Kenneth Forbus also argues for testing multiple dimensions of intelligence in his article Software Social Organisms: Implications for Measuring AI Progress.

In the article Principles for Designing an AI Competition, or Why the Turing Test Fails as an Inducement Prize, Stuart Shieber discusses several desirable features for an inducement prize contest, contrasting them with the current Turing test.

Douglas Lenat in WWTS (What Would Turing Say?) takes a step back and focuses instead on synergy between human and machine, and the development of conjoint superhuman intelligence.

While the articles included in this issue propose and discuss several kinds of tests, and we hope to see many of them being deployed very soon, they should be considered merely as a starting point for the AI community. Challenge problems, well chosen, can drive media interest in the field, but also scientific progress. We hope therefore that many AI researchers participate actively in formalizing and refining the initial proposals described in these articles and discussed at our initial workshops.

In the meantime, we have created a website[1] with pointers to presentations, discussions, and most importantly ways for interested researchers to get involved, contribute, and participate in these successors to the Turing test.

Note

[1] www.math.unipd.it/~frossi/btt.

Gary Marcus is founder and chief executive officer of Geometric Intelligence and a professor of psychology and neural science at New York University.

Francesca Rossi is a research scientist at the IBM T.J. Watson research center, (on leave from the University of Padova).

Manuela Veloso is the Herbert A. Simon University Professor in the Computer Science Department at Carnegie Mellon University.