FIRST ANCHOR BOOKS EDITION, FEBRUARY 2012

Portions of this work were previously published in *The Atlantic*.

Grateful acknowledgment is made to Richard Wilbur for permission to reprint a portion of "The Beautiful Changes"

The Library of Congress has cataloged the Doubleday edition as follows:
Christian, Brian, 1984–
The most human human / Brian Christian.
p. cm.
1. Philosophical anthropology. 2. Human beings. 3. Turing test. I. Title.
BD450.C5356 2011
128—dc22
2010048572

Anchor ISBN: 978-0-307-47670-8

Book design by Michael Collica

www.anchorbooks.com

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

0. Prologue

Claude Shannon, artificial intelligence pioneer and founder of information theory, met his wife, Mary Elizabeth, at work. This was Bell Labs in Murray Hill, New Jersey, the early 1940s. He was an engineer, working on wartime cryptography and signal transmission.

She was a computer.

1. Introduction: The Most Human Human

I wake up five thousand miles from home in a hotel room with no shower: for the first time in fifteen years, I take a bath. I eat, as is traditional, some slightly ominous-looking tomatoes, some baked beans, and four halves of white toast that come on a tiny metal rack, shelved vertically, like books. Then I step out into the salty air and walk the coastline of the country that invented my language, despite my not being able to understand a good portion of the signs I pass on my way—LET AGREED, one says, prominently, in large print, and it means nothing to me.

I pause, and stare dumbly at the sea for a moment, parsing and reparsing the sign in my head. Normally these kinds of linguistic curiosities and cultural gaps interest and intrigue me; today, though, they are mostly a cause for concern. In the next two hours I will sit down at a computer and have a series of five-minute instant-message chats with several strangers. At the other end of these chats will be a psychologist, a linguist, a computer scientist, and the host of a popular British technology show. Together they form a judging panel, and my goal in these conversations is one of the strangest things I've ever been asked to do.

I must convince them that I'm human.

Fortunately, I *am* human; unfortunately, it's not clear how much that will help.

The Turing Test

Each year, the artificial intelligence (AI) community convenes for the field's most anticipated and controversial annual event—a competition called the Turing test. The test is named for British mathematician Alan Turing, one of the founders of computer science, who in 1950 attempted to answer one of the field's earliest questions: *Can machines think?* That is, would it ever be possible to construct a computer so sophisticated that it could actually be said to be thinking, to be intelligent, to have a mind? And if indeed there were, someday, such a machine: How would we know?

Instead of debating this question on purely theoretical grounds, Turing proposed an experiment. A panel of judges poses questions by computer terminal to a pair of unseen correspondents, one a human "confederate," the other a computer program, and attempts to discern which is which. There are no restrictions on what can be said: the dialogue can range from small talk to the facts of the world (e.g., how many legs ants have, what country Paris is in) to celebrity gossip and heavy-duty philosophy—the whole gamut of human conversation. Turing predicted that by the year 2000, computers would be able to fool 30 percent of human judges after five minutes of conversation, and that as a result "one will be able to speak of machines thinking without expecting to be contradicted."

Turing's prediction has not come to pass; at the 2008 contest, however, held in Reading, England, the top program came up shy of that mark by just a single vote. The 2009 test in Brighton could be the decisive one.

And I am participating in it, as one of four human confederates going head-to-head (head-to-motherboard?) against the top AI programs. In each of several rounds, I, along with the other confederates, will be paired off with an AI program and a judge—and will have the task of convincing the latter that I am, in fact, human.

The judge will talk to one of us for five minutes, then the other,

and then has ten minutes to reflect and make his choice about which one of us he believes is the human. Judges will also note, on a sliding scale, their confidence in this judgment—this is used in part as a tie-breaking measure. The program that receives the highest share of votes and confidence from the judges each year (regardless of whether it "passes the Turing test" by fooling 30 percent of them) is awarded the "Most Human Computer" title. It is this title that the research teams are all gunning for, the one that the money awards, the one with which the organizers and spectators are principally concerned. But there is also, intriguingly, another title, one given to the *confederate* who elicited the greatest number of votes and greatest confidence from the judges: the "Most Human Human" award.

One of the first winners, in 1994, was *Wired* columnist Charles Platt. How'd he do it? By "being moody, irritable, and obnoxious," he says—which strikes me as not only hilarious and bleak but also, in some deeper sense, a call to arms: How, in fact, do we be the most human humans we can be—not only under the constraints of the test, but in life?

Joining the Confederacy

The sponsor and organizer of the Turing test (this particular incarnation of which is known as the Loebner Prize) is a colorful and somewhat curious figure: plastic roll-up portable disco dance floor baron Hugh Loebner. When asked his motives for backing and orchestrating this annual Turing test, Loebner cites *laziness*, of all things: his utopian future, apparently, is one in which unemployment rates are nearly 100 percent and virtually all of human endeavor and industry is outsourced to intelligent machines. I must say, this vision of the future makes me feel little but despair, and I have my own, quite different ideas about what an AI-populated world would look like and reasons for participating in the test. But in any event, the central question of how computers are reshaping our sense of self, and what the ramifications of that process will be, is clearly the crucial one.

Not entirely sure how to go about becoming a confederate, I started at the top: by trying to reach Hugh Loebner himself. I quickly found his website, where, amid a fairly inscrutable amalgam of material about crowd-control stanchions,[1] sex-work activism,[2] and a scandal involving the composition of Olympic medals,[3] I was able to find information on his eponymous prize, along with his email address. He replied by giving me the name of Philip Jackson, a professor at the University of Surrey, who is the one in charge of the logistics for this year's Loebner Prize contest in Brighton, where it will be held under the auspices of the 2009 Interspeech conference on speech and communication science.

I was able to get in touch via Skype with Professor Jackson, a young, smart guy with the distinct brand of harried enthusiasm that characterizes an overworked but fresh-faced academic. That and his charming Briticisms, like pronouncing "skeletal" so it'd rhyme with "a beetle": I liked him immediately.

He asked me about myself, and I explained that I'm a nonfiction writer of science and philosophy, specifically of the ways in which science and philosophy intersect with daily life, and that I'm fascinated by the idea of the Turing test and of the "Most Human Human." For one, there's a romantic notion as a confederate of *defending the human race*, à la Garry Kasparov vs. Deep Blue—and soon, Ken

---

1. Crowd-control stanchions seem to have recently replaced portable disco dance floors as the flagship product of Loebner's company, Crown Industries, which is the Loebner Prize's chief sponsor.

2. Surely I'm not the only one who finds it ironic that a man who's committed himself to advancing the progress of interaction with *artificial* entities has resigned himself—as he has discussed openly in the pages of the *New York Times* and on several television talk shows—to paying, whether happily or unhappily, for *human* intimacy?

3. Apparently the "gold" medals are actually silver medals *dipped in gold*—which is, admittedly, a bit bizarre, although it seems to have caused Loebner more than a decade of outrage, which over the years has vented itself in the form of picketing, speeches, and a newsletter called *Pants on Fire News*.

Jennings of *Jeopardy!* fame vs. the latest IBM system, Watson. (The mind also leaps to other, more *Terminator*– and *The Matrix*–type fantasies, although the Turing test promises to involve *significantly* fewer machine guns.) When I read that the machines came up shy of passing the 2008 test by just one single vote, and realized that 2009 might be the year they finally cross the threshold, a steely voice inside me rose up seemingly out of nowhere. *Not on my watch.*

More than this, though, the test raises a number of questions, exciting as well as troubling, at the intersection of computer science, cognitive science, philosophy, and daily life. As someone who has studied and written about each of these areas, and who has published peer-reviewed cognitive science research, I find the Turing test particularly compelling for the way it manages to draw from and connect them all. As we chatted, I told Professor Jackson that I thought I might have something rather unique to bring to the Loebner Prize, in terms of both the actual performance of being a confederate and relating that experience, along with the broader questions and issues raised by the test, to a large audience—which would start what I think could be a fascinating and important conversation in the public culture at large. It wasn't hard to get him to agree, and soon my name was on the confederate roster.

After briefing me a bit on the logistics of the competition, he gave me the advice I had heard from confederates past to expect: "There's not much more you need to know, really. You *are* human, so just be yourself."

"Just be yourself"—this has been, in effect, the confederate motto since the first Loebner Prize in 1991, but seems to me like a somewhat naive overconfidence in human instincts—or at worst, fixing the fight. The AI programs we go up against are often the result of decades of work—then again, so are we. But the AI research teams have huge databases of test runs of their programs, and they've done statistical analysis on these archives: they know how to deftly guide the conversation away from their shortcomings and toward their strengths, what conversational routes lead to deep exchange and which ones

fizzle—the average confederate off the street's instincts aren't likely to be so good. This is a strange and deeply interesting point, of which the perennial demand in our society for conversation, public speaking, and dating coaches is ample proof. The transcripts from the 2008 contest show the judges being downright apologetic to the human confederates that they can't make better conversation—"i feel sorry for the [confederates], i reckon they must be getting a bit bored talking about the weather," one says, and another offers, meekly, "sorry for being so banal"—meanwhile, the computer in the other window is apparently charming the pants off the judge, who in no time at all is gushing lol's and :P's. We can do better.

So, I must say, my intention from the start was to be as thoroughly disobedient to the organizers' advice to "just show up at Brighton in September and 'be myself'" as possible—spending the months leading up to the test gathering as much information, preparation, and experience as possible and coming to Brighton ready to give it everything I had.

Ordinarily, there wouldn't be very much odd about this notion at all, of course—we train and prepare for tennis competitions, spelling bees, standardized tests, and the like. But given that the Turing test is meant to evaluate *how human* I am, the implication seems to be that being human (and being oneself) is about more than simply showing up. I contend that it is. What exactly that "more" entails will be a main focus of this book—and the answers found along the way will be applicable to a lot more in life than just the Turing test.

Falling for Ivana

A rather strange, and more than slightly ironic, cautionary tale: Dr. Robert Epstein, UCSD psychologist, editor of the scientific volume *Parsing the Turing Test,* and co-founder, with Hugh Loebner, of the Loebner Prize, subscribed to an online dating service in the winter of 2007. He began writing long letters to a Russian woman named

Ivana, who would respond with long letters of her own, describing her family, her daily life, and her growing feelings for Epstein. Eventually, though, something didn't feel right; long story short, Epstein ultimately realized that he'd been exchanging lengthy love letters for *over four months* with—you guessed it—a computer program. Poor guy: it wasn't enough that web-ruffians spam his email box every day, now they have to spam his heart?

On the one hand, I want to simply sit back and laugh at the guy—he *founded* the Loebner Prize, for Christ's sake! What a chump! Then again, I'm also sympathetic: the unavoidable presence of spam in the twenty-first century not only clogs the inboxes and bandwidth of the world (roughly 97 percent of *all email messages* are spam—we are talking tens of billions a day; you could *literally* power a small nation[4] with the amount of electricity it takes to process the world's daily spam), but does something arguably worse—it erodes our sense of trust. I hate that when I get messages from my friends I have to expend at least a modicum of energy, at least for the first few sentences, deciding whether it's really *them* writing. We go through digital life, in the twenty-first century, with our guards up. All communication is a Turing test. All communication is suspect.

That's the pessimistic version, and here's the optimistic one. I'll bet that Epstein learned a lesson, and I'll bet that lesson was a lot more complicated and subtle than "trying to start an online relationship with someone from Nizhny Novgorod was a dumb idea." I'd like to think, at least, that he's going to have a lot of thinking to do about why it took him four months to realize that there was no actual exchange occurring between him and "Ivana," and that in the future he'll be quicker to the real-human-exchange draw. And that his *next* girlfriend, who hopefully not only is a bona fide *Homo sapiens* but also lives fewer than eleven time zones away, may have "Ivana," in a weird way, to thank.

---

4. Say, Ireland.

The Illegitimacy of the Figurative

When Claude Shannon met Betty at Bell Labs in the 1940s, she was indeed a computer. If this sounds odd to us in any way, it's worth knowing that nothing at all seemed odd about it to them. Nor to their co-workers: to their Bell Labs colleagues their romance was a perfectly normal one, typical even. Engineers and computers wooed all the time.

It was Alan Turing's 1950 paper "Computing Machinery and Intelligence" that launched the field of AI as we know it and ignited the conversation and controversy over the Turing test (or the "Imitation Game," as Turing initially called it) that has continued to this day—but modern "computers" are nothing like the "computers" of Turing's time. In the early twentieth century, before a "computer" was one of the digital processing devices that so proliferate in our twenty-first-century lives—in our offices, in our homes, in our cars, and, increasingly, in our pockets—it was something else: a job description.

From the mid-eighteenth century onward, computers, frequently women, were on the payrolls of corporations, engineering firms, and universities, performing calculations and doing numerical analysis, sometimes with the use of a rudimentary calculator. These original, human computers were behind the calculations for everything from the first accurate predictions for the return of Halley's comet—early proof of Newton's theory of gravity, which had only been checked against planetary orbits before—to the Manhattan Project, where Nobel laureate physicist Richard Feynman oversaw a group of human computers at Los Alamos.

It's amazing to look back at some of the earliest papers in computer science, to see the authors attempting to explain, for the first time, what exactly these new contraptions were. Turing's paper, for instance, describes the unheard-of "digital computer" by making analogies

to a *human* computer: "The idea behind digital computers may be explained by saying that these machines are intended to carry out any operations which could be done by a human computer." Of course in the decades to come we know that the quotation marks migrated, and now it is the digital computer that is not only the default term, but the *literal* one. And it is the *human* "computer" that is relegated to the illegitimacy of the figurative. In the mid-twentieth century, a piece of cutting-edge mathematical gadgetry was "like a computer." In the twenty-first century, it is the *human* math whiz that is "like a computer." An odd twist: we're *like* the thing that used to be *like* us. We imitate our old imitators, one of the strange reversals of fortune in the long saga of human uniqueness.

The Sentence

Harvard psychologist Daniel Gilbert says that every psychologist must, at some point in his or her career, write a version of "The Sentence." Specifically, The Sentence reads like this: "The human being is the only animal that _____." Indeed, it seems that philosophers, psychologists, and scientists have been writing and rewriting this sentence since the beginning of recorded history. The story of humans' sense of self is, you might say, the story of failed, debunked versions of The Sentence. Except now it's not just the animals that we're worried about.

We once thought humans were unique for having a language with syntactical rules, but this isn't so;[5] we once thought humans were

---

5. Michael Gazzaniga, in *Human,* quotes Great Ape Trust primatologist Sue Savage-Rumbaugh: "First the linguists said we had to get our animals to use signs in a symbolic way if we wanted to say they learned language. OK, we did that, and then they said, 'No, that's not language, because you don't have syntax.' So we proved our apes could produce some combinations of signs, but the linguists said that wasn't enough syntax, or the right syntax. They'll never agree that we've done enough."

unique for using tools, but this isn't so;[6] we once thought humans were unique for being able to do mathematics, and now we can barely imagine being able to do what our calculators can.

There are several components to charting the evolution of The Sentence. One is a historical look at how various developments—in our knowledge of the world as well as our technical capabilities—have altered its formulations over time. From there, we can look at how these different theories have shaped humankind's sense of its own identity. For instance, are artists more valuable to us than they were before we discovered how difficult art is for computers?

Last, we might ask ourselves: Is it appropriate to allow our definition of our own uniqueness to be, in some sense, *reactionary* to the advancing front of technology? And why is it that we are so compelled to feel unique in the first place?

"Sometimes it seems," says Douglas Hofstadter, "as though each new step towards AI, rather than producing something which everyone agrees is real intelligence, merely reveals what real intelligence is *not*." While at first this seems a consoling position—one that keeps our unique claim to thought intact—it does bear the uncomfortable appearance of a gradual retreat, the mental image being that of a medieval army withdrawing from the castle to the keep. But the retreat can't continue indefinitely. Consider: if *everything* of which we regarded "thinking" to be a hallmark turns out not to involve it, then . . . what is thinking? It would seem to reduce to either an

---

6. Octopuses, for instance, were discovered in 2009 to use coconut shells as "body armor." The abstract of the paper that broke the news tells the story of our ever-eroding claim to uniqueness: "Originally regarded as a defining feature of our species, tool-use behaviours have subsequently been revealed in other primates and a growing spectrum of mammals and birds. Among invertebrates, however, the acquisition of items that are deployed later has not previously been reported. We repeatedly observed soft-sediment dwelling octopuses carrying around coconut shell halves, assembling them as a shelter only when needed."

epiphenomenon—a kind of "exhaust" thrown off by the brain—or, worse, an illusion.

Where is the keep of our *selfhood*?

The story of the twenty-first century will be, in part, the story of the drawing and redrawing of these battle lines, the story of *Homo sapiens* trying to stake a claim on shifting ground, flanked on both sides by beast and machine, pinned between meat and math.

And here's a crucial, related question: Is this retreat a good thing or a bad thing? For instance, does the fact that computers are so good at mathematics in some sense *take away* an arena of human activity, or does it *free* us from having to do a nonhuman activity, liberating us into a more human life? The latter view would seem to be the more appealing, but it starts to seem less so if we can imagine a point in the future where the number of "human activities" left to be "liberated" into has grown uncomfortably small. What then?

Inverting the Turing Test

> *There* are *no broader philosophical implications . . .*
> *It doesn't connect to or illuminate anything.*
>
> —NOAM CHOMSKY, IN AN EMAIL TO THE AUTHOR

Alan Turing proposed his test as a way to measure the progress of technology, but it just as easily presents us a way to measure our *own*. Oxford philosopher John Lucas says, for instance, that if we fail to prevent the machines from passing the Turing test, it will be "not because machines are so intelligent, but because humans, many of them at least, are so wooden."

Here's the thing: beyond its use as a technological benchmark, beyond even the philosophical, biological, and moral questions it poses, the Turing test is, at bottom, about the act of communication. I see its deepest questions as practical ones: How do we connect meaningfully with each other, as meaningfully as possible, within the limits

of language and time? How does empathy work? What is the process by which someone comes into our life and comes to mean something to us? These, to me, are the test's most central questions—the most central questions of being human.

Part of what's fascinating about studying the programs that have done well at the Turing test is that it is a (frankly, sobering) study of how conversation can work in the total absence of emotional intimacy. A look at the transcripts of Turing tests past is in some sense a tour of the various ways in which we demur, dodge the question, lighten the mood, change the subject, distract, burn time: what shouldn't pass as real conversation at the Turing test probably shouldn't be allowed to pass as real human conversation, either.

There are a number of books written about the technical side of the Turing test: for instance, how to cleverly design Turing test programs—called chatterbots, chatbots, or just bots. In fact, almost everything written at a practical level about the Turing test is about how to make good bots, with a small remaining fraction about how to be a good judge. But nowhere do you read how to be a good confederate. I find this odd, since the confederate side, it seems to me, is where the stakes are highest, and where the answers ramify the furthest.

Know thine enemy better than one knows thyself, Sun Tzu tells us in *The Art of War.* In the case of the Turing test, knowing our enemy actually *becomes* a way of knowing ourselves. So we will, indeed, have a look at how some of these bots are constructed, and at some of the basic principles and most important results in theoretical computer science, but always with our eye to the human side of the equation.

In a sense, this is a book about artificial intelligence, the story of its history and of my own personal involvement, in my own small way, in that history. But at the core, it's a book about living life.

We can think of computers, which take an increasingly central role in our lives, as nemeses: a force like *Terminator*'s Skynet, or *The Matrix*'s Matrix, bent on our destruction, just as we should be bent on theirs. But I prefer, for a number of reasons, the notion of

rivals—who only ostensibly want to win, and who know that competition's main purpose is to raise the level of the game. All rivals are symbiotes. They need each other. They keep each other honest. They make each other better. The story of the progression of technology doesn't have to be a dehumanizing or dispiriting one. Quite, as you will see, the contrary.

In the months before the test, I did everything I could to prepare, researching and talking with experts in various areas that related back to the central questions of (a) how I could give the "most human" performance possible in Brighton, and (b) what, in fact, it means to be human. I interviewed linguists, information theorists, psychologists, lawyers, and philosophers, among others; these conversations provided both practical advice for the competition and opportunities to look at how the Turing test (with its concomitant questions of humanhood) affects and is affected by such far-flung fields as work, school, chess, dating, video games, psychiatry, and the law.

The final test, for me, was to give the most uniquely human performance I could in Brighton, to attempt a successful defense against the machines passing the test, and to take a run at bringing home the coveted, if bizarre, Most Human Human prize—but the ultimate question, of course, became what it *means* to be human: what the Turing test can teach us about ourselves.