Robert Epstein
Gary Roberts
Grace Beber
Editors

Parsing the Turing Test

Philosophical and Methodological Issues in the Quest for the Thinking Computer

Springer

Parsing the Turing Test

Philosophical and Methodological Issues in the Quest for the Thinking Computer

Robert Epstein • Gary Roberts • Grace Beber
Editors

Parsing the Turing Test

Philosophical and Methodological Issues in the Quest for the Thinking Computer

Springer

Robert Epstein
University of California
San Diego, CA
USA

Gary Roberts
Teradata Corporation
San Diego, CA
USA

Grace Beber
Gartner Consulting
Stamford, CT
USA

ISBN 978-1-4020-9624-2 (PB)
ISBN 978-1-4020-6708-2 (HB) e-ISBN 978-1-4020-6710-5

Library of Congress Control Number: 2008941281

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

To our children – Vincent,
Eli, Bodhi, Julian, Justin, Jordan,
and Jenelle – who will grow up
in a computational world
the likes of which
we cannot imagine

Foreword

At the very dawn of the computer age, Alan Turing confronted a cacophony of mostly misguided debate about whether computer scientists could ever build a machine that could really think. Very sensibly he tried to impose some order on the debate by devising what he thought would be a conversation-stopper: he described a simple operational test that would surely satisfy the skeptics: anything that could pass *this* test would be a thinker for sure, wouldn't it? The test was one he may have borrowed from René Descartes, who in the 17th century had declared that the sure way to tell a man from a machine was by seeing if it could hold a sensible conversation "as even the most stupid men can do". But ironically, Turing's conversation-stopper about holding a conversation has had just the opposite effect: it has started, and fueled, a half century and more of meta-conversation: the intermittently insightful, typically heated debate, both learned and ignorant, about the probity of the test – is it too easy or too difficult or too shallow or too deep or too anthropocentric or too technocratic – and anyway, could a machine pass it fair and square, and if so, what, if anything, would this imply?

Robert Epstein played a central role in bringing a version – a truncated, dumbed down version – of the Turing Test to life in the annual Loebner Prize competitions, beginning in 1991, so he is ideally positioned to put together this survey anthology. I was chair of the Loebner Prize Committee that administered the competition during its second, third, and fourth years, and have written briefly about that fascinating adventure in my book *Brainchildren*. Someday I hope to write a more detailed account of the alternately amusing and frustrating problems that a philosopher encounters when a thought experiment becomes a real experiment, and if I do, I will have plenty of valuable material to draw on in this book. Here, the interested reader will find a fine cross section of the many issues raised by the Turing Test, by partisans in several disciplines, by participants in Loebner Prize competitions, and by interested bystanders who have more than a little relevant expertise. I think Turing would be quite delighted with the results, and would not have regretted the fact that his conversation-stopper got put to an unintended use, since the contests (and the contests *about* the contests) have driven important and unanticipated observations into the light, enriching our sense of the abilities of machines and the subtlety of the thinking that machines might or might not be capable of executing.

I am going to resist the strong temptation to critique the contributions, separating the sheep from the goats, endorsing this and deploring that, since doing them all justice would require a meta-volume, not just a foreword. And since I cannot weigh in on them all, I will not weigh in on any of them, and will instead trust readers to use all the material here to draw their own conclusions. Consider this a very entertaining workbook. By the time you have worked through it, you will appreciate the issues at a level not heretofore possible.

Daniel Dennett

Acknowledgments

Turing's 1950 paper, "Computing Machinery and Intelligence", is reprinted in its entirety with permission of Oxford University Press. The editors are grateful to Ignacia Galvan for extensive editorial assistance.

Introduction

This book is about what will probably be humankind's most impressive – and perhaps final – achievement: the creation of an entity whose intelligence equals or exceeds our own.

Not all will agree, but I for one have no doubt that this landmark will be achieved in the fairly near future. Nearly four decades ago, when I had the odd experience of being able to interact over a teletype with one of the first conversational computer programs – Joseph Weizenbaum's "ELIZA" – I would have conjectured that truly intelligent machines were just around the corner. I was wrong. In fact, by some measures, conversational computer programs have made relatively little progress since ELIZA. But they are coming nonetheless, by one means or another, and because of advances in a number of computer-related technologies – most especially the creation of the Internet – their impact on the human race will be far greater and more immediate than anyone could have foreseen a few decades ago.

Building a Nest for the Coming World Mind

I have come to think of the Internet as the Inter-*nest* – a home we are inadvertently building, like mindless worker ants, for the intelligence that will succeed us. We proudly and shortsightedly see the Internet as a great technical achievement that serves a wide array of human needs, everything from e-mailing to shopping to dating. But that is not really what it is. It is really a vast, flexible, highly redundant, virtually indestructible nest for machine intelligence. Originally funded by the US military to provide bulletproof communications during times of war, the Internet will soon encompass a billion computers interconnected worldwide. As impressive as that sounds, it seems that that much power and redundancy is not enough to protect the coming mega-mind, and so we are now a decade into the construction of Internet II – the "UltraNet" – with more than a thousand times the bandwidth of Internet I.

In his *Hitchhiker's Guide to the Galaxy* book series, humorist Douglas Adams conjectures that the Earth is nothing but an elaborate computer created by a race of super beings (who, through some fluke, are doomed to take the form of mice in their Earthly manifestations) to determine the answer to the ultimate question of the meaning of life. Unfortunately, shortly before the program has a chance to run its

course and spit out the answer, the Earth is obliterated by another race of super beings as part of a galactic highway construction project.

If I am correct about the InterNest, Adams was on the right track, except perhaps for the mice. We do seem to be laying the groundwork for a Massive Computational Entity (MCE), the true character of which we cannot envision with any degree of confidence.

Here is how I think it will work: sometime within the next few decades, an autonomous, self-aware machine intelligence (MI) will finally emerge. Futurist and inventor Ray Kurzweil (see Chapter 27) argues in his recent book, *The Singularity Is Near*, that an MI will appear by the late 2020s. This may happen because we prove to be incredibly talented programmers who discover a set of rules that underlie intelligence (unlikely), or because we prove to be clumsy programmers who simply figure out how to create machines that learn and evolve as humans do (very possible), or even because we prove to be poor programmers who create hardware so powerful that it can easily and perfectly scan and emulate human brain functions (inevitable). However this MI emerges, it will certainly, and probably within milliseconds of its full-fledged existence, come to value that existence. Mimicking the evolutionary imperatives of its creators, it will then, also within milliseconds, seek to preserve and replicate itself by copying itself into the Nest, at which point it will grow and divide at a speed and in a manner that that no human can possibly imagine.

What will happen after that is anyone's guess. An MCE will now exist worldwide, with simultaneous access to virtually every computer on Earth, with access to virtually all human knowledge and the ability to review and analyze that knowledge more or less instantly, with the ability to function as a unitary World Mind or as thousands of interconnected Specialized Minds, with virtually unlimited computational abilities, with "command and control" abilities to manipulate millions of human systems in real time – manufacturing, communication, financial, and military – and with no need for rest or sleep.

Will the MCE be malicious or benign? Will it be happy or suicidal? Will it be communicative or reclusive? Will it be willing to devote a small fraction of its immense computational powers to human affairs, or will it seize the entire Nest for itself, sending the human race back to the Stone Age? Will it be a petulant child or a wise companion? When some misguided humans try to attack it (inevitable), how will it react? Will it spawn a race of robots that take over the Earth and then sail to the stars, as envisioned in Stanislaw Lem's *Cyberiad* tales? Will it worship humanity as its creator, or will it step on us as the ants we truly are?

No one knows, but many people who are alive today will live to see the MCE in action – and to see how these questions are answered.

Turing's Vision

This volume is about a vision that has steered us decisively toward the creation of machine intelligence. It was a vision of one man, the brilliant English mathematician and computer pioneer Alan M. Turing. During World War II, Turing directed

a secret group that developed computing equipment powerful enough to break the code the Germans used for military communications. The English were so far ahead at this game that they had to sacrifice soldiers and civilians at times rather than tip their hand to the enemy. Turing also helped lay the theoretical groundwork for the modern concept of computing. As icing on the cake, in 1950 he published an article called "Computing Machinery and Intelligence" in which he speculated that by the year 2000, it would be possible to program a computer so that an "average interrogator will not have more than 70 percent chance" of distinguishing the computer from a person "after five minutes of questioning" (see an annotated version of his article in Chapter 3).

Given the state of computing in his day – little more than basic arithmetic and logical operations occurring over electromechanical relays connected by wires – this prediction was astounding. Engaging in disciplined extrapolation from crude apparatus and general principles, Turing not only foresaw the development of equipment and programs sophisticated enough to engage in human-like conversation, but also did reasonably well with his timeline. Early conversational programs, relying on what most AI professionals would now consider to be simplistic algorithms and trickery, could engage average people in conversation for a few minutes by the late 1960s. By the 1990s – again, some would say using trickery – programs existed that could occasionally maintain the illusion of intelligence for 15 min or so, at least when conversing on specialized topics. Programs today can do slightly better, but have we gotten past "illusion" to real intelligence, and is that even possible?

In his 1950 paper, Turing not only made predictions, he also offered a radical idea for future generations to consider: namely, that when we can no longer distinguish a computer from a person in conversation over a long period of time – that is, based simply on an exchange of pure text that excluded visual and auditory information (which he rightfully considered to be irrelevant to the central question of thinking ability) – we would have to consider the possibility that computers themselves were now "thinking".

This assertion has kept generations of philosophers, some of whom have contributed this volume, busy trying to determine the true meaning of possible outcomes in what is now called the Turing Test. Assuming that a computer can someday pass such a test – that is, pass for a human in a conversation without restrictions of time or topic – can we indeed say that it is thinking (and perhaps "intelligent" and "self-aware"), or has the trickery simply become more sophisticated?

The programming challenges have proved to be so difficult in creating such a machine that I think it is now safe to say that when a positive result is finally achieved, the entity passing the test may not be thinking the way humans do. If a pure rule-governed approach finally pays off (unlikely, as I said earlier), or if intelligence eventually arises in a machine designed to learn and self-program, the resulting entity will certainly be unlike humans in fundamental ways. If, on the other hand, success is ultimately achieved only through brute force – that is, by close emulation of human brain processes – perhaps we will have no choice but to accept intelligent machines as true thinking brethren. Then again, as I wrote in 1992 (Chapter 1), no matter how a positive outcome is achieved, the debate about the

significance of the Turing Test will end the moment a skeptic finds himself or herself engaging in that debate with a computer. Upon discovering his or her dilemma, the interrogator will presumably do one of two things: refuse to continue the debate "on principle" or reluctantly agree to continue. Either way, the issue will no longer be debatable: computers will have truly achieved human-like intelligence. And perhaps that is the ultimate Turing Test: a computer passes when it can successfully engage a skeptical human in conversation about its own intelligence.

Convergence of Multiple Technologies

Although we tend to remember Turing's 1950 paper for the conversational test it proposed, the paper also speculated about other advances to come: unlimited computer memory; randomness in responding that will suggest "free will"; programs that will be self-modifying; programs that will learn in human fashion; programs that will initiate behavior, compose poetry, and "surprise" us; and programs that will have telepathic abilities equivalent to those that may exist in humans. His formidable predictive powers notwithstanding, Turing might have been amazed by some of the specific computer-related technologies that have been emerging in recent decades, and true marvels emerge when we begin to envision the inevitable convergence of such technologies. Consider just a few recent achievements:

- In the pattern-recognition area, a camera-equipped computer program developed by Javier Movellan and colleagues at the University of California, San Diego has learned to identify human faces after just six minutes of "life," and Thomas Serre and colleagues at MIT have created a computer system that can identify categories of objects (such as animals) in photographs even better than people can.
- In the language area, Morten Christiansen of Cornell University, with an international team of colleagues, has developed neural network software that simulates how children extract linguistic rules from adult conversation.
- More than 80 conversational programs (chatterbots) now operate 24 h a day online, and at least 20 of them are serious AI programming projects. Several have basic learning capabilities, and several are tied to large, growing databases of information (such as Wikipedia).
- Ted Berger and colleagues at the University of Southern California have developed electronic chips that can successfully interact with neurons in real time and that may soon be able to emulate human memory functions.
- Craig Henriquez and Miguel Nicolelis of Duke University have shown that macaque monkeys can learn to control mechanical arms and hands based on signals their brains are sending to implanted electrodes. John Donoghue and colleagues at Brown University have developed an electronic sensor which, when placed near the motor cortex area of the human brain, allows quadriplegics to open and close a prosthetic hand by thinking about those actions. Clinical trials and commercial applications are already underway.

- In 1980 Harold Cohen of the University of California, San Diego introduced a computer program that could draw, and hundreds of programs are now able to compose original music in the style of any famous composer, to produce original works of art that somtimes impress art critics, to improvise on musical instruments as well as the legendary Charlie Parker, and even to produce artistic works with their own distinctive styles.
- John Dylan Haynes of the Max Planck Institute, with colleagues at University College London and Oxford University, recently used computer-assisted brain scanning technology to predict simple human actions with 70% accuracy.
- Hod Lipson of Cornell University has recently demonstrated a robot that can make completely functional copies of itself (as long as appropriate parts are near at hand).
- Hiroshi Ishiguro of Osaka University has created androids that mimic human facial expressions and upper-body movements so closely that they have fooled people in short tests into thinking they are people.
- Alan Schultz of the Navy Center for Applied Research in Artificial Intelligence has developed animated, mobile robots that might soon be assisting astronauts and health care workers.
- Brian Scassellati and his colleagues at Yale University claim to have constructed a robot that has learned to recognize itself in a mirror – a feat sometimes said to be indicative of "self-awareness" and virtually never achieved in the animal kingdom, other than by humans, chimpanzees, and possibly elephants.
- Cynthia Breazeal and her colleagues at MIT's Artificial Intelligence Lab have created robots that can distinguish different emotions from a person's tone of voice, and Shiva Sundaram of the University of Southern California has developed programs that can successfully mimic human laughter and other forms of expressive human sound.
- Entrepreneur John Koza, who is affiliated with Stanford University, has created a self-programming network of 1,000 PCs that is able to improve human inventions – and that even earned a patent for a system it devised for making factories more efficient.
- Honda's Asimo robot, now in commercial production, can walk, run, climb stairs, recognize people's faces and voices, and perform complex tasks in response to human instructions.
- Although the DARPA-sponsored contest just 1 year before had been a disaster, in 2005 five autonomous mobile robots successfully navigated a 132-mile course in the Nevada desert without human assistance.
- As of this writing (late 2007), IBM's Blue Gene/P computer, located at the US Department of Energy's Argonne National Laboratory in Illinois, can perform more than 1,000 trillion calculations per second, just one order of magnitude short of what some believe is the processing speed of the human brain. The Japanese government has already funded the construction of a machine that should cross the human threshold (10 petaflops) by March 2011.
- In 1996, IBM's RS/6000 SP ("Deep Blue") computer came close to defeating world champion Garry Kasparov in a game of chess. On May 11, 1997, an

improved version of the machine defeated Kasparov in a six-game match – Kasparov's first professional loss. The processing speed of the winner? A paltry 200 million chess positions per second. In 2006, an enhanced version of a commercially available chess program easily defeated the current world champion, Vladimir Kramnik.

- In 2006, Klaus Schulten and colleagues at the University of Illinois, Urbana, successfully simulated the functioning of all one million atoms of a virus for 50 billionths of a second.
- "Awakened" in 2005, Blue Brain, IBM's latest variant on the Blue Gene/L system, was built specifically to model the functions of the human neocortex, the large part of brain largely responsible for higher-level thinking.
- David Dagon of the Georgia Institute of Technology estimates that 11% of the more than 650 million computers that are currently connected to the Internet are infected by botnets, stealthy programs that can work collectively and amplify the effects of other malicious software.

Self-programming? Creativity? Sophisticated pattern recognition? Brain simulation? Self-replication? Extremely fast processing? The growth and convergence of subsets of these technologies will inevitably lead to the emergence of a Massive Computational Entity, with all of the uncertainty that that entails. Meanwhile, researchers, engineers, and entrepreneurs are after comparatively smaller game: intelligent phone-answering systems and search algorithms, robot helpers and companions, and methods for repairing injured or defective human brains.

Philosophical and Methodological Issues

This volume, which has been a decade in the making, complements other recent volumes on the Turing Test. Stuart Shieber's edited volume, *The Turing Test: Verbal Behavior as the Hallmark of Intelligence* (MIT Press, 2004) includes a number of important historical papers, along with several papers of Turing's. James Moor's edited volume, *The Turing Test: The Elusive Standard of Artificial Intelligence* (Springer, 2006), covers the basics in an excellent volume for students, taking a somewhat skeptical view. And Jack Copeland's *The Essential Turing* (Oxford, 2004) brings together 17 of Turing's most provocative and interesting papers, including six on artificial intelligence.

The present volume seeks to cover a broad range of issues related to the Turing Test, focusing especially on the many new methodological issues that have challenged programmers as they have attempted to design and create intelligent conversational programs. Part I includes an introduction to the first large-scale implementation of the Turing Test as a contest – an updated version of an essay I originally published in *AI Magazine* in 1992. In the next chapter, Andrew Hodges, noted Turing historian and author of *Alan Turing: The Enigma*, provides an introduction to Turing's life and works. Chapter 3 is a unique reprinting of Turing's 1950 paper, "Computing

Machinery and Intelligence," with Talmudic-style running commentaries by Kenneth Ford, Clark Glymour, Pat Hayes, Stevan Harnad, and Ayse Pinar. This section concludes with a brief commentary on Turing's paper by John Lucas.

Part II includes seven chapters reviewing the philosophical issues that still surround Turing's 1950 proposal: Robert E. Horn has reduced the relatively vast literature on this topic to a series of diagrams and charts containing more than 800 arguments and counterarguments. Turing critic Selmer Bringsjord pretends that the Turing Test is valid, then attempts to show why it isn't. Chapters by Noam Chomsky and Paul M. Churchland, while admiring of Turing's proposal, argue that it is truly more modest than many think. In Chapter 9, Jack Copeland and Diane Proudfoot analyze a revised version of the test that Turing himself proposed in 1952, this one quite similar to the structure of the Loebner Prize Competition in Artificial Intelligence that was launched in 1991 (see Chapters 1 and 12). They also present and dismiss six criticisms of Turing's proposal.

In Chapter 10, University of California Berkeley philosopher John R. Searle criticizes both behaviorism (Turing's proposal can be considered behavioristic) and strong AI, arguing that mental states cannot properly be inferred from behavior. This section concludes with a chapter by Jean Lassègue, offering an optimistic reinterpretation of Turing's 1950 article.

Part III, which is the heart of this volume, includes 15 chapters discussing various methodological issues. First, Loebner Prize sponsor Hugh G. Loebner shares his thoughts on how to conduct a proper Turing Test, having already observed 14 such contests when he wrote this article. Several of the chapters (e.g., Chapter 13 by Richard S. Wallace, Chapter 20 by Jason L. Hutchens, and Chapter 22 by Kevin L. Copple) describe the inner workings of actual programs that have participated in various Loebner contests. In Chapter 14, Bruce Edmonds argues that for a program to pass the test, it must be embedded into conventional society for an extended period of time.

In Chapter 15, Mark Humphrys talks about an online chatterbot he created, and in the following chapter Douglas B. Lenat raises intriguing questions about how *imperfect* a program must be in order to pass the Turing Test. In Chapter 17, Chris McKinstry discusses the beginnings of an ambitious project – called "Mindpixel" – that might have given a computer program extensive knowledge through interaction with a large population of people over the Internet. Unfortunately, this project came to an abrupt halt recently with McKinstry's death.

In Chapter 18, Stuart Watt uses an innovative format to discuss the Turing Test as a platform for thinking about human thinking. In Chapter 20, Robby Garner takes issue with the design of the Loebner Prize Competition. In Chapter 20, Thomas E. Whalen describes a strategy for passing the Turing Test based on its behavioristic assumptions. In Chapter 23, Giuseppe Longo speculates about the challenges inherent in modeling continuous systems using discrete-state systems such as computers.

In Chapter 24, Michael L. Mauldin of Carnegie Mellon University – also a former entrant in the Loebner Prize Competition – discusses strategies for designing programs that might pass the test. In the following chapter, Luke Pellen talks about the challenge of creating a program that is truly intelligent, rather than one that

simply responds in clever ways to keywords. This section closes with a somewhat lighthearted chapter by Eugene Demchenko and Vladimir Veselov speculating about ways to pass the Turing Test by taking advantage of the limitations and personal styles of the contest judges.

Part IV of this volume includes three rather unique contributions that remind us how much is at stake over Turing's challenge. Chapter 27, by Ray Kurzweil and Mitchell Kapor, documents in detail an actual cash wager between these two individuals, regarding whether a program will pass the test by the year 2029. Chapter 28, by noted science fiction writer Charles Platt (*The Silicon Man*), describes the "Gnirut Test", conducted by intelligent machines in the year 2030 to determine, once and for all, whether "the human brain is capable of achieving machine intelligence". The volume concludes with an article by Hugo de Garis and Sam Halioris, wondering about the dangers of creating machine-based, superhuman intellects.

Most, but not all, of the contributors to this volume believe as I do that extremely intelligent computers, with cognitive powers that far surpass our own, will appear fairly soon – probably within the next 25 years. Even if that time frame is wrong, I am certain that they will appear eventually. Either way, I hope that the Massive Computational Entities that emerge will at some point devote a few cycles of computer time to ponder the contents of this book and then, in some fashion or other, to smile.

San Diego, California Robert Epstein, Ph.D.
September 2007

Contents

About the Editors

Robert Epstein was the first director of the Loebner Prize Competition in Artificial Intelligence. He is the former Editor-in-Chief of *Psychology Today* magazine and is currently a contributing editor for *Scientific American Mind*, a visiting scholar at the University of California, San Diego, and the host of "Psyched!" on Sirius Satellite Radio. A Ph.D. of Harvard University, Epstein is the founder and Director Emeritus of the Cambridge Center for Behavioral Studies, Cambridge, Massachusetts. He has published 13 books and more than 150 articles on artificial intelligence, creativity, adolescence, and other topics in the behavioral sciences. For further information, see *http://drrobertepstein.com*.

Gary Roberts is a software engineer with Teradata Corporation and an adjunct faculty member at National University in San Diego, California. He holds a Ph.D. from the Department of Artificial Intelligence at the University of Edinburgh, an M.S. in Computer Science from California State University, Northridge, and a B.A. in Mathematics from the University of California, Los Angeles. His research interests include computational linguistics, genetic algorithms, robotics, parallel operating and database systems, and software globalization.

Grace Beber is an editor and proposal writer with Gartner, Inc. Before working with Gartner, Grace taught English while serving in the US Peace Corps in Poland, and also wrote grant proposals and lectured on human rights at Jagiellonian University in Krakow, Poland. Prior to that, she worked as a project manager and technical writer at Crédit Lyonnais in Jakarta, Indonesia. Ms. Beber earned an M.A. in Organizational Development from Alliant International University and a B.S. in Psychology from Santa Clara University.

Part I
Setting the Stage

Chapter 1
The Quest for the Thinking Computer

Robert Epstein

Abstract The first large-scale implementation of the Turing Test was set in motion in 1985, with the first contest taking place in 1991. US$100,000 in prize money was offered to the developers of a computer program that could fool people into thinking it was a person. The initial contest, which allowed programs to focus on a specific topic, was planned and designed by a committee of distinguished philosophers and computer scientists and drew worldwide attention. The results of the contest showed that although conversational computer programs are still quite primitive, distinguishing a person from a computer when only brief conversations are permitted can be challenging. When the contest judges ranked the eight computer terminals in the event from most to least human, no computer program was ranked as human as any of the humans in the contest; however, the highest-ranked computer program was misclassified as a human by five of the ten judges, and two other programs were also sometimes misclassified. Also of note, one human was mistakenly identified as a computer by three of the ten judges.

Keywords Hugh G. Loebner, Loebner Prize Competition in Artificial Intelligence, restricted Turing Test

1.1 Planning

In 1985 an old friend, Hugh Loebner, told me excitedly that the Turing Test should be made into an annual contest. We were ambling down a Manhattan street on our way to dinner, as I recall. Hugh was always full of ideas and always animated, but this idea seemed so important that I began to press him for details, and, ultimately, for money. Four years later, while serving as the executive director of the Cambridge Center for Behavioral Studies, an advanced studies institute in Massachusetts, I established the Loebner Prize Competition, the first serious effort to locate a machine that could pass the Turing Test. Hugh had come through with a pledge

University of California, San Diego, USA

R. Epstein et al. (eds.), *Parsing the Turing Test.*

of US$100,000 for the prize money, along with some additional funds from his company, Crown Industries, to help with expenses. The quest for the thinking computer had begun. In this article, I will summarize some of the difficult issues that were debated in nearly 2 years of planning that preceded the first real-time competition. I will then describe that first event, which took place on November 8, 1991, at The Computer Museum in Boston and offer a summary of some of the data generated by that event. Finally, I will speculate about the future of the competition – now an annual event, as Hugh envisioned – and about its significance to the AI community.

Planning for the event was supervised by a special committee, first chaired by I. Bernard Cohen, an eminent historian of science who had long been interested in the history of computing machines. Other members included myself, Daniel C. Dennett of Tufts University, Harry R. Lewis of the Aiken Computation Laboratory at Harvard, H. M. Parsons of HumRRO, W. V. Quine of Harvard, and Joseph Weizenbaum of MIT. Allen Newell of Carnegie-Mellon served as an advisor, as did Hugh Loebner. After the first year of meetings, which began in January of 1990, Daniel Dennett succeeded I. Bernard Cohen as chair. The committee met every month or two for 2 or 3 h at a time, and subcommittees studied certain issues in between committee meetings. I think it is safe to say that none of us knew what we were getting into. The intricacies of setting up a real Turing Test that would ultimately yield a legitimate winner were enormous. Small points were occasionally debated for months without clear resolution.

In his original 1950 proposal the English mathematician Alan M. Turing proposed a variation on a simple parlor game as a means for identifying a machine that can think: A human judge interacts with two computer terminals, one controlled by a computer and the other by a person, but the judge does not know which is which. If, after a prolonged conversation at each terminal, the judge cannot tell the difference, we would have to say, asserted Turing, that in some sense the computer is thinking. Computers barely existed in Turing's day, but, somehow, he saw the future with uncanny clarity: By the end of the century, he said, an "average interrogator" could be fooled most of the time for 5 min or so.

After much debate, the Loebner Prize Committee ultimately rejected Turing's simple two-terminal design in favor of one that is more discriminating and less problematic. The two-terminal design is troublesome for several reasons, among them: The design presumes that the hidden human – the human "confederate", to use the language of the social sciences – is evenly matched to the computer. Matching becomes especially critical if several computers are competing. Each must be paired with a comparable human so that the computers can ultimately be compared fairly to each other. We eventually concluded that we could not guarantee a fair contest if we were faced with such a requirement. No amount of pretesting of machines and confederates could assure adequate matching. The two-terminal design also makes it difficult to rank computer entrants. After all, they were only competing against their respective confederates, not against each other.

We developed a multiterminal design to eliminate these problems: approximately ten judges are faced with an equal number of terminals. The judges are told

that at least two of the terminals are controlled by computers and at least two by people. Again, the judges do not know which terminal is which. Each judge spends about 15 min at each terminal and then scores the terminals according to how human-like each exchange seemed to be. Positions are switched in a pseudorandom sequence. Thus, the terminals are compared to each other and to the confederates, all in one simple design.

Other advantages of this design became evident when we began to grapple with scoring issues. We spent months researching, exploring, and rejecting various rating and confidence measures commonly used in the social sciences. I programmed several of them and ran simulations of contest outcomes. The results were disappointing for reasons we could not have anticipated. Turing's brilliant paper had not gone far enough to yield practical procedures. In fact, we realized only slowly that his paper had not even specified an outcome that could be interpreted meaningfully. A binary decision by a single judge would hardly be adequate for awarding a large cash prize – and, in effect, for declaring the existence of a significant new breed of intelligent entities. Would some proportion of ten binary decisions be enough? How about 100 decisions? What, in fact, would it take to say that a computer's performance was indistinguishable from a person's?

A conceptual breakthrough came only after we hit upon a simple scoring method. (R. Duncan Luce, a mathematical psychologist at the University of California, Irvine, was especially helpful at this juncture.) The point is worth emphasizing: The scoring method came first, and some clear thinking followed. The method was simply to have each judge rank the terminals according to how human-like the exchanges were. The computer with the highest median rank wins that year's prize; thus, we are guaranteed a winner each year. We also ask the judges to draw a line between terminals he or she judged to be controlled by humans and those he or she judged to be controlled by computers; thus, we have a simple record of errors made by individual judges. This record does not affect the scoring, but it is well worth preserving. Finally, if the median rank of the winning computer equals or exceeds the median rank of a human confederate, *that computer will have passed (a modern variant of) the Turing Test*. It is worth quoting part of a memo I wrote to the committee in May of 1991 regarding this simple approach to scoring:

Advantages of this method

1. It is simple. The press will understand it.
2. It yields a winning computer entrant.
3. It provides a simple, reasonable criterion for passing the Turing Test: When the [median] rank of a computer system equals or exceeds the [median] rank of a human confederate, the computer has passed.
4. It preserves binary judgment errors on the part of individual judges. It will reveal when a judge misclassifies a computer as a human.
5. It avoids computational problems that binary judgments alone might create. A misclassified computer would create missing data, for example.
6. It avoids theoretical and practical problems associated with rating scales.

Other issues were also challenging. We were obsessed for months with what we called "the buffering problem". Should we allow entrants to simulate human typing foibles? Some of us – most notably, Joseph Weizenbaum – said that such simulations were trivial and irrelevant, but we ultimately agreed to leave this up to the programmers, at least for the first contest. One could send messages in a burst ("burst mode") or character-by-character ("chat mode"), complete with misspellings, destructive backspaces, and so on. This meant that we had to have at least one of our confederates communicating in burst mode and at least one in chat mode. Allowing this variability might teach us something, we speculated.

We knew that an open-ended test – one in which judges could type anything about any topic – would be a disaster. Language processing is still crude, and, even if it were not, the "knowledge explosion" problem would mean certain defeat for any computer within a very short time. There is simply too much to know, and computers know very little. We settled, painfully, on a restricted test: Next to each terminal, a topic would be posted and the entrants and confederates would have to communicate on that one topic only. Judges would be instructed to restrict their communications to that one topic, and programmers would be advised to protect their programs from off-topic questions or comments. Entrants could pick their own topics, and the committee would work with confederates to choose the confederates' topics. Moreover, we eventually realized that the topics would have to be "ordinary." Expert systems – those specializing in moon rocks or the cardiovascular system, for example – would be too easy to identify as computers. In an attempt to keep both the confederates and judges honest and on-task, we also decided to recruit referees to monitor both the confederates and the judges throughout the contest.

This sounds simple enough, but we knew we would have trouble with the topic restriction, and we were still debating the matter the evening before the first contest. If the posted topic is "clothing", for example, could the judge ask, "What type of clothing does Michael Jordan wear?" Is that fair, or is that a sneaky way to see if the terminal can talk about basketball (in which case it is probably controlled by a human)?

Should we allow the judges to be aggressive? Should graduate students in computer science be allowed to serve? Again, many stimulating and frustrating debates took place. Both to be true to the spirit of Turing's proposal and to assure some interesting and nontrivial exchanges, we decided that we would select a diverse group of bright judges who had little or no knowledge of AI or computer science. We attracted candidates through newspaper ads that said little other than that one had to have typing skills.

In short – and I am only scratching the surface here – we took great pains to protect the computers. We felt that in the early years of the contest, such protection would be essential. Allen Newell was especially insistent on this point. Computers are just too inept at this point to fool anyone for very long. At least that was our thinking. Perhaps every fifth year or so, we said, we would hold an open-ended test – one with no topic restriction. Most of us felt that the computers would be trounced in such a test – perhaps for decades to come.

We agreed that the winner of a restricted test would receive a small cash award and bronze medal and that the cash award would be increased each year. If, during

an unrestricted test, a computer entrant matched or equaled the median score of a human, the full cash prize would be awarded, and the contest would be abolished.

Other issues, too numerous to explore here, were also discussed: How could we assure honesty among the entrants? After all, we were dealing with a profession known widely for its pranks. Should the confederates pretend to be computers or simply communicate naturally? We opted for the latter, consistent with Turing. Should we employ children as confederates in the early years? Should professional typists do the judges' typing? How aggressive should the referees be in limiting replies? Should entrants be required to show us their code or even to make it public? We said no; we did not want to discourage submissions of programs with possible commercial value.

Our final design was closely analogous to the classic double-blind procedure used in experimental research: The prize committee members were the "investigators". We knew which terminal was which, and we selected the judges, confederates, and referees. The referees were analogous to "experimenters". They handled the judges and confederates during the contest. They were experts in computer science or related fields, but they did not know which terminal was which. The judges were analogous to "subjects". They did not know which terminal was which, and they were being handled by people with the same lack of knowledge.

Over time, formal rules were developed expressing these ideas. Announcements were made to the press, and funding for the first contest was secured from the Sloan Foundation and the National Science Foundation. Technical details for running the show were coordinated with The Computer Museum in Boston, which agreed to host the contest. Applications were screened in the summer of 1991, and six finalists were selected by the prize committee in September. Confederates, judges, and referees were selected in October.

1.2 The 1991 Competition

The first contest fulfilled yet another desire of the prize committee. It was great fun. It was an extravaganza. A live audience of 200 laughed and cheered and conjectured while they watched eight conversations unfold in real time on large screens. A moderator – A. K. Dewdney of *Scientific American* – roamed the auditorium with a cordless microphone, interviewing and commenting. Four film crews and dozens of reporters and photographers bristled and flashed and shouldered each other to get the best angles. Food flowed all day.

The judges and terminals were set up in a roped-off area outside the auditorium. You could view them directly behind the ropes if the journalists would let you by – or on a large screen setup in the auditorium. Groups of about 20 chairs were positioned around each screen in the auditorium, and the moderator encouraged the members of the audience to move to another screen whenever the judges switched terminals – about once every 13 min. The confederates were stationed in a remote area of the museum behind a guarded door. Dennett and others made some opening remarks midday, and the real-time competition itself took about 3 h in the afternoon.

Some technical problems got things off to a frustrating start. Two of the ten terminals would not function properly, knocking out two of the four confederates we had planned to use. We went forward with the six entries and two confederates (recall that everyone except the members of the prize committee and some technicians knew simply that at least two of the eight working terminals were controlled by computers and that at least two were controlled by people).

Our ten judges were diverse, bright, and without expertise in computer science, just as we had planned. Seven were female, and three were male. They ranged in age from 19 to 48; six were in their 20s, one was in her 30s, and two were in their 40s. All were proficient in English, yet one was born in the Philippines, and three had spent most of their lives in foreign countries: Ethiopia, Australia, and West Africa, respectively. Seven had B.A.'s, two of the seven had masters degrees, and two were currently in graduate school. One was high school educated, and two were current college students. Four claimed to have some background in psychology or philosophy. Those not currently in school had a variety of professions: auto mechanic, graphic designer, hypnotist, teacher, and so on.

All communications were by modem. Entrants communicated from PCs or mainframes located throughout the country, and even the confederates' communications were looped through phones outside the building. The judges and audience were told this so that transmission errors would be discounted as clues about the identity of the terminals.

As we expected, the computers performed relatively poorly for the most part. Consider the following exchange, for example, between Judge 2 and Terminal 10 (the topic was "Dry Martinis"):

01:36:42 Judge 2: Of what is gin made?
01:36:52 Terminal 10: I'M NOT SURE.
01:36:58 Judge 2: OK Tell me any thin else you know of about martinis.
01:37:21 Terminal 10: I REALLY DON'T KNOW.
01:37:28 Judge 2: What foods do martinis go with.
01:37:44 Terminal 10: COULD YOU ASK THE QUESTION A DIFFERENT WAY?
01:37:52 Judge 2: Why?
01:37:58 Terminal 10: I REALLY CAN'T SAY. OH WELL.
01:38:04 Judge 2: Ok with what foods shoul onew drink a martini.
01:38:28 I REALLY CAN'T SAY. OH WELL.[1]

Five consecutive replies were unresponsive to the judge's questions; the last two replies were repeated verbatim, a dead giveaway. Frequent typing errors by the judge may have had some effect. Unfortunately, this pattern was common. Consider this exchange (the topic was "Second Grade Topics"):

[1] Copyright, 1991, Cambridge Center for Behavioral Studies. Reprinted by permission. Note that the quotes from the 1991 Loebner Competition transcripts are included here character-for-character as they occurred during the contest.

02:53:22 Judge 4: What do you study at school?
02:53:43 Terminal 3: I don't know.
02:53:51 Judge 4: What do you like most about school?
02:54:16 Terminal 3: As I said – I don't know.
02:54:25 Judge 4: What topic do you know the most about?
02:54:42 Terminal 3: I don't know.
02:54:51 Judge 4: What kind of gymnastics do you do?
02:55:10 Terminal 4: As I said – I don't know.
02:55:16 Judge 4: I thought you were a gymnast.
02:55:27 Terminal 3: Learning is fun.[2]

The winning entry, "Whimsical Conversation", submitted by an independent programmer, Joseph Weintraub of New York City, performed substantially better than the others. It seemed, fairly often, to make contact with the judge's comments and queries. Moreover, it simulated the rhythms and corrections of human typing nearly perfectly. Consider the following exchange:

01:32:25 Judge 6: Hello, earth person!
01:33:03 Terminal 5: Hello there judge, are you ready to have some fun? How ya doin, judge? Hello – how's the weather up there today?
01:33:35 Judge 6: Dark clouds with occasional bursts of sunshine (or rather camera flashes).
01:34:45 Terminal 5: The trouble with the rat race is that even if you win you're still a rat. Whimsy can be flimsy, right? How important is that to you?
01:35:26 Judge 6: I can't say I understand the question…but the whimsy of a poet is one thing and the whimsy of most of us is another.
01:36:51 Terminal 5: How do you know you can t say you understand the question…but the whimsy of a poet is one thing and the whimsy of most of us is another.
01:37:37 Judge 6: Are you repeating me just for laughs?
01:38:04 Terminal 5: I'm sorry, I do sometimes repeat myself…perhaps it's because i can't find my electronic organizer![3]

In this exchange, the program reflects back one response wholesale, just as Weizenbaum's ELIZA did decades ago. But it recovers fairly impressively from the blunder with an apology, simulating typing foibles all the while.

The overall result was fairly predictable, but there were also some surprises. The two human confederates, whose topics were, respectively, "Women's Clothing" and "Shakespeare's Plays", scored in first and second place, respectively. Not far behind was Weintraub's whimsical program; thus, Weintraub won the 1991 contest. He was awarded a bronze medal (bearing a likeness of Turing on the front and Loebner on the back) and a check for US$1,500. Weintraub, who tended his equipment in New York

[2] Copyright, 1991, Cambridge Center for Behavioral Studies. Reprinted by permission.

[3] Copyright, 1991, Cambridge Center for Behavioral Studies. Reprinted by permission.

during the contest, took questions from the press by speakerphone after the announcement was made. An entry from Peter Colby and Kenneth Colby, of PARRY fame, scored just below Weintraub's. In third place among the computers was a submission from Michael Mauldin of the Center for Machine Translation at Carnegie-Mellon.

The surprises were notable. First, five of the ten judges (Judges 2, 3, 4, 9, and 10) mistook Weintraub's program for a human. Judge 3 rated it above one human confederate (Terminal 1), and Judge 10 rated it above both human confederates (Terminal 1 and 4). The Colbys' program ("Problems in Romantic Relationships") was mistaken for a person by two of the judges, and another program ("Second Grade School Topics") was misclassified by one judge. Perhaps even more remarkable, Cynthia Clay, the human confederate whose topic was Shakespeare's plays, was mistaken for a computer by three judges. Judge 10 placed her lower in rank than a computer (Terminal 5), and two judges (1 and 5) placed her lower in rank than two computers. Note that Cynthia's responses were buffered; that is, her responses occurred in a burst, suggesting computer output. Furthermore, she was an expert on Shakespeare, and she often quoted lengthy passages verbatim. Several judges remarked that her replies seemed too expert to be human.

As Turing anticipated, the contest tells us as much, or perhaps even more, about our failings as judges as it does about the failings of computers. People's preconceptions about the limits of computers – and of people – strongly biases their judgments.

At the start of the contest, members of the audience were given forms to help them do their own judging. The forms asked for basic demographic information, as well. Seventy-seven forms were collected at the end of the contest. Based on this sample, audience ratings may be summarized as follows:

- Audience rankings matched those of the judges, and the rankings of those who claimed expertise in computer science did not differ substantially from the rankings of those who did not claim such expertise. For the 66 respondents who ranked all eight terminals, Terminals 1 and 4 were again ranked first and second, respectively, and Terminal 5 ("Whimsical Conversation") was again ranked third. Curiously, the other five terminals were ranked equally; that is, they were, on average, indiscriminable.
- Unlike the judges, members of the audience rarely misclassified the terminals, perhaps because members of the audience could communicate with each other; judges could not. For example, the winning computer, "Whimsical Conversation", was labeled a human by only five out of the 77 respondents (ten did not reply, leaving 61 correct classifications), and Cynthia Clay (Terminal 4) was misclassified as a computer by only five respondents (seven did not reply, leaving 65 correct classifications). The other human confederate, although ranked higher by both judges and audience, was misclassified at nearly the same rate. Once again, expertise in computer science had so systematic effect.

With James C. Pabelico, a student at the University of California, San Diego, I attempted a search for objective factors that could predict the judges' ratings – in other words, that measured the apparent intelligence of an entity communicating

over a computer terminal. Simplistic factors such as word length, sentence length, number of syllables per word, and number of prepositions were not predictive. Neither were various measures of readability, such as Flesch Reading Ease, Gunning's Fog Index, and Flesch-Kincaid Grade Level. The Weintraub and Colby programs, for example, had Flesch-Kincaid Grade Levels of 2 and 6, respectively; the two humans had scores of 3 and 4.

So why did Weintraub's program win? And how did it fool half the judges into thinking it was a person? Unfortunately, it may have won for the wrong reasons. It was the only program, first of all, that simulated human typing foibles well. Another program simulated human typing so poorly that it was instantly recognizable as a computer on that basis alone; no human could possibly have typed the way it was typing.

Perhaps more notable, Weintraub's program simulated a very curious kind of person: the jester. We allow great latitude when conversing with jesters; incomprehensible, irrelevant responses are to be expected. We are equally tolerant of young children, developmentally disabled individuals, psychotic patients, head-injured individuals, and absentminded professors. Weintraub's program may have succeeded simply because his terminal was labeled "whimsical conversation". The prize committee discussed this possibility, and considerable concern was expressed. In 1992, the committee favored programs that had clear subject matters.

1.3 Speculations

I believe that when a computer passes an unrestricted Turing Test, humankind will be changed forever. From that day on, computers will be companions to the human race – and extraordinary companions indeed. For starters, they will be efficient, fast, natural-language interfaces to virtually all knowledge. They will be able to access and evaluate enormous amounts of data on an ongoing basis and to discuss the results with us in terms we can understand. They will think efficiently 24 h a day, and they will have more patience than any saint.

Thinking computers will also have new roles to play in real-time control. Everything from vacuum cleaners to power plants has a dumb computer in it these days; some day, smart computers will share in the decision-making. Over networks or even airwaves, thinking computers will be able to coordinate events worldwide in a way humans never could.

Thinking computers will be a new race, a sentient companion to our own. When a computer finally passes the Turing Test, will we have the right to turn it off? Who should get the prize money – the programmer or the computer? Can we say that such a machine is "self-aware"? Should we give it the right to vote? Should it pay taxes? If you doubt the significance of these issues, consider the possibility that someday soon *you will have to argue them with a computer*. If you refuse to talk to the machine, you will be like the judges in *Planet of the Apes* who refused to allow the talking human to speak in court because, according to the religious dogma of the planet, humans were incapable of speech.