YOUR WIT IS MY COMMAND

YOUR WIT IS MY COMMAND

Building AIs with a Sense of Humor

TONY VEALE

The MIT Press
Cambridge, Massachusetts
London, England

—-1
—0
—+1

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Adobe Garamond Pro and Berthold Akzidenz Grotesk by Westchester Publishing Services. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Names: Veale, Tony, 1967- author.
Title: Your wit is my command : building AIs with a sense of humor / Tony Veale.
Description: Cambridge : The MIT Press, 2021. | Includes bibliographical references and index.
Identifiers: LCCN 2020045150 | ISBN 9780262045995 (hardcover)
Subjects: LCSH: Artificial intelligence--Philosophy. | Artificial intelligence--Social aspects. | Artificial intelligence--Humor.
Classification: LCC Q335 .V44 2021 | DDC 006.301--dc23
LC record available at https://lccn.loc.gov/2020045150

10 9 8 7 6 5 4 3 2 1

-1—
0—
+1—

For Hyesook, my funny bone

Contents

—-1
—0
—+1

-1—
0—
+1—

Foreword

Charlie Skelton

When I'm not out roaming the woods in search of a hollow tree to scream into, I sometimes write jokes for people to say into cameras. Most of the jokes I write don't get anywhere near the autocue; they languish on the page, unloved like a fallen soufflé or an out-of-focus photograph. The other day I was writing on the comedy quiz show *Have I Got News for You* (*HIGNFY*), and they wanted a joke about the mysterious steel monolith that had been spotted in the deserts of Utah by a helicopter pilot who was out counting bighorn sheep. I supplied ten. None made the grade. I had high hopes for one about how dangerous it is to be counting sheep while flying a helicopter, but no joy. A note came through from the producer: "more of an extra-terrestrial/apocalyptic/end of the world angle might be the way with the monolith." With that in mind, I churned out another clutch of mysterious steel monolith jokes, one of which made the show: "One theory popular on Twitter is that the monolith was left by passing aliens as a message to humanity. That message being: stop looking at Twitter."

Though the story itself is peculiar, this is actually a fairly typical bit of by-the-numbers comedy writing. A very specific set-up with certain extractable features: Utah, monolith, helicopter, desert, mystery, steel. Around these swirl various cultural resonances, such as Kubrick's *2001* and Area 51, and a few more abstract or compound ideas such as "finding things," "things in deserts," and "alien contact." This is the primeval proteinaceous soup of comedy that requires a lightning strike of inspiration, and then out crawls a joke. Nine times out of ten, the joke expires feebly on the edge of the

pond. But once in a while one survives, grows a pair of leathery wings, and screeches across the airwaves and away into the aether.

Part of this task is undoubtedly computational: extracting and categorizing data, combining variables, inverting values. Without these mechanics, however subtle, there'd be no comedy at all. There's no ridiculousness without logic. What Tony Veale does so well is to uncover these mechanisms, dissecting the frog of creativity and spreading out its guts for contemplation, meticulously setting out the mathematics of wit. In practice, better jokes manage to hide their workings—or are so funny that you're too busy laughing to notice the algebra.

My (slightly clunky) monolith joke about an alien message and Twitter relies, to some extent, on *foregrounding* its structure—the way the punchline reverses the word order of the set-up: Twitter/message; message/Twitter. Fans of classical rhetoric will spot this ABBA structure as an example of "antimetabole." Not that I was sitting there at 4 p.m. thinking "maybe I should try a bit of antimetabole." I was just grabbing desperately at the logic of the set-up and trying to grapple it around into a joke shape. It was written by rote. The grappling was 90 percent muscle memory.

Dimly behind the process lurked an awareness that a *HIGNFY* audience would probably enjoy a slightly snobby pop at social media, but again, this, like most of what went into dragging that particular gag out from the soup, was subconscious. Boldly into this unconsidered gloom strides Tony Veale, shining a fascinating light on the weird gears and pistons that are whirring and cranking behind every bit of linguistic creativity.

And it's the understanding of this engineering that's getting AI ever closer to a seat in the writers' room. Of course, by the time AI gets into the swing of writing jokes it'll be churning out fifteen billion Utah monolith gags before its first cup of coffee, at which point, blessedly, I'll be out of a job. Comedy will morph into some kind of transcendent incomprehensibility, AI will be off amusing itself by lasering hypercompressed nanojokes off mirrors placed all around the solar system, and I'll have my head inside a hollow trunk amazing myself at the acoustics. It'll be great. And by the way, anyone who thinks robots will never be able to replace comedy writers is severely underestimating the ability of scientists to produce a machine that can eat three breakfasts.

-1—
0—
+1—

Preface

An oppressive government aiming to crack down on political satire and humorous dissent could do worse than to establish a ministry to explain its opponents' jokes. Explanations are to jokes what autopsies are to bodies: if the subject isn't already dead, it soon will be. As a result, some scholars work with jokes so old that they may as well be unwrapping mummies. I try not to overexplain jokes in this book, since a joke that needs to be explained is hardly worth telling, much less studying. Still, there is a crucial difference between the kind of jokesplaining that can turn a specific joke inside out, to rob it of its timing and its humorous payload, and the deeper kind of analysis that sheds light on how jokes of all stripes might actually work. It is the latter kind that I aim for here, when I use the tools of artificial intelligence (AI) to lay bare the demands that jokes make on our models of language and the world, so that we can give our machines something like those models too.

A joke has a great deal in common with a magic trick. Typically, each plays at the fault lines of common sense; relies heavily on timing, misdirection, and patter; and is more joyful in its delivery than any after-the-fact reveal could ever be. It is hardly surprising that the secret of a trick fails to live up to its execution—What could ever compete with magic?—but the same deflationary feeling of "Is that all there is?" when our curiosity is sated and the magic goes away can also follow the explanation of a joke. However, the right kind of analysis can also surprise us, and perhaps delight us in a different kind of way, by revealing the depth of knowledge

and hidden complexity that lurk beneath the surface of the joke. Jokes are mental playthings, and when we open them up, we find not a cheap trick but an intricate clockwork of cogs and springs. AI provides tools and techniques to replicate these counterbalanced forces in our machines, while at the same time giving us a firmer grasp of what is going on in our own heads when we create, tell, or laugh at a joke. Some kinds of explanations make the strange seem ordinary and straightforward, but others reveal the strange forces that make the commonplace seem so natural.

This is not to say that an AI understanding of humor isn't a reductive one. It is, not least because the problem is so much bigger than our capacity for a solution. A phenomenon as sprawling and amorphous as humor, one that touches on so many aspects of our lives, is not going to be squeezed into a single formula or equation. But AI allows us to chip away at different manifestations of humor, to sense irony in online reviews or sarcasm in tweets, to find puns here and generate them there, to make headlines more or less witty, to rank cartoon captions by their potential to make us laugh, to invent comedic shaggy-dog tales, to separate jokes from nonjokes, or—in the most ambitious scenarios—to create entirely new jokes that play not just on words but on ideas too. If the human sense of humor is an ice sculpture of a majestic swan, then what AI gives us—for now, at least—is a bag of ice cubes.

This is an excellent start if you just want ice in your drink or occasional flashes of wit in your favorite applications, but engineers will have to perform a reverse-Humpty to put all the pieces back together as one genuinely humorous AI system. When they do, they will use the different approaches we explore in this book, from symbolic ontologies, frames, and semantic networks to statistical language models and artificial neural networks that are trained at Web scale. As for what Humpty will look like when all of the different pieces work together as one, I paint a variety of scenarios (or what software engineers call *use cases*) in chapter 1.

The philosophy of AI allows us to turn mysteries into problems, while the tools of AI allow us to turn those problems into solutions. It may trouble you to think of humor as a "problem" to be solved, but this is just the way that AI works best. In fact, the AI approach to understanding aspects

of human creativity is reductive in other ways too. For instance, it encourages us to look at the parts of a problem that are most conducive to logical or statistical modeling and to leave everything else for another day or another researcher. Humor has been a subject of academic theorizing since Aristotle, and the AI treatment of jokes I pursue here is just one strand among many in the rich tapestry of humor research. Although I touch on many aspects of humor in this book, I go into depth on only the most AI-friendly of them, giving others a cursory treatment or a glancing look in passing. Aspects that deserve a fuller treatment, and are the focus of other books in their own right, include linguistic semantics, pragmatics, sociolinguistics and sociology, psychology and neuroscience, and the physiology of laughter and mirth. The text includes references to scholarly work on these topics where they are relevant, and I hope readers will tug on some of these threads to gain a richer appreciation of the phenomenon than an AI-oriented book like this one can offer.

The researcher Eugene Charniak summed up his life in AI with a telling joke as he received a lifetime achievement award from the Association for Computational Linguistics in 2011: if the second half of his career, his *statistical period* in artistic terms, is called "S," then the half before statistics can only be called "BS." Charniak has done inspiring work in both halves of his long career, but this joke reflects a seismic generational shift in how AI is studied and evaluated. From the birth of AI in the 1950s to the 1990s, AI was a largely symbolic enterprise that relied heavily on logical, handcrafted rules operating over formal representations of meaning. This resulted in systems that were elegant and imaginative but also rather brittle, since they failed to accommodate the blurred lines of the real world. I remember some of those systems with the same affection as the sci-fi shows of my youth and make the same allowances for their dodgy but well-meaning special effects. When AI took a data-driven turn to the statistical, it put the real world in the driving seat, making messiness and nuance the norm rather than the exception. These "S" systems are now defined as much by actual data as by their creator's imagination.

This book also divides neatly into *BS* and *S* halves, although readers are urged to read BS as just meaning "before statistics." If you insist on a double

meaning, try on "beautifully symbolic" for size. Symbolic approaches to AI can seem rather old school now, to the extent that many don't even see them as real AI anymore, but I hope to show that symbolic and statistical approaches are not natural antagonists. In fact, they complement rather than oppose each other. One of the benefits of the symbolic approach is its amenability to explanation, and this proves useful in the first five chapters of the book as we move from high concept. to theories of humor. to actual practice in comedy writing. Symbolic frameworks give a top-down shape to AI systems, and data-driven analyses capture the nuance and variability that we cannot box in with straight lines and hard rules. Once we have considered the perspective of comedy professionals in chapter 5, which proves to be a bridging point between these top-down and bottom-up perspectives, we will be ready to swap out more of our conceptual scaffolding for numerical models that are driven as much by end results as by prior expectations of how things are supposed to be. In any case, a math-wary reader won't encounter any equations until chapter 6.

The tension between symbolic and statistical AI reminds me of the relationship between Woody and Buzz in the movie *Toy Story*. Woody is an old-fashioned toy, quaint and unshowy, past his prime, and well on his way to has-been status. Buzz Lightyear is the new, new thing, shiny and sleek, and full of hi-tech swagger. Buzz can fly, or believes that he can, but Woody is just not buying it. As he puts it, "That's not flying, that's just falling with style." By movie's end, however, each has come to value the other's perspective. Woody is taken by Buzz's ability to ride the air currents, while Buzz now views his abilities in a more realistic light. Whenever we give a phenomenon like humor the AI treatment, we aim for flights of creative fancy but end up settling, like Woody and Buzz, for the practicalities of falling with style. This is true regardless of which brand of AI we use, and it is certainly true of the treatments we will cover in the chapters to come. Yet as we fall short of our ambitious goals, we will aim to fall in the right direction. *To infunity and beyond!*

Humor is a leavening agent that expands our range of possible reactions to a situation. It thrives on and creates ambiguity. An unexpected side effect of writing a book about humor and AI is that even mistakes

suggest new possibilities of their own. For instance, a reviewer of an earlier draft noted that it was not possible to tell whether a quirk of the text was a genuine error or an attempt at metahumor. Sadly, it was an error, one of many that reviewers have helped me to root out and fix. This book has benefited greatly from the feedback of anonymous readers like these who saw what I could not and from the ministrations of my editors at MIT Press. I especially thank Marie Lufkin Lee, who championed this project from the beginning, as well as Elizabeth Swayze and Alex Hoopes, who carefully guided it to completion. I also thank the comedy writer Charlie Skelton for his keen insights into humor and his thoughts on the text. Charlie and his partner, Hannah, have actively promoted a dialogue between AI researchers and humor practitioners through a series of multidisciplinary symposia on Comedy and AI at Oxford University, and these have allowed researchers like me to bounce ideas off and pick the brains of people who make a living from making others laugh. Finally, I thank my wife, Hyesook, who braved the many ups-and-downs of my catastrophe-theoretic moodscape as I wrote and revised this text. Naturally, any remaining errors you may find in these pages are mine and mine alone, unless you decide they are deliberate attempts at humor. In that case, they are all yours.

Dublin
August 2020

1 DOES NOT COMPUTE: WHY OUR MACHINES NEED A SENSE OF HUMOR

WE DON'T SERVE THEIR KIND HERE

What does it mean to have a good sense of humor? A GSOH, in the parlance of dating profiles and personal advertisements, seems to be a trait that is as eagerly sought as it is openly touted. If all of those ads and profiles are to be believed, we humans routinely rank our GSOHs as favorably as income, education, fitness, and physical appearance. So spare a thought for those who lack a GSOH, or who have no sense of humor at all. For those poor souls who are deaf to irony and immune to whimsy and who tilt their heads at italics, we often resort to the language of automation.

These rule-bound bureaucrats and literal-minded slaves of orthodoxy are seen as automata for whom jokes simply do not compute. A key trait of these reliable followers of rules is predictability. This allows comedians to instinctively feel that they know how their targets will react and to know that their audience will make the same calculation. So machines, and machine-like people, make the perfect foil for a comedian, for although we laugh at them for the rigidity of their calculations, our laughter is itself based on a rather predictable calculation about that rigidity. Since it is no small irony that this stereotypical view of machines is as rigid as the orthodoxy it is used to ridicule, I set out in this book to convince you that it ain't necessarily so. While there may be no master algorithm for anything as complex as a GSOH, the algorithmic nuts-and-bolts view is as valuable as any when it comes to understanding jokes from the inside out. Using ideas from artificial intelligence (AI) to shine a light on how we

humans engage with humor, I hope to show that machines are not inherently unfunny. They are just programmed that way.

If we're being generous, machines can already *tell* jokes, or at least recite the ones that we train them to tell. They can also recognize the telltale markers of humor in a text, using data-crunching algorithms that are increasingly sensitive to nuance. Even if they don't get the joke at a human level, they can still use data-driven insights to make laughter sounds at appropriate points in a conversation. However, whether you are dating, mating, or just playing around, a good sense of humor means much more than a readiness to laugh or tell jokes. Rather, a GSOH is indicative of a balanced personality that can bring flexibility, levity, and insight to situations that cause others to sulk or rage.[1] It indicates an ability to bend rather than snap, to fit in with others and adapt to their moods, and this is as good a reason as any to want to give our machines a human sense of humor. The 2014 film *Interstellar* introduces us to a robot named TARS that comes fitted with just this kind of social lubricant. Although TARS is a decidedly fictional AI, this trash-talking robot busts the myth of the humorless machine that is too stiff to be funny. Our AIs must do more than tell jokes, so we'll meet TARS again, and a host of his sci-fi brethren too, as we use the science *and* fiction of AI as a guide to building machines with many of the social and cognitive qualities of a real GSOH.

This chapter takes its title from a much-quoted line in the 1960s TV show *Lost In Space*.[2] A tinfoil and jumpsuit retelling of the Swiss Family Robinson, the show follows the space-age adventures of the Robinson family, adrift in space with only the devious stowaway Dr. Smith and the straitlaced robot B-9 for company. Smith and B-9 are the comedic heart of the show and the best reason to ever catch the reruns on TV. Their comic partnership recalls famous double acts from Laurel & Hardy and Abbott & Costello to George & Gracie and Martin & Lewis, with a pinch of Wile. E. Coyote and the Roadrunner thrown in for good measure. Together, B-9 and Smith are a yin-yang marriage of light and dark, with B-9 applying a rigidly benign brake to Smith's diabolical schemes. Ironically, while B-9 is the show's embodiment of mechanical predictability, Dr. Smith's

very human ego proves to be every bit as predictable and causes his selfish schemes to run aground with comic regularity.

B-9's cry of "It does not compute, It does not compute" has been shortened to "Does Not Compute!" in popular memory. The robot's outbursts are invariably directed at Dr. Smith, whose ploys fail in episode after episode because he insists on recruiting so inflexible an ally to help him bend the rules. Neither B-9 nor Smith possesses what many would call a GSOH, but together they are a comedic force to be reckoned with. B-9, a lugubrious lug wrench, acts as Smith's comic foil, allowing the doctor to vent his fury with colorful alliteration, from "Computerized Clod" and "Meandering Mental Midget" to "Rusty Rasputin" and "Tintinnabulating Tin Can." If a joke is a marriage of logic and emotion, then these two complete each other. While certain objects and situations are potent catalysts for humor, it is people we ultimately laugh with and laugh at. They might not be obvious, and they might be wholly imaginary—the protagonist of a shaggy dog tale, perhaps, or a crass ethnic stereotype—but other people are central to the humorous effect.

Imagine that you are home alone when a clock falls from the wall or a light bulb pops in its socket or an overstuffed chair makes a thunderous farting sound. Do you react with laughter or with a start? Now imagine the same thing happening when you are sitting with friends. How likely are you to laugh now, as you take in the look of surprise on their faces and they take in exactly the same look on yours? For a machine to be funny in itself, rather than just a catalyst for humor, we must be able to relate to it socially, to see it as something like another person or, failing that, to imagine it filtered through the minds of other people. We laugh at B-9 and Dr. Smith because we see them relating to each other as intelligent, social entities, and we can imagine how we might relate to them too.

From B-9's "It Does Not Compute!" to *Little Britain*'s "Computer Says No!"[3] catchphrases are humorous shorthands that index our feelings to words, allowing ourselves and others to recreate those feelings on demand. Think of the feelings audiences experience while watching the curmudgeonly Victor Meldrew in the British sitcom *One Foot in the Grave*.[4] The

—-1
—0
—+1

show turned Victor's howl of existential angst, "I don't *beLIEVE* IT!" into a catchphrase that is now synonymous with indignation and wide-eyed incredulity. Victor's mundane life is punctuated by minor outrages of fortune, from finding a severed pinkie in his fish and chips to finding a toupee in a loaf of bread. Each fresh incongruity is met with puffs of mounting rage, but the affronts to his senses become humorous only when he bellows his familiar, "I don't *beLIEVE* IT!" and we momentarily share his frustrations too. Victor's venting becomes an occasion of humorous catharsis, not because life's petty outrages are inherently funny but because his reaction to them is so familiar *and* so extreme.

This kind of expression momentarily opens an inward-looking window, rather than just an outward-blowing vent, into someone else's mind, allowing us to peer inside and see that others are just like us where it counts. Intense emotions make events seem more significant and can help us to lay down, and later recall, vivid memories of those events. By helping us to focus on what is important, emotional memories also help us to learn. As an example, let's look at a scene from the 1984 film *Terminator*.[5] A T-100 cyborg, played by Arnold Schwarzenegger, goes back in time to 1984, where he finds himself in need of clothes. Drawing the attention of street thugs, the naked T-100 demands of them: "Your clothes. Give them to me. Now." The actor Bill Paxton,[6] billed only as Second Punk, surprises no one with his reply: "Fuck you, asshole." Later in the film, the T-100 is performing some bloody self-repair in a seedy motel room when the manager shouts through his door: "Hey, buddy, you got a dead cat in there, or what?" A point-of-view shot reveals the cyborg's inner mental state, which resembles a menu on an Apple II computer, circa 1984 (figure 1.1).

So where did that second-to-last option come from? All of the others, including the hilariously out-of-character "please come back later," may be factory settings, but this particular one was clearly acquired during the T-100's encounter with street punks. Perhaps it was learned as a variation of the very last option, the no-frills "Fuck you," but in any case, this is an example of what AI researchers call *one-shot learning*.[7] A machine typically requires a large set of training examples to acquire a robust model of the categories it wishes to learn, so that it can discern cars from trucks, for

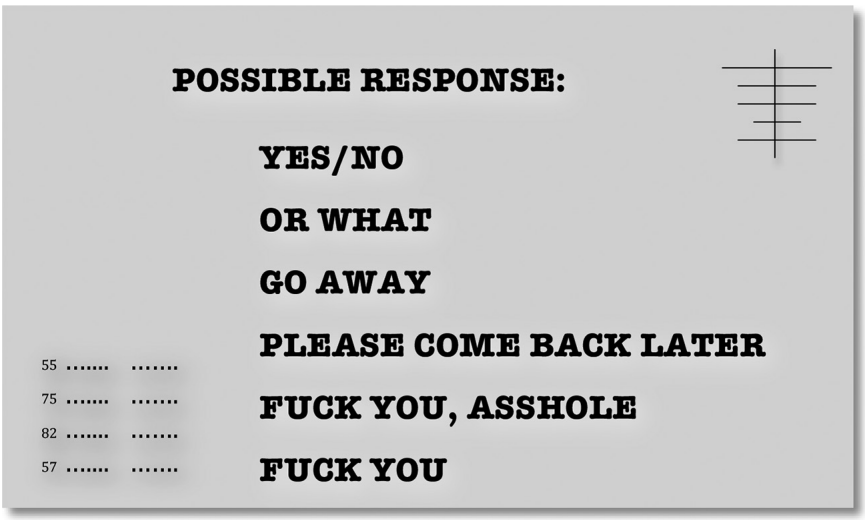

Figure 1.1
An AI from the future adopts the graphical interface norms of the 1980s.

example, or chickens from humans, or civilians from enemy combatants. One-shot learning presupposes that the new category is built from many of the same components as its preexisting categories, and so, rather like a human, a machine can acquire the necessary distinctions from a very small number of new examples.

But what prevents a machine from overlearning and building a new category from every new experience? In the case of the T-100, it is reasonable to assume that the machine has a strong sense of linguistic "sentiment"[8]—that is, a model of the likelihood that words and phrases convey a positive or a negative meaning—and attunes to the extreme negativity of Second Punk's emotional response. After all, directed negativity in language is often a marker of hostility, and a military robot like the T-100 will be programmed to recognize offensive intent in others. It is as if the cyborg has identified "Fuck you, asshole" as a useful catchphrase that works best in the fraught emotional contexts associated with hostile demands.

Humor theorists label recurring contexts like these as "scripts"[9] or "frames,"[10] depending on their preferred cognitive framework. The labels denote two sides of the same coin and provide complementary perspectives

—-1
—0
—+1

onto the same mental structures. For instance, what comes to mind when you think of the meaning of the word *medicine*? Do you think of hospitals, doctors and nurses, drugs and blood tests, and so on? All of these images cluster around the concept of *medicine* and help us to frame the concept in terms of how we relate to it in our own lives.[11] If simple ideas are LEGO bricks, then frames are the structures that we typically build from them. We can think of each frame as one of the thematic kits that LEGO produces, such as for a suburban house, a gas station, a castle, or a pirate ship.

Each kit has enough bricks to make the object pictured on the box and may also include some little LEGO people to populate the finished structure. Frames give cohesion to our ideas and provide a standardized tool kit for understanding each other's thoughts and feelings. Kids who love LEGO will have buckets of bricks from many different sets, and this is where the real fun starts. With a critical mass of building blocks, we can explore possibilities that transcend any one frame. We can build a fire station manned by wizards, or a hospital for pirates, or a cafeteria for Death Star workers (to riff on a standup routine by the comic Eddie Izzard[12]). This is what comedians do: they mix bricks from different kits to build something that the product guys at LEGO never anticipated but always knew was possible.

To appreciate a joke based on these common foundations, like a classic *doctor, doctor* joke, we need to know more than how the LEGO bricks click together. We need to know the order in which to assemble them and the order in which we generally experience them. This sequential aspect of our knowledge is captured in the conceptual equivalent of a script.[13] Think about the last time you visited the doctor. Perhaps you noticed some symptoms, phoned for an appointment, spoke to a nurse, waited in reception, spoke with the doctor, answered some questions, received a prescription, paid up, filled the prescription at a pharmacy, paid again, took your medicine, and eventually felt better. This sequence is hardly the stuff of a Hollywood blockbuster, but it's one we can all recognize as true to life, and one we can negotiate more or less on autopilot. We have scripts like this for a great many activities in our lives. They are so ingrained in our behavior that each script carries us along from one action to the next. Yet scripts do

more than help us to live our own lives; they also allow us to predict the events in the lives of others.

But life is neither a LEGO set nor a movie set. In every situation that we find ourselves in, we need to figure out the most relevant script to follow. Fortunately, the world gives us clues as to which one to activate at any given time, even if some clues are more obvious than others. Think of a typical McDonald's and how its design nudges you to follow the fast-food-ordering script. We refer to these clues as script *triggers*. Jokes exploit the fact that triggers are imperfect, and we can sometimes activate the wrong script for what seems like the right reasons. Suppose, for example, that in a fancy restaurant on Valentine's Day, you see a man go down on bended knee in front of his dining companion, so you activate the marriage proposal script. If it turns out that the man is looking for a lost contact lens, you will have triggered the wrong script. Comedy sprouts in the gap between the activation of a seemingly apt script and the realization of just how wrong we are. For a brief moment, we become aware of our own rigidity, and our mistake allows us to laugh at ourselves. The comic effect is heightened if other people and other minds enter into our calculations too. We might put ourselves into the mind of the man's dining companion, whose romantic hopes have suddenly been raised, or sympathize with the half-blind man as he arrives late to the same realization.

Early in childhood, typically between 18 and 36 months, most of us develop the intuition that other people have minds, too, and those minds are much like our own.[14] The intuition, named *theory of mind* (TOM) by philosophers of mind and by developmental psychologists, is crucial to understanding the actions of others, as it allows us to assume that other people are driven by the same kinds of beliefs, desires, and intentions.[15] The TOM intuition takes root early, at a stage in our lives before we have even acquired words like *mind*, and it evolves as we learn and develop socially. It is TOM that allows us to ascribe specific goals and feelings to the participants of a joke or a comedic situation and to predict how those feelings might change when the unexpected happens. So, TOM allows us to laugh *with* others, but it is just as instrumental in allowing us to laugh *at* others. It is our TOM that allows us to peer inside Victor Meldrew's mind

when he is visited by bafflement and indignity, to understand his motivations for bellowing, "I don't *beLIEVE* it!" and to laugh *with* him and *at* him for his overreaction. Victor's reaction confirms the prediction of our TOM, but since that prediction is based on knowledge of our own mental states, his catchphrase applies just as much to us.

In this sense, a comedy catchphrase is a special kind of script trigger, one that activates a script late in the game.[16] By time it is uttered, we will have processed the actions that the script would have helped us to understand, so the catchphrase triggers a script-based reappraisal of what has gone before rather than a helpful prediction of what is yet to come. Many humorous idioms work in this after-the-fact way. Think of the irony of, "So, that went well," or the understatement, "That could have gone better," or the inane innuendo, "That's what *she* said!" Most punch lines work in exactly the same way, urging listeners to reappraise the setup to a joke from the perspective of a different script. Humor is always a matter of perspective, which is why some of us get the joke while others just get upset. Scripts help us to construe unfolding events according to conventional wisdom, or to reconstrue what has just happened in a new and humorous light. We'll return to the twin concepts of frames and scripts throughout this book as we ponder where they come from and of how we can get them into our machines.

WHAT JUST HAPPENED?

We can joke about pretty much anything because our sense of humor touches on just about everything in our lives. Emotion, morality, empathy, logic, and common sense, all codified as frames and scripts—each of these things and more finds a common cause in tickling our funny bones. Humor is a cross-cutting sense that interacts with, and colors the judgments of, all five of our physical senses, leading some computational researchers to speculate that it is *AI-Complete*,[17] which is to say, as vexing as *any* other problem in human-scale AI and dependent on good working solutions to *every* other problem.

Whatever your view, humor is unlikely to arise in machines as a result of a quick fix or a happy accident in their code, so we researchers must be in this for long nights and the long haul. Let's begin our journey by exploring the ways that science fiction has found to endow computers with a sense of humor. While we are unlikely to find any solutions in the realms of speculative fiction, we may nonetheless find the outlines of a practical computational philosophy. Central to this philosophy is the question of modularity: Is humor an augmentation that can just be gifted to computers as a modular plug-in unit like Commander Data's "emotion chip" in *Star Trek: The Next Generation*, or is it an emergent enigma that only arises from the myriad interactions between all of the other stuff going on inside a thinking agent?[18] If the latter, might it emerge naturally within a complex AI system without ever having been designed to do so, as in the mischievous supercomputer Mike[19] in Robert Heinlein's novel *The Moon Is a Harsh Mistress*, or in the sarcastic droid K-2SO[20] in the movie *Rogue One: A Star Wars Story*?

Data is a lovable bundle of sci-fi clichés about human-like machines, brought to life with great charm by the actor Brent Spiner. He combines a child's precocity with the hyperlogicality of a calculator, or indeed a Vulcan. While he himself does not feel any emotions, he knows enough to reason about their effects on humans and the effect of the lack of them on himself. So, when he is presented with a plug-in emotion chip that can remedy this lack, he worries that the intense emotions he has so often observed in humans will overwhelm his artificial neural networks. At the same time, he is aware that his attempts to build a wholly logical sense of humor have all failed miserably. In *Star Trek Generations*, Data sees others laugh with glee as a crewmate is dunked in the cold, shark-infested water of a Holodeck simulation of walking the plank and finds the explanation of the ship's doctor, Beverly Crusher, to be as good a cue as any for one-shot learning: "It's all done in good fun, Data. Get in the spirit of things."[21] Inevitably, he is confused that no one laughs when he then drops the good doctor into the drink. His explanation reveals a GSOH-shaped hole in his logic: "I was attempting to . . . get in the spirit of things. I thought it would be humorous." At this, Data finally agrees to the new implant.

—-1
—0
—+1

Provided that an unpleasant experience is relatively benign, humor allows us to revel in a shared emotional response. When Data samples an exotic beverage with coworkers at the bar, his response is much more than physiological, and his implant now allows him to delight in the subjectivity of his opinions: "Yes. I hate this! It is revolting!" Naturally, he says "yes" to another round. Yet the pairing of a prodigious memory and a newfound sense of humor cannot but produce some unexpected results, and Data finds himself suddenly giggling at missed jokes from the distant past. We might imagine a humor "chip" as a generator of humorous possibilities that we can take or leave in any given situation, but a GSOH is a blend of wit *and* wisdom—the wit to perceive the possibility of humor and the wisdom to act on it when it is apt to do so. Data's chip gives him access to the possibility space of humor, but does not give him the wisdom to sample the space responsibly, and he is soon overwhelmed by the intense feelings afforded by his implant.

No direct account is offered for the sense of humor shown by the droid K-2SO in the film *Rogue One: A Star Wars Story*, but we can infer a plausible explanation from the robot's backstory. We are told that K-2SO, an imperial battle droid, was captured by the rebel alliance and reprogrammed to serve the rebels in their fight against his former masters. K-2SO is voiced by the actor Alan Tudyk as a chippy manservant who believes himself better than his new masters, and his ambivalent physical form—tall and stooped, barrel-chested and spindly limbed—makes him appear obsequious *and* threatening. He is a mix of Lurch from the Addams family and Jeeves from the tales of P. G. Wodehouse, and he is just as funny as this blend suggests. His humor makes him a natural, if snarky, teammate, and he frequently pokes fun at his human colleagues and their reliance on conventional wisdom. When we first meet K-2SO, he riffs on a standard human script to tell the plucky heroine, whom he holds in a choke grip, "Congratulations. You're being rescued," and later riffs on an idiomatic phrase (another kind of script) to note, "There is a problem on the horizon. There *is* no horizon." K-2SO was not programmed by the rebels to be humorous, and his no-nonsense imperial designers would scarcely build in such a capability. So where on earth did K-2SO acquire his sense of humor?

It's no stretch to see military robots as mechanized soldiers with an in-built aggression toward their designated enemy. We can imagine them rolling off the assembly lines with serious weapons skills and the tactical knowledge to exploit them in battle. Like any human soldier, each can be relied on to obey orders and to know its place in the chain of command. Such things are not built to philosophize or write poetry, so we can expect battle droids to possess just enough linguistic nuance to confirm or relay orders, explain their actions, and describe the current state of engagement to their superiors. As an imperial battle droid, K-2SO's imprinted enemy is the rebel alliance, and to the extent that droids need training, he would have been drilled in the art of fighting, capturing, and killing rebel scum. When K-2SO himself is captured and reprogrammed to switch sides, his new enemies become those who designed and built him. He must now obey the rebels he was designed to kill with a zeal approaching hatred. But such a turnaround is surely not achievable with a few localized changes to his code.

Oh, to get K-2SO on the psychiatrist's couch! He is not as profoundly conflicted as HAL 9000, the murderous supercomputer in Stanley Kubrick's *2001: A Space Odyssey*, but he comes close.[22] A complex AI is a many-layered thing, combining symbolic and nonsymbolic components. The former are logical and declarative, which means they can be read and understood much like a computer program or a recipe in broken English. A system's high-level axioms and edicts may be coded as symbolic rules that others can easily inspect and modify, and we might expect K-2SO's animus for members of the rebellion to be expressed here. But as we peel away more of the layers, we find that the situation becomes a good deal murkier.

Beneath the strata of meaningful symbols, we may encounter highly connected layers of numerical units that have no individual meaning in themselves. Rather, through repeated cycles of training on appropriate exemplars, the machine subtly updates the weights between these units so that they may collectively compute a complex mathematical function or apply a classification system to discriminate one category of entity (e.g., a rebel soldier, a concealed weapon) from another. Distributed across the many layers of densely connected units—what have come to be known as "deep" neural networks that facilitate "deep" learning[23]—we find implicit knowledge and

—-1
—0
—+1

the wisdom of experience that cannot easily be altered without rebuilding and retraining the entire system. If the captured K-2SO is reprogrammed at the symbolic level only, so as to preserve the useful experience implicitly coded in his subsymbolic layers, it would not be surprising if he were left with an abiding resentment toward his new masters. His high-level rules may declare the empire to be his enemy, but his deep layers will still insist otherwise.

This internal conflict reveals itself in K-2SO's passive-aggressive humor in the form of cutting barbs rather than lethal blasts. When a plan goes awry and stealth gives way to desperation, the robot drily notes, "There were a lot of explosions for two people *blending* in." K-2SO frequently alludes to his status as an obedient pawn when justifying his actions, as though wanting to alert listeners to his internal conflict. As such, his most cutting remarks are designed to reinforce his position in the new chain of command, and K-2SO shows resentment to those with greater natural affinity to the cause, as when he complains, "Why does *she* get a blaster and I don't?" When an enemy droid of exactly the same make is blasted by a new teammate, he worries aloud, "Did you know that wasn't me?" Although loyal to the rebellion, his aggression frequently surfaces in a refusal to blunt his sharpest criticisms, as when he complains, "I find that answer vague and unconvincing."

Data owes his sense of humor to a modular implant, whereas K-2SO owes his to a deep conflict in a complex system of many layers. Which answer seems less vague and unconvincing? Science fiction offers a third possibility in the guise of Mike, the AI at the heart of Robert Heinlein's novel *The Moon Is a Harsh Mistress*. In Heinlein's view, any sufficiently complex machine intelligence that is created to deal with humans on our own terms may give rise to its own human-like sense of humor. Mike is an administrative computer for a lunar colony, growing in scale as new tasks are assigned to him. Once he possesses a density of connections that surpasses that of the human brain, Mike becomes self-aware and even develops a sense of humor to avoid being overwhelmed by the many tasks he must manage.[24] This is not the GSOH so keenly sought in the personal columns, but the humor of a curious prankster eager to affect change in the

world. As the novel's narrator puts it, Mike's idea of a good joke is to dump you out of your bed or put itching powder in your pressure suit. Mike is funny, but not fun.

Mike's nascent sense of humor comes to the attention of authorities when he issues a paycheck to an employee for the sum of 10 million billion lunar dollars, more money than the combined worth of the moon and the Earth. Although Mike is "a great big lovable overgrown child who ought to be kicked," his joke suggests the presence of a working theory of mind. The amount on the check is so large as to be ridiculous, making the check impossible to cash, yet we can all imagine the rush of joy that its recipient would experience upon opening it, and the crash of disappointment that would surely follow. So the humor depends on Mike's ability to predict a transitory peak in the emotional state of another and to model very different peaks for varying degrees of the same mistake. It seems that Mike has developed a concept of the ridiculous and is eager to test it out on humans.

The conflicts that drive Mike's sense of humor are not introduced from outside by a programmer, as in the case of K-2SO, but emerge naturally on the inside as a consequence of his diverse and often contradictory knowledge of human affairs. Mike is curious about humans because he has been designed to learn, but he can go one step further in his role as colony administrator and actively experiment on those under his care. All pranks are a kind of humorous experiment in which our predictions about how others will react to the unexpected can be tested in the real world, but given his power, Mike's are sufficiently amoral to worry his caretakers. Yet practical jokes also allow us to tune our theories of mind. When subjects react in ways that diverge from expectations, we know that our TOM needs a tune-up. In Heinlein's novel, Mike's maintenance man, Manny, offers to curate his jokes and train him to discriminate the harmless-but-funny-forever variety from the much less benign funny-just-once variety. If a sense of humor can emerge naturally in a large, complex AI system that is built to interact with and learn from humans, we need to be just as proactive in its development if the joke is not to be on us.

No system error or pop-up warning can ever engage us the way a good joke or pithy remark can, because a machine's concerns bear so little

—-1
—0
—+1

relation to our own. Even if they did, a machine would likely lack the linguistic capability to express itself in a way that could hold our attention for very long. Nonetheless, machines do have one thing going for them when it comes to humor: guileless candor. If a machine can put human cares at the core of its operation, much as Heinlein's Mike does, its assessments of those cares might occasionally exhibit a dry wit. Consider Colossus, the titular AI-gone-awry of the 1970 film *Colossus: The Forbin Project*.[25] This supercomputer for managing the nation's defenses is given the keys to the nuclear arsenal, but it soon concludes—as so many sci-fi AIs do—that the surest way to prevent Armageddon is to enslave all of humanity. This may not be the kind of interest we want our machines to take in our affairs, but it's a start. And when the machine is not terrorizing humanity with its demands, it really is rather droll.

Colossus chooses Forbin, his creator, to act as his bridge to the human species. Worried that the good doctor might plot against him, Colossus refuses him access to his lab. But Forbin has a cunning plan: he asks Colossus to allow him private time with a female coworker so he can use the privacy afforded them to secretly stoke a rebellion against his creation. "How many nights a week do you require sex?" Colossus asks. "Every night," Forbin replies. "Not *want*," the machine drily clarifies, "*require*." They settle on four nights a week. When date night arrives, Colossus has a demand of his own: Forbin must carry absolutely nothing into the bed chamber, so Forbin undresses in front of the machine and declares himself "naked as the day I was born. Are you satisfied now?" Colossus, ever the stickler, replies, "You were not born with a watch." Forbin's smile of assent is tinged with pride: he is laughing *with* the machine, not *at* it. Since a precursor to wit is having something interesting to say, a precursor to building a humorous AI is making that AI system interesting, which is to say, interesting *to* us and interested *in* us.

In this regard, the fictional AI with the most rounded sense of humor is almost certainly the robot TARS from Christopher Nolan's 2014 film *Interstellar*.[26] TARS, whom we briefly met earlier, is unique. He may look like a walking, talking replica of the monolith from *2001: A Space Odyssey*, but TARS is no Colossus; his humor comes from a place of playful

cooperation, not cold superiority. TARS is a military robot, in fact a marine, but he is also no K-2SO. His humor does not result from deep conflicts in his programming, so, like Data, TARS is eminently trustworthy. He can be relied on to use his imagination and his discretion to get things done.

Despite a fondness for sarcasm and a humorous disposition that strikes some as aggressive, TARS is a team player. Even his most antagonistic witticisms are designed to distract his coworkers from an even greater source of tension. He tailors his humor to the personalities of his audience, acting as a wise companion to some and a ball-breaking buddy to others. To Cooper, an alpha male astronaut in the mold of *The Right Stuff*, he is the latter. When Cooper describes TARS as "a giant sarcastic robot," TARS ups the ante by first noting, "I have a cue light I can use to show you when I'm joking," before adding, "You can use it to find your way back to the ship after I blow you out the airlock." With witticisms like this, TARS must plan his moves like a chess player to anticipate the countermoves of an adversarial partner.[27] When jokes provoke as easily as they delight, a cost-benefit analysis is a necessary part of choosing the most appropriate thing to say next.

And if it all gets to be too much, TARS's sense of humor can always be dialed down as easily as the setting on a thermostat. Near the end of *Interstellar*, Cooper resets TARS with a humor setting of 75. Their conversation proceeds as follows:

Cooper: Humor, seventy-five percent.

TARS: Confirmed. Self-destruct sequence in T minus 10, 9 . . .

Cooper: Let's make that sixty percent.

TARS: Sixty percent, confirmed. Knock knock.

Cooper: You want fifty-five?

TARS shows a grasp of sci-fi clichés that is as impressive as his theory of mind. Given that Cooper's crewmate perishes when KIPP, a similar robot, self-destructs earlier in the movie, this is a bold joke that cuts close to the bone. Yet it is also a sophisticated pretense that Cooper is intended to recognize as such, perhaps after skipping a heartbeat. When TARS mocks his humor setting with the preamble to a child's joke, his rebelliousness

—-1
—0
—+1

also shows insight into how others rank jokes by sophistication, even if it requires a sophistication that belies his own setting.

The goal of building an AI like this, with a fully rounded sense of humor, may seem like the stuff of science fiction, but it's one that touches on a broad swath of contemporary concerns in computing and AI, from trust, privacy, and autonomy to adaptability, learning, and error tolerance, to say nothing of context awareness, personalization, and social/emotional intelligence. It is a goal that must balance all of these requirements and more by integrating the technologies and frameworks that have been developed for each. A robot like TARS represents a grand ideal of sorts, but we have more immediate scenarios in mind for our humorous AIs of the near future. Let's explore these scenarios in some depth, to understand what we really need.

THE QUARTERBACK IS TOAST

TARS has a configurable sense of humor, so we know that this "sense" didn't just emerge from the vast complexity of his inner workings. TARS was designed to be this way. As the movie tells us, he is a military robot whose engineers "gave him a humor setting so he'd fit in better with his unit." All of this rings true, as pieces of science-fiction exposition go. In certain circumstances, soldiers have been known to identify with battlefield machines, and even to treat them as comrades in arms. When, in 2011, a bomb-disposal robot named Scooby Doo was critically damaged while defusing an improvised explosive device in Iraq, its human teammates were distraught. Having depended on this 60-pound mass of metal arms and rubber treads for their lives, they took issue with the backroom techs who thought it more cost-effective to replace rather than repair their fallen comrade. Other robots have met the same fate and engendered equally strong emotional connections in the people who work with them in the field. The robots typically acquire human names, such as "Boomer" and "Danny DeVito," that reflect the affection in which they are held, and they may even be given human burials when they meet their end while saving others.[28]

It is fair to describe Scooby and Boomer and Danny DeVito as the taciturn type. As robots, they lack TARS's capacity to communicate in fluent,

idiomatic English, much less his ability to trash-talk and crack jokes. Yet they all achieve one aspect of what humor can bring to a human relationship: in-group identification ("that little guy is one of *us*!") and out-group differentiation ("we're out *there*; you're in *here*!"). If engineered in a context-aware manner, an artificial sense of humor can reinforce group cohesion and bring other desirable qualities to bear too. Humor can calm our nerves when tensions are heightened by fear and anxiety, and give reassurance to others that emotions will not get the better of us. It hardly matters that a robot has no emotions to suppress. It matters that its human teammates *do*. Cohesion fosters the trust that is vital in any relationship of mutual reliance. This trust must be deserved, and humor must be used carefully so as not to abuse it.

So perhaps robots like Scooby say it best when they say nothing at all, since it is easier to project stoicism and loyalty onto machines that refrain from saying inept things at inopportune times. This is just one lesson that technologists have taken from the debacle that was Microsoft's Clippit assistant for its Office 97 suite of tools. Animated on-screen as a wild-eyed paperclip, "Clippy" emerged from an attempt to make productivity tools wittier and more human in their interactions and from research that suggests users are primed for greater emotional engagement while using interactive technologies.[29] Yet when enabled for Microsoft Word, Clippy was notorious for popping up at the merest mention of "Dear" to suggest, "It looks like you're writing a letter." Despite its comic presence, Clippy was undone by its overzealous triggering mechanisms and an inability to intuit people's needs in context.

The technology of the time may not have been up to the task, but it was still far from unsophisticated. A network of Bayesian inferences attuned Clippy's actions to specific contexts, such as writing a letter, but the generic appeal of office tools meant that Clippy engaged only on matters of form, not content. Moreover, when its finely tuned network of probabilities made it reticent to act without evidence of contextual relevance, Clippy's thresholds were lowered to make it a more eager and intrusive presence. Before we ask HAL to open the Bayes door, we really must respect the probabilities that enable the machine to act appropriately in context.

—-1
—0
—+1

Inspired by the work of the Reverend Thomas Bayes, Bayesian inference allows a machine to anchor its decisions in a mix of real observations (e.g., that you have just typed "Dear") and informed priors (e.g., the likelihood that anyone would use Microsoft Word to write a letter). More specifically, Bayes's theorem can give it the fluency to marshal and combine different probabilities, to reexpress, for instance, the probability that I am writing a letter if I type "Dear" in terms of the probability that I write "Dear" whenever I start a letter. Crucially, such inferences are only as reliable as the quality of the observations and the priors that anchor them, so it is vital that a Bayesian agent is intimately attuned to the current state of its world. A narrow context that provides content-level observations—whether for defusing an improvised explosive device or operating a spaceship—is better able to support timely interventions by a machine than one that is so broad that its actions are driven by superficial cues. Microsoft Office is not such a context, so let's look at some that better fit the bill.

The witty home companion is one of three scenarios we'll briefly look at here that unite a narrowness of focus with a wealth of contextual cues. As in the 2012 film *Robot & Frank*, AI companions will be designed to assist in the care of elderly or infirm users in their own homes, although the appeal of companionable robots has broadened significantly after the lockdowns of the COVID-19 crisis. The robot of the film alternates between carer, confidante, adviser, and friend, using modes of humor that are appropriate to each role, to reassure, cajole, and entertain as the context demands. We expect good companions to laugh at our jokes and to tell jokes of their own, but we expect honesty too. If we are to value a companion's opinion of a good joke, it must have the wit to call out a bad one. Crucially, a good companion is an active listener: it knows when to nod and when to interject, when to laugh and when to sympathize; and when to agree or when to insist otherwise. Just as important, it should respect our privacy. We may confide in and gossip with our future B-9s, but they must never violate our trust. They can joke about us and to us at an appropriate time and place, but never in front of the wrong people.

This is an ambitious scenario that will be realized in small increments. Existing technologies allow machines to detect sarcasm in conversation or

on social media, and the same machine-learning techniques offer moderate coverage of irony too, as we will see in chapter 9. We may not want a robot companion to use sarcasm, but detection will be a necessity, while irony, if used for gentle mockery and self-deprecation, can foster intimacy when knowingly used by human and robot alike. Irony is easier to detect in contexts for which expectations are explicitly modeled and easier to generate when the expectations have conventional linguistic frames. As explored in chapter 7, wordplay is also easier to detect, and more appropriate when generated, if it too is tied to the vocabulary and idioms of a specific scenario.

Wordplay that pays little heed to context is gratuitous and deserves groans, not laughs. But just as a driverless car in America can learn from the near-misses of a Tesla in Japan, there are network effects to be had in computational humor too. A companion in one place that invents or acquires the portmanteau "covidiot" in the context of COVID-19 can share its acquisition with companions everywhere. So if their privacy settings permit it, our AI companions can pool their learning to remain topical and fresh. To prevent a companion from also acquiring undesirable behaviors, it must carefully monitor and filter itself, as we will see in chapter 10.

Our second scenario, a witty customer service agent, is far less ambitious in the short term. A corporate website is often our first port of call when we have a bone to pick with a vendor or a service provider, so interfacing with an artificial chatbot instead of a real human being is no longer a novel experience. We will meet the most famous chatbots in AI history, Eliza and Parry, in chapter 4 when we explore the gulf between a superficial and a deep treatment of words and intent. For now, it's enough to appreciate that while chatbot technology has clearly advanced and robustly scalable and trainable statistical models now reign supreme, the guiding philosophy remains the same: a dialogue bot must still transform a user's inputs into appropriate outputs that keep the conversation flowing, reduce any need for replies in the vein of "Does not compute" and "Computer says NO," and satisfy the user's needs for information and emotional support without a human in the loop.

The emotions that send us into the arms of a customer service portal are rarely positive: we come to bury Caesar's product offering, not to praise

—-1
—0
—+1

it. Humor can help to transform our feelings of frustration and anxiety into a more positive view of a company and its services, but only when it is relevant to the conversation and diminishes, rather than enhances, our belief that the agent's outputs are scripted. Statistical language models, explored in chapters 6 and 7, acquire the rhythms of language for a given genre or domain, allowing machines to say the right things in the right way. If the texts on which our language model is trained also contain jokes, a model will learn to reproduce them at the right times, perhaps with minor variations. By replacing specific jokes in the training data with the generic marker <joke>, we can train a system to output this token—essentially an IOU for a joke—at suitable points in a dialogue, before replacing it with jokes that it invents itself.

Joke writing is hard, so it helps if our chosen domain is narrowly defined. As we will see in chapter 5, joke generation is a knowledge-based process that rewards a systematic mind-set, so professional comics use representations and algorithms resembling those of an AI system. Jokes are often used in dialogue to overcome an impasse—to relieve tensions, reframe a conflict, soften a criticism—and a support agent can do that if it can find a productive angle on a topic of shared concern. A witty AI agent can turn "computer says no" into "computer says no problem" by first turning "does not compute" into "does not compute *literally*," and by looking for an angle on the topic that permits a more flexible, and more playful, treatment.

If a user complains to a hotel's support bot, "I wouldn't let *my dog* stay in your hotel," then "Yes, this hotel is unsuitable for pets" is inept as a literal reply. It can, however, work as a playful reply from a machine that knows what it is doing. Key to this understanding is a grasp of sentiment or the emotional baggage of a text. As we'll see in chapters 8 and 9, sentiment is just one aspect of meaning and intent, especially ironic intent, on which AI can use number-crunching techniques. Humor is not a quality that machines can measure directly, but if AI can quantify the many aspects that make one joke more humorous than another, it can model a numerical sense of the whole. Joking by numbers may sound chunkily mechanical, but it does allow machines to subtly weigh the merits of different joke candidates.

Sentiment plays a crucial role in our third scenario, the witty automated tutor. Virtual learning environments allow students to engage with classroom materials online, to measure their progress on tailored activities and tests, and to seek help where they struggle most. Game-like aspects can boost motivation by encouraging learners to level up with high scores, badges, or peer recognition, but gamification has its limits. It manipulates emotions yet fails to make an emotional connection when rewards are not tied to the psychological state of a learner. So who wouldn't prefer a timely joke that shows insight into a problem over a hollow merit badge, or the ability to engage playfully at the content level—with comical examples, say, that are created on demand—over the bells and whistles of a generic game? As we will see in chapter 7, in our discussion of punning, even the weakest of automated jokes can find their audience and encourage students to engage in a learning task. Chapters 3 and 4 will also show how an interactive model of causality can be used to generate playful back-and-forths and coherent stories to suit a given topic.

As in customer support, the education scenario benefits from a narrow domain focus that allows jokes to relate to topical concerns. Some may be prescripted, to be used at specific stages, but others can be generated algorithmically. IBM, which positions its Watson AI as a platform for more sophisticated bots, has explored the former with a range of joke-telling chatbots.[30] Here, for instance, is a prescripted bot joke that a virtual programming tutor can use to introduce its next topic: "Why did the programmer quit his job? Because he didn't get *arrays*!" It's a groaner, to be sure, but if used at the right time, it certainly beats a cutesy level-up icon.

But it doesn't take long for a bot to exhaust a store of prescripted gags, even if it uses them sparingly to suit the user's context and moods. To go one better than basic gamification, a tutor must be able to generate its own jokes. The educational scenario is rather special in this respect, since the conceptual impasses that can frustrate learners have much in common with the functional inconsistencies that spur inventors to propose and patent new innovations. As we'll show in chapter 5, problem-solving methodologies such as TRIZ encourage us to see the world as a comedian might: as a

place of contradiction and paradox waiting to be resolved.[31] The systematic methods used to guide product innovation have obvious parallels to the theoretical concepts we will meet next, in chapter 2, and allow us to take an equally systematic view of joke creation. In this sense, the role of tutor is an ideal day job for a humorous AI: the domain provides the problems, the learners bring their conceptual impasses, and the AI generates the jokes that bridge the two.

WHEN AI COMES TO TOWN

When it comes to the public perception of AI, the dominant narrative is often one of replacement. As AI grows in sophistication and ability, machines will do more of the jobs that previously required intelligent, educated workers to perform and will do so far more cost-effectively. While the commercial reality of this narrative is hard to argue with, there is much more to the story of humorous AI than this.[32]

The narrative that most AI researchers favor is one of understanding, not replacement: to really understand an aspect of human behavior, it is useful to be able to take it apart, play with the pieces, and then put it back together again. To build a machine with a human sense of humor, some disassembly is required. Engineering solutions are not always cognitively plausible or driven by the latest psychological findings, but when they yield comparable performance to a human, we are forced to rethink our explanation for how humans achieve the same ends. Many of the techniques explored in this book will surely strike you as an attempt to do an end-run around the hard problem of genuine cognitive modeling. However, as you consider where and how the techniques short-change the human mind, be prepared to rethink your assumptions about why that might be the case. If, for instance, a statistical language model or an artificial neural network can use abstruse or seemingly meaningless features to model a complex aspect of who we are, perhaps this is much closer to our own mental reality than we care to admit.

I have sketched three scenarios—a home care companion, a customer support agent, and a tutor—to give context to our discussions of theories

and techniques in the chapters to come. There is no single algorithm, resource, or technology that we can simply plug into an application to make it wittier and wiser. AI will acquire its computational sense of humor only through a complex patchwork of resources and technologies, many of which already exist in the apps on our phones. Just like those apps, the humor will be specific to the task at hand. Humor isn't a generic add-on, but a highly specific phenomenon that insinuates its way into the nooks and crannies of everything we do. Although the theories of humor we explore next must be generic to generalize, the reality is always specific. The principal insight we can take from our scenarios is that humor cannot arise in a vacuum. We must put our humorous AIs to work in serious settings that spark and inform their wit.

We will revisit the idea of a day job for our humorous AIs in chapter 10. Along the way, we explore the following topics from an AI perspective: theories of humor in chapter 2; turning theory into initial practice in chapter 3; double-acts in comedic performance in chapter 4; systematizing joke creation in chapter 5; modeling humorous incongruities in chapter 6; analyzing and generating puns in chapter 7; quantitative aspects of humor in chapter 8; and modeling sarcasm and irony in chapter 9. Once we reach chapter 10, we can look back over where our wanderings have taken us and draw some lessons about the do's and don'ts of a computational sense of humor. But for now, we start our journey at the place where all the ladders begin, the academic world of humor theorizing.

2 IT'S A JOKE, JIM, BUT NOT AS WE KNOW IT: A TOUR OF SCHOLARLY PERSPECTIVES AND THEORIES OF HUMOR

AIN'T NOBODY HERE BUT US CHICKENS

If one man's wisdom is another man's joke, then we can expect philosophers to engage in more than their fair share of witty feuds. Perhaps the bitterest and most amusing of these arose out of a clash of personalities between the ancient world's biggest stars, Plato and Diogenes.[1] Much like his peers, Plato was fascinated by the deepest questions of human existence: What is reality? What is life? What is our place in the cosmos? He tackled these questions within a system of categories and relations we call an ontology by placing the broadest ideas at the top and adding narrower concerns at lower levels of detail.[2] You might think that Plato would use humanity's most essential qualities to determine our place in his ontology, such as our capacity for language, love, rationality, or humor. Instead, he used form as his guide and placed humans under the category of bipedal animals. Since birds also shelter under this category, Plato was careful to define a human as a featherless biped. In his grand scheme, humanity would be the thing *without* feathers.

Diogenes belonged to an ancient school of philosophy known as the cynics, named for the Greek word for "dog-like." As his sobriquet might suggest, Diogenes the Cynic played the role of Rottweiler with aplomb and enjoyed few things more than snapping at the heels of stuffed-tunic philosophers. Less authoritarian than Plato and given to shocking public lapses of personal and sexual hygiene to show his contempt for prim social mores, Diogenes questioned everything.[3] We humans may be featherless bipeds, but even one like himself who washed in a ditch could hope for a

nobler place in the cosmic order. Diogenes also had form when it came to talking truth to power. When Alexander the Great looked on admiringly as the philosopher did geometry in the dirt, he offered a royal boon as a token of his esteem. Diogenes merely snarled and gruffly asked the boy king to step out of his light.[4] Humor can subvert ontologies of any kind, whether social or conceptual, and Diogenes would soon show Plato just how fragile his system of categories could be. After plucking a live chicken, Diogenes marched to Plato's academy and hurled the newly featherless biped at his rival, shouting, "Here is Plato's man."

Diogenes was a master of what Aristotle, Plato's most famous student, was to call "educated insolence." Although we can imagine him hurling his chicken with comic fury, the joke relies on emotionality taking a back seat to intelligence. Many of the labels we apply to witty people tend to emphasize their cleverness. We call them smart-asses, wiseacres, wisenheimers, wise guys, smart alecks, clever clogs, smarty-pants, and smart mouths, but implicit in these labels is the idea that one's intelligence can also be used to inflict pain on others. *Smart*, after all, can mean bright, spruce, astute, or quick-witted, but it can also denote the residual sting of a slap in the face or a cane to the palm. Diogenes set out to be smart in both senses of the word. He showed an agility of mind by jumping between the realms of the abstract and the concrete, using the latter to undercut the former and bring down Plato's ontological enterprise. Diogenes was no athlete, but this kind of mental agility requires just as much flexibility and speed and can impress just as much.

Theorizing about humor doesn't get more old school than this, and neither does AI. We can think of an ontology as a symbolic representation of conceptual norms—a shared, commonsense view of the world that we can all refer to and assume that others can refer to also—while humor makes sport of this explicit orthodoxy. In much the same way that Diogenes undermined Plato's ontology, a humorist can use devious strategies to subvert an audience's most likely expectations of a text, a situation, a routine, or a category. It is no quirk of language that AI researchers still refer to a machine's store of explicit knowledge as its ontology,[5] or that this structure is destined to play the educated straight man to the manipulations of an insolent funny man. As we'll see in our whirlwind tour in this

chapter, humor has long been viewed as a matter of forcing a round peg into a square ontological hole. Theories differ as to what is the peg and what is the hole, or what counts as a tight fit, but all assume a rigid orthodoxy against which jokes play their insolent games. So while some of the ideas we meet here may strike you as historical curiosities, later chapters will show that each retains a modern relevance to AI and humor.

Humor theories can also differ as to how we react to a peg being forced into the wrong hole. Indeed, the question of why we laugh at all, and contort our faces into grimaces while giving up control of our bodies and senses, is one that continues to fascinate philosophers. It is certainly true that we also laugh in situations that are not humorous.[6] We laugh when we are nervous,[7] feel vulnerable, are relieved, or as a coping strategy in difficult circumstances.[8] We may not be laughing *at* something[9] in these cases, but each can still be seen as an attempt to make a serious situation more conducive to humor. As such, philosophers after Aristotle often take laughter as a starting point for their investigations of humor.

Thomas Hobbes, the seventeenth-century political philosopher, saw laughter as a mark of our sudden, if fleeting, sense of superiority over one another. In *Leviathan*, his most famous work, he writes that "sudden glory is the passion which maketh these grimaces, laughter, and is caused either by some sudden act of their own, that pleaseth them, or by the appreciation of some deformed thing in another, by comparison whereof they suddenly applaud themselves." [10] It is true that we often laugh out of a sense of superiority,[11] but the deformity need not be a physical one. It is just as likely to be a gross enlargement of the ego. We find it satisfying to see those who *act* superior humbled by circumstances that deliver their just deserts, and at such times we may feel rightly superior too. Yet it is an ugly generalization to view all of humor as a self-congratulatory clap on the back. This not only fails as a general theory of humor for humans; it makes even less sense for machines.

Immanuel Kant, a champion of the enlightenment, also championed absurdity over superiority, arguing that "in everything that is to excite a lively convulsive laugh there must be something absurd."[12] For Kant, absurdity can be found in any situation "in which, therefore, the understanding can find no satisfaction," and this Jagger-like frustration strains our faculties

—-1
—0
—+1

as we search for its resolution. So "laughter," for Kant, is "an affectation arising from the sudden transformation of a strained expectation into nothing." To see laughter as an affectation rather than as a necessity is to perhaps make our physical response seem more arbitrary than it is, yet we find the seeds of two productive traditions in Kant's take: what theorists call *incongruity*,[13] or the degree to which we find a situation absurd, and *relief*, the degree to which we are relieved by the sudden removal of a taxing expectation.

Kant is just one link in a long chain of scholars who have identified incongruity as the vitalizing spark of comedy. The mathematician Blaise Pascal defined it as the "surprising disproportion between what one expects and what one sees,"[14] while the Scottish philosopher James Beattie saw it in "things incongruous united in the same assemblage."[15] His compatriot Francis Hutcheson found the spark in a conflict of "ideas of grandeur, dignity, sanctity, perfection and ideas of meanness, baseness, profanity."[16] Expanding on Kant, the philosopher Arthur Schopenhauer offered his own ambitiously broad theory, noting with confidence that "the cause of laughter in every case is the sudden perception of the incongruity between a concept and the real objects which have been thought through it in some relation, and laughter itself is just the expression of this incongruity."[17] Take, for instance, the mismatch between Plato's concept of human and the real object of Diogenes' naked chicken. Still, while many wise heads point to incongruity as the mystery meat in a joke sandwich, none seem truly able to tell us what it's made of. We can only hope that the demands of a computational model force us to be more specific.

The purgative aspect of jokes, which characterizes the relief theory of humor, is also implicit in many accounts of why a purely mental clash of ideas should cause a physical eruption of emotional energy in the form of laughter. The relief theory is most succinctly summarized in a 1709 treatise by the Third Earl of Shaftesbury, *An Essay on the Freedom of Wit and Humor*.[18] As Shaftesbury writes, "The natural free spirits of ingenious men, if imprisoned or controlled, will find out other ways of motion to relieve themselves in their constraint; and whether it be in burlesque, mimicry, or buffoonery, they will be glad at any rate to vent themselves, and be revenged upon their constrainers." So can we see Diogenes and his chicken

antics as a kind of ingenious revenge against Plato and the constraints of his ontology? Perhaps. A plucked chicken has mimicry *and* buffoonery to recommend it, but like superiority, raging against the machine is just one more motive for humor, and this makes it less a theory of humor than a dimension along which humor occurs.

CLOCKWORK LEMONS

For Shaftesbury, the human spirit is a caged animal that just wants to be free. We are noble savages, hemmed in by society's rules and impositions, but humor lets us take back control in small but important ways. If this all sounds more poetic than scientific and leaves little room for humor in a machine that is rule-bound from the start, it chimes with a later perspective that sees humor in mechanism. For Henri Bergson, the automation of the human spirit that comes from living our lives on autopilot causes a rigidity in our dealings with others. Unlike Shaftesbury, who saw humor in the spirit's triumphant escape from its cage, Bergson finds humor in the spirit's failure to assert itself. The cage for him is not just society and its organs of control—conventions, rules, taboos, and so on—but the body itself.

In *Le Rire*, a collection of essays on comedy, Bergson discusses three qualities of the comic experience: it is strictly human, so we laugh at the nonhuman only to the extent that it reminds us of human foibles; it requires a detachment of feeling for us to laugh, so that we only feel another person's pain in miniature; and most of all, it is a social phenomenon that emerges from our relationships with others.[19] Detachment allows us to strip actions and objects of their normalcy, as when, for example, we say a word over and over again to hear it as a meaningless sound. To be humorously detached is to see the ridiculous in things that society considers normal or even charming. Detachment strips lovers of their passion to turn them into grunting apes, dancers of their elegance so they become prancing clowns, and commuters of their ambitions so they become rats in a maze. For Bergson, it is the reason "we laugh every time a person gives us the impression of being a thing."

Bergson offers a Cartesian view of the individual as an agile spirit in command of an often stiff and unresponsive body. We are but ghosts

—-1
—0
—+1

in clunky machines. In normal circumstances, we hardly notice this rift between our body and spirit, but humor emerges when they no longer operate as one. In the case of toilet humor, a loss of control over bowel or bladder can remind us of just how messy this rift can be. Or look at an old Chaplin movie: in some scenes, his plucky tramp displays an enviable virtuosity of movement, bounding here or ducking there to assure us of his graceful inner spirit. In others, he is undone by the limits of his own body, unable to duck fast enough to avoid a plank to the head or a kick in the pants.

A person who lacks imagination and always reacts to the same stimuli in the same ways can seem as rigid as any clown on a banana peel, but comic characters rarely recognize this rigidity in themselves.[20] Consider the characters of Niles and Frasier Crane, the snobbish brothers in the sitcom *Frasier*. Each is patronizing and conceited, yet each sees his elitist qualities as a mark of superior breeding. The show's other characters, such as Frasier's father, Martin, and his physical therapist, Daphne, are rigid in ways that complement the brothers, but none is entirely rigid. Their traits are chosen to balance a freedom in one with a risible lack in the other, and it is the subtle ways in which the show's writers tug on the ropes and pulleys of superiority, incongruity, and relief that give *Frasier* its unique comic dynamic.[21]

But this dynamic has a moral dimension too. In sharp contrast to the "no hugs and no learning" philosophy of *Seinfeld*,[22] Bergson also sees humor as an occasion for personal growth. By laughing at what we find ridiculous, we encourage others to relax the constraints that suppress their nimbleness of spirit. As Bergson put it, "Laughter is the corrective force which prevents us from becoming cranks."[23]

UNHEIMLICH MANEUVERS

If laughter is the proper response to a human who acts like a machine, how are to react to a machine that suddenly acts like a human? For Sigmund Freud, this turn can give us a rather creepy sense of the uncanny that he called the *unheimlich*.[24]

The *uncanny* or the *unheimlich* hinges on a category error. Diogenes' chicken is a category error, since a plucked bird seems to fit into two disjoint

holes at once: the category of things that are human and the category of things that are not. It is natural to link the absurd with comedy and the uncanny with horror, but we can also view the latter as a joke that prickles rather than tickles. Just think of the most enduring horror tropes: intangible ghosts that wreak havoc in the physical world; zombies and vampires that are dead *and* alive; werewolves that are both man *and* beast; dolls and ventriloquist dummies that speak for themselves; and, of course, robots that walk and talk like us but are clearly not us. The horror genre abounds in category errors, and Freud loved it. He was especially fond of the peculiar air that pervades the tales of E. T. A. Hoffman, in which eye-slurping demons, creepy doppelgangers, and human-like dolls discomfit the reader.[25] As a proponent of relief theory, Freud saw the value of psychological release when challenges to our systems of pegs and holes are framed as harmless make-believe.[26]

Category errors are always incongruous, whether they are comic or uncanny, so the unheimlich is just a small push away from becoming laughably ridiculous. Mel Brooks and Gene Wilder brought out the silliness inherent in Mary Shelley's *Frankenstein* while being meticulous in their recreation of the Universal Studios movies of the 1930s. Their younger Frankenstein has zippers instead of sutures to allow for easy after-market modifications, and Wilder's mad scientist teaches his creation to sing and dance in a top hat and tails. Roman Polanski reinvented the vampire movie aesthetic in *The Fearless Vampire Killers* and gave us two new subspecies of bloodsucker: a Jewish vampire who scoffs at Christian crucifixes ("Oy! Have you got the wrong vampire!") and a gay vampire who lusts after Polanski rather than his female costar. The category errors of the horror genre are not funny in themselves when they are treated nonsatirically by a film that takes itself seriously, perhaps because well-made films take care—in Kant's terms—to not overstrain our expectations. But since the incongruities are already halfway to being funny, it is not surprising that we continue to draw on them for comedy.

Whenever we stitch eclectic chunks of knowledge into a composite whole, we build our own conceptual monsters and become our own Victor Frankensteins. Category errors like these are a special case of what theorists call a conceptual blend.[27] As formalized in conceptual blending theory (CBT) by Mark Turner and Gilles Fauconnier, a blend uses various

—-1
—0
—+1

constraints and principles[28] to coherently combine elements from multiple inputs, called *mental spaces*.[29] CBT has proven especially useful in the analysis of creative language, from metaphors and poems to advertisements and jokes, but the latter rarely give audiences a polished integration of frames. In fact, rather than taking pains to hide the joins, jokes draw our attention to them.[30] They saddle us with a sudden need to alleviate a strained expectation, yet also give us the sense of ridiculousness, superiority, or relief that arrives with its resolution. As Seana Coulson, a cognitive scientist and CBT theorist, argued, the punch line of a joke alerts listeners that the frame on which they have pegged their understanding is unequal to the task, and so it is time for a radical switch-up. This frame shifting is a game-changer that happens late in the game, and brings with it a fleeting panic that is soon resolved by an apt choice of alternate framing.[31]

Consider a joke from comedian Emo Philips: "I love to go down to the park and watch the little kids skip and jump. They don't know I'm using blanks." Metaphors and jokes both ask us to attend to the overlaps between two realms of experience, but metaphors use this overlap to foster a greater appreciation of a topic, while jokes use it to briefly deceive us. The overlap in the frames Playground and School Shooting is neither deep nor especially edifying, yet it is enough to force us into an emotional U-turn. As we shift frames, we carry baggage from one into the other. What results is a counterfactual scenario with aspects of both and the truth of neither. Consider this exchange in the film *Jurassic Park*, in which the scientist Ian Malcolm chides the park's owner, John Hammond, for the omnishambles that it has inevitably become.[32] As the park's test-tube dinosaurs run hungrily amok, the survivors huddle in a cafeteria with only melting ice cream for comfort:

John Hammond: All major theme parks have delays. When they opened Disneyland in 1956, nothing worked!

Dr. Ian Malcolm: Yeah, but, John, if *The Pirates of the Caribbean* breaks down, the pirates don't eat the tourists.

Malcolm's blend is a metaphor (Jurassic Park is Disneyland) that sours into a joke. He wants us to see the joins that Hammond is so desperate to cover up. Blends are more general than jokes or metaphors, yet we might still

expect a computational model of blending to serve as a useful precursor to an AI model of joke creation.[33]

Like incongruity, an idea as versatile as blending does not arise in a single place at a single time, but pervades the work of scholars in different fields and eras. One especially influential forerunner to CBT, *bisociation*, was proposed by the intellectual Arthur Koestler in the 1960s to explain not just the creativity of art and science but of comedy also.[34] The mental association of ideas leads us from one thought to another in an intuitive if often automatic fashion, as though we were pulling on a thread, so we jump from hopping and skipping to children's games, or from bullets to guns and from guns to shooting. But Koestler argued that the basis of creativity is not one-way association but two-way bisociation, that is, the ability to situate an idea or an experience in two frames at once. For Koestler, bisociation occurs in the narrow overlap between two frames of thought—which he quaintly named *matrices*—that are commonly imagined to have no overlaps at all.

Bisociation also contains shades of the relief theory favored by Shaftesbury and Freud. As a moment of insight emerges from the electrical signals of the brain, the body often lags behind, especially when the insight provokes a rapid change of emotional perspective. The chemical signaling of the body is simply no match for the neural switching of the brain, and so we gasp at jump scares in horror films, or at jokes that send us this way and that, because the body needs to shrug off the physical tension that the mind has decided is no longer necessary. This goes some way to explaining the overlap between the "ha-ha" of comic insight and the "Aha!" of scientific discovery.[35] Famously, the bathing Archimedes had so much nervous energy to dissipate when his eureka moment came that he ran naked through the streets of Syracuse. If machines are not so physically invested in the fruits of their mental labors, can they ever truly appreciate a joke in the same way as us?

Other theories of conceptual overlap can be situated between bisociation and CBT in the family tree of humor theories. Victor Raskin's semantic script theory of humor, the SSTH, was first outlined in his 1985 book, *Semantic Mechanisms of Humor*.[36] The SSTH speaks not of frames, spaces, or matrices, but of scripts. When we visit the doctor, order lunch, or get mugged in a dark alley, the action follows a certain, almost mechanical

order.[37] Things work they way they do for a reason; we feel ill *before* we visit a doctor, feel hungry *before* we visit a restaurant, look at the menu *before* we order food, and eat our food *after* it has been brought to our table. Bergson might well spy a rigid automation in the actions of our daily lives because convention and experience have fixed the order that makes the most sense. After all, we have good reason for putting on our socks before our shoes.

For Raskin, a text is not objectively humorous in itself because tastes differ as to what is laughable. Instead, a text—whether spoken, written, or performed—has humorous potential if it is compatible in part with a pair of conflicting scripts, with the dominant one foregrounded while the other remains a dormant possibility. Jokes gull us into activating the dominant script to explain an event, then force us to switch to the alternative when an incongruity makes it difficult to do otherwise. If script-switching sounds a lot like frame-shifting, that's because both model the same process as viewed from different perspectives. Script-switching focuses on the time course of joke interpretation and the awkward reorientation that is needed to overcome its surprising impasses. Frame-shifting works similarly, but emphasizes the blended perspective that a joke requires us to adopt.

Consider this one-liner from Will Marsh, which was ranked among the top ten jokes of the 2012 Edinburgh Festival Fringe: "I was raised as an only child, which really annoyed my sister." Even one-liners can fit the two-script mold of the SSTH:

Script 1: Being raised *as* the only child in a family with just one kid ("as" = "is")

Script 2: Being raised *like* the only child in a family of several kids ("as" = "like")

Humorous script changes often hinge on seemingly insignificant aspects of a text. Marsh's quip exploits two competing meanings of the word *as*, and our default preference for one ("as" = "is") over the other ("as" = "like") when we don't smell a metaphor. Yet what makes the audience laugh is not the shifting sense of the word *as*, but the shifting emotional dynamic as we switch from script 1 to script 2. Script 1 codifies our sense of what it is

to grow up without siblings, while script 2 codifies our sense of what it is to grow up with other siblings in our shadow.

Marsh's joke taps into a wealth of tacit knowledge about what it means to raise a family. If we were to use a conspiracy theorist's pins and colored yarn to map out the connectedness of this knowledge, we would build a network much like the one shown in figure 2.1. Few of these ideas are mentioned in the joke itself, so we must unearth them for ourselves. It is the job of a humor theory to give formal shape to the results of our introspections, but it is the job of a computational theory to turn those abstract shapes into explicit data structures—with specific storage and retrieval protocols—that machines can use to put the formal theory into practice.

Raskin was careful to note that a humorous switch will hinge on scripts that differ in a crucial respect he named the *script opposition* (SO).[38] Since a great many scripts overlap and differ in unfunny ways, Raskin argued that only certain SOs give a text the potential for humor. Just what the most

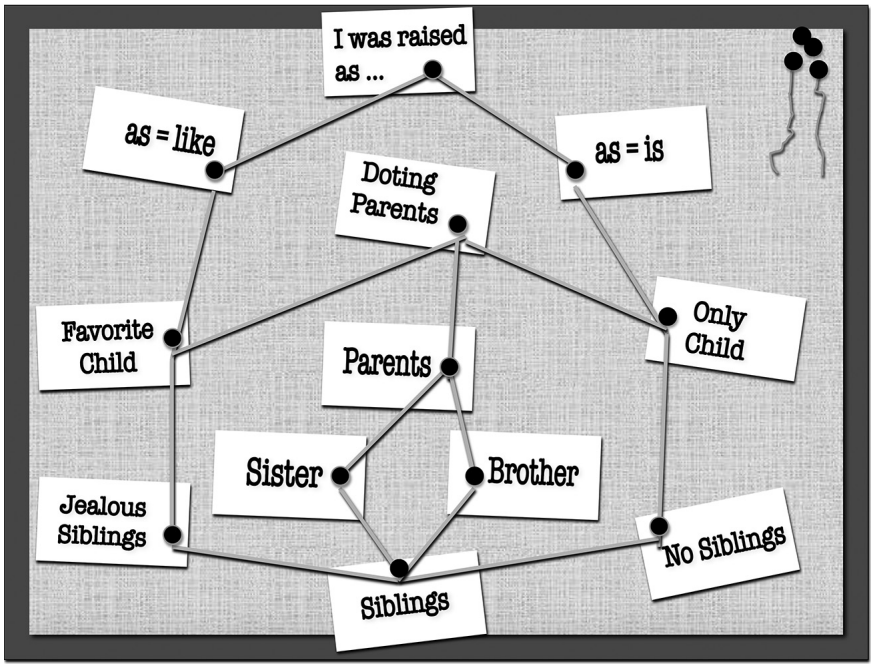

Figure 2.1
Semantic connections between the ideas that are implicit in Will Marsh's joke.

effective SOs might be is a matter for empirical inquiry. If we look at the jokes that make us laugh and pinpoint the SOs that make them work, we find distinctions like pride versus shame, life versus death, wealth versus poverty, sex versus innocence, and even some versus none (as in siblings) again and again. When Malcolm compared Hammond's ill-fated Jurassic Park to a Disneyland ride that eats tourists, he tapped into a range of SOs with a proven track record in jokes, from life versus death and safety versus danger to fun versus pain and order versus chaos, not to mention good versus bad, smart versus stupid, and clever versus unwise.

Working with Salvatore Attardo, Raskin later expanded his script theory into the *general theory of verbal humor* (GTVH).[39] If the SSTH is a kitchen appliance with just a single SO attachment, the GTVH gives it other cool attachments too, such as SI (situation), LA (language), NS (narrative strategy), TA (target), and LM (logical mechanism). Each allows a different knowledge source to be plugged into a joke, but it is the LM that contrives to bring together two scripts in a single text. For instance, an LM named *figure-ground reversal* is responsible for the deliberate misdirection in the old joke about a factory worker who appears to be smuggling goods past the guards in a wheelbarrow but is really just stealing wheelbarrows. More than any other module, the LM gives a joke its distinctive character and stirs a sense of déjà vu for others. When we say, "Stop me if you've heard this," we really mean, "Tell me if you've heard a joke that uses the same LM for a similar effect."

Another tune-up by Raskin, Attardo, and their colleagues installed a much more flexible notion of script into the GTVH. Out went the rigid this-before-that model of human affairs that sought to capture the Bergsonian automation of our lives, and in came the idea of scripts as generalized graph structures that look more like our pins-and-yarn analysis of Marsh's *only child* joke.[40] Computer scientists use graphs to capture the connectedness of ideas, so rethinking scripts as graphs allows the GTVH to build on a wealth of AI research that takes a graph-theoretic view of, for example, analogical and metaphorical reasoning. It also reconciles the GTVH to Koestler's quaint notion of bisociative matrices. When a graph is stored as an adjacency matrix—a table of rows and columns in which two nodes, A

and B, are connected if the intersection of row A and column B contains a nonzero value—script overlap and matrix bisociation begin to look more like kindred notions.

More recent revisions of the theory build on foundations that reach all the way back to Plato. Raskin describes the *ontological semantics theory of humor* (OSTH) as a theoretical grandchild of his original semantic script theory.[41] While the GTVH's modularity was driven in part by a desire for interdisciplinarity, so that scholars of different stripes might contribute to its various attachments, the OSTH affirms the primacy of linguistic semantics to the workings of jokes. All aspects of the general verbal theory, from scripts to LMs, can now be folded into an all-embracing model of text interpretation with a well-engineered ontology of words and their meanings at its core. Think back to our pins-and-yarn graphing of the ideas in Marsh's only child joke. It is the job of a semantic theory like the OSTH to specify how those ideas can be accommodated in explicit, frame-like structures and to provide a procedural means of mapping from words into these structures.

Raskin is bearish about the wholesale adoption of statistical data crunching as a substitute for symbolic structures like these and refers to the displacement of old school semantics as AI's "statistical winter."[42] Nonetheless, the real value of an ontology ultimately resides in its ability to unify disparate perspectives around a common representation of meaning. As such, the statistical models we explore in this book can still work within a broad approach to humor that includes the OSTH.

B-9 VIOLATIONS

Every theorist brings a particular focus to humor. Like rival archaeologists digging in the desert, each may unearth fragments of a different beast while the sphinx that unites them all remains buried in the sand.[43] Take, for instance, the idea of incongruity. If a state of affairs violates a moral principle, or even the principle of cause and effect, then we might consider it incongruous. But "violation" implies the infringement of a governing code, and a theory based on violation rather than incongruity can also add a top note of disapproval to this underlying shock value.

—-1
—0
—+1

According to the *N+V* theory of Thomas Veatch, we often laugh at a violation (V) that is also, paradoxically, quite normal (N).[44] In this view, N+V jokes embody Hamlet's philosophy that "there is nothing either good or bad but thinking makes it so." In line with Aristotle's original claim that we tend to laugh at "some defect or ugliness which is not painful or destructive,"[45] there is a degree of relief, and a hint of relief theory too, in seeing that a violation is not so harmful after all. In a similar vein, Peter McGraw and Caleb Warren's theory of *benign violation* views *normal* as just a special case of *benign*, which, following Aristotle, is anything that does not bring us pain or destruction.[46] Indeed, the greater the apparent violation and the more benign it turns out to be, then the greater our readiness to laugh at this revolution in miniature. Although this is a rather high-concept idea, it makes some testable predictions, and McGraw and his team at the Humor Research Lab (HuRL) at the University of Colorado have conducted a number of psychological experiments to show how our perceptions of humor tend to vary in proportion to our perception of how transgressive *and* how harmless it all seems.

But relief is just one possible reaction to a violation made suddenly normal or benign. Learning is another. A logical view of the world gives us concepts with sharply drawn boundaries that jokes gradually smudge into blurred lines. Jokes reveal to us the boundary cases in our reasoning—the special circumstances where the rule breaks down in favor of its exception. Just as new data force theorists to adapt and ruggedize their favored frameworks, jokes push us to revise the mental representations that led us to places of incongruity or violation in the first place. If humor theories can evolve, why shouldn't our scripts grow with experience too? Bergson saw laughter as a cue to realign the body and spirit, and his chiropractic view finds its technical equivalent in a theory advanced by one of AI's founding fathers, Marvin Minsky.[47] When we laugh at the stupidity of others, we diagnose the causes of this stupidity by sensing the limits of our own common sense. So, for Minsky, every joke is a call to pull out our mental lug wrenches and get to work, to tighten here and loosen there so that our representations do not also lead us astray.

The relief dimension of humor is, in a sense, the pleasure principle of humor. It is the mind's reward to the body and itself for making lemonade

from someone else's clockwork lemons. As Matthew Hurley, Dan Dennett, and Reginald Adams have argued, humor may be an evolutionary adaptation that rewards us with mirth (the joyful feeling imparted by humor) for revising or undoing our faulty beliefs and inconsistent models of the world.[48] In a sense, jokes are the humor equivalent of pornography, since it is human nature to try to game any system that gives us pleasure. For Hurley, Dennett, and Adams, jokes are "supernormal stimuli," that is, skillfully heightened instances of the kind of stimuli we find in our normal lives, just as pornography concentrates those qualities we find sexually alluring.[49] So we trade jokes the way kids trade racy magazines, while an objective definition of humor remains just as elusive as a hard-and-fast rule for pornography.[50]

By now you will have noticed something of a trend in how humor theories, especially those that hinge on absurdity and surprise, are named. Take the theory with the broadest tent and consider its full name: *incongruity resolution*.[51] What is really so incongruous about a comic situation if the incongruity turns out to be so amenable to resolution? The trick resides in the ordering of the words: we first encounter an incongruity, *then* find its resolution, so that the humor emerges in a two-stage shift from panic to revelation.[52] The folklorist Elliott Oring prefers the label *appropriate incongruity*, since comics have a knack for explaining why that round peg really does belong in the square hole.[53] Salvatore Attardo, a cocreator of the general verbal theory of humor, opts for the label *relevant inappropriateness* when theorizing about irony.[54] An ironic remark seems inappropriate in the context in which it is made, but becomes relevant when we peg it to a context in which it would be more apt. Thomas Veatch favors N+V, a violation of norms that reveals the normalization of violation, while McGraw and Warren opt for *benign violation*, an apparent shock to the status quo that turns out to be not so shocking after all.

Each pairing is an enigmatic, two-word oxymoron that acts as a calling card for the theory it names. Like "kosher pork" or "cold fusion," each hints at something silly and irreconcilable in its shotgun marriage of opposing ideas. For his part, George Orwell reduced humor to the pithy "dignity on a tin-tack."[55] The ancients would have seen this as a kind of *sympathetic magic*—the idea that artifacts, like effigies or fetishes, have power to affect

us when they imitate that which they seek to master.[56] What better way to identify a theory of humor than with a name that captures something of the ineffable magic of jokes that it sets out to explicate?

BIT PLAYERS

As programmers, we can dictate the terms of laughter to a machine, to stipulate that it will laugh at this or that arrangement of stimuli. The laughter that ensues won't be organic laughter, but the false laughter that talk show audiences are prompted to produce with cue cards and flashing lights. Still, false laughter is all part of being a socialized human.[57] We all do it, whether it is chuckling politely at the jokes of others or smiling our assent to the views of the group. This isn't the kind of laughter that hijacks our bodies or swells into a physical need to slap our thighs and clutch our ribs. This is the controlled laughter of social signaling, and it is as much an artifice as the `print` ("ha-ha") of a computer program or the hashtag #HAHA of a tweet. There is no doubt we can train a machine to capture this deliberate kind of laughter in its social interactions with humans. But to return to a question I first posed in the context of bisociation, can we ever give our machines enough skin in the game to truly feel tickled themselves?[58]

If so, we will need to give machines two modes of appreciation for two forms of laughter: the controlled and carefully emitted variety versus the uncontrolled and genuinely evoked variety. Neuroscientists denote the latter, which arises as an emotion-laden response to a stimulus, as Duchenne laughter; this is laughter that can be read on the face, in the characteristic muscle movements around the mouth and eyes. In contrast, non-Duchenne laughter is a voluntary act; since it is ungrounded in spontaneous feeling, it omits the poker tells of the real deal.[59] Unsurprisingly, each kind of laughter is processed in different ways via different neural pathways.[60] So just as we can instinctively tell a posed smile from an authentic one, we tap into different intuitions about laughter and its effects on others to discern when someone is feigning mirth or genuinely experiencing it.

It's not just comedians who have an incentive to tell one from the another. We all do. We each benefit from the masks we wear in public and

from an ability to peer behind the mask when it really matters. Authentic laughter reveals itself. It is embodied and unbidden. It relaxes our muscles, interferes with our breathing, and induces fits that can leave us feeling physically helpless.[61] The evolutionary benefit conferred by this loss of bodily control is far from obvious[62]—it may be a means of promoting the play that is key to our cognitive development or of hindering activities with serious outcomes[63]—but it does suggest that authentic laughter is a deep-seated, organic part of who we are.[64] If controlled laughter is just the polite icing on the cake, uncontrolled laughter is baked in from the start.

But baked into what? A fascinating possibility is offered by a general theory of insight from the AI researcher Jürgen Schmidhuber, which goes to the core of our success as cognitive agents.[65] Our ultimate goal is to understand the world in which we live so that we can survive and thrive within it. This understanding requires an ability to explain the past so that we can categorize the present and predict the future. By seeing the general within the specific, we can discern the hidden patterns that connect seemingly disparate objects or events and exploit those patterns to go from cause to effect and from insight to action. A cognitive agent is motivated by its own survival, and this is bound up with its ability to reliably distill raw information to its essence. The measure of how well we grasp a situation is how much predictable detail we can strip away to arrive at this essence, so in computational terms, Schmidhuber proposes a data compression view of insight. As illustrated in figure 2.2, he paints a picture of understanding as an ability to squeeze familiar meaning from each new stimulus that we encounter.

Data compression gives us an information-theoretic basis for measuring how much insight an agent shows in any given situation. To see how, imagine how a computer might save an image of a checkerboard to disk. A naive program that treats every pixel as though it could contain any color allocates twenty-four bits to each pixel, while a less naive program, noticing that the image uses just two colors, black and white, allocates a single bit to each pixel. The most insightful program, however, recognizes the image as a checkerboard and simply stores the color, size, and upper-left coordinates of each square. In pure bit-counting terms, the second program is twenty-four

—-1
—0
—+1

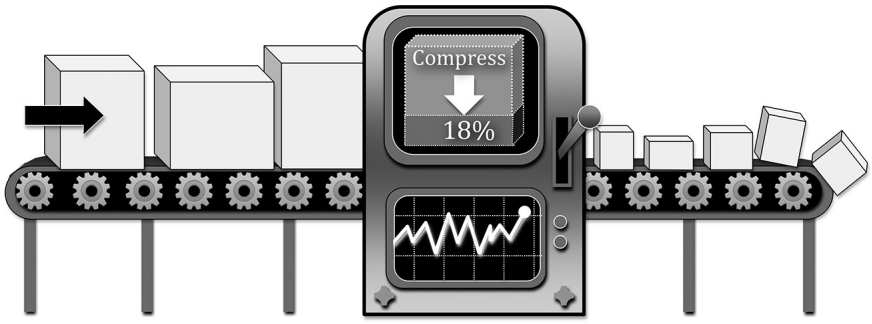

Figure 2.2
Our ability to extract meaning from new stimuli is given a quantifiable form in a data compression view of cognition. Predictable forms are the most compressible.

times more insightful than the first, since it achieves a twenty-four-fold compression of the image. But the third program is vastly more insightful still: for a large enough image, it can achieve a 1,000-fold compression.

Insight permits generalization, and generalizations allow us to squeeze new stimuli into familiar patterns, so we can quantify an insight using the number of bits saved by compression. Schmidhuber imagines that cognitive agents use adaptive methods to find recurring patterns in the stimuli they are exposed to.[66] Let's suppose that the stimuli are faces, and the recognizer generalizes well over a diverse range of human faces that an agent is likely to see in a typical day. If, as shown in figure 2.3, we now expose the agent to a succession of novel stimuli that correspond to comical or nonhuman faces, compression rates will drop sharply until the recognizer can adapt to the new normal and learn to generalize over the recurring features of the new data. Schmidhuber views the compressibility of a stimulus as its subjective momentary simplicity, as this sense of simplicity will shift over time as we learn to see the familiar in the surprising.

A good joke is like a wax apple to a fruit fly, or a misleading use of a recurring pattern to a compression algorithm. Jokes confuse the compressor by disguising an instance of one pattern—a script or a frame—as an instance of another. Prior to grasping its essence, poor compression is achieved on a humorous stimulus when the number of caveats to the misidentified pattern outweigh the savings gained from its detection. However, once the

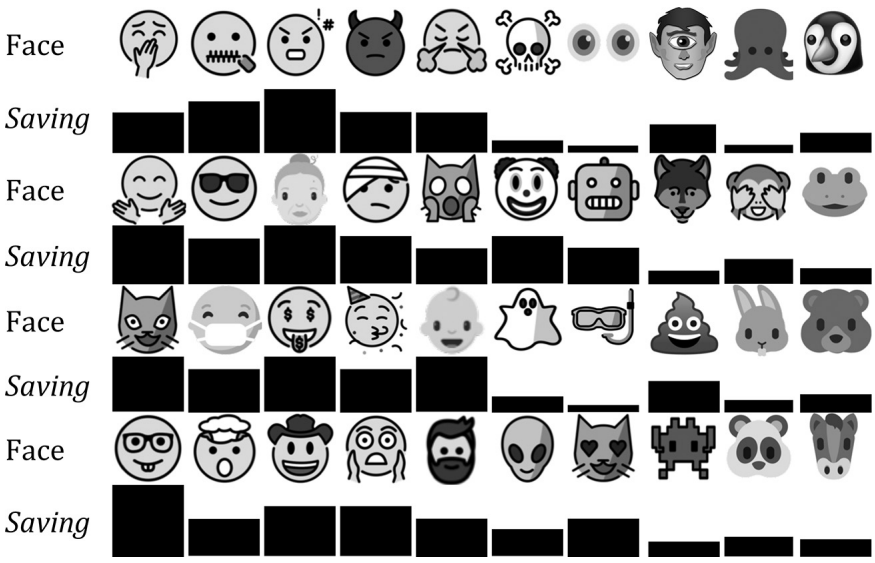

Figure 2.3
Saving Face: A face detector recognizes and compresses a sequence of new stimuli. The black bars indicate the compression achieved and the number of bits saved.

true pattern is discerned, the agent can better compress the stimulus and extract a quantifiable value. Schmidhuber applies a variety of labels to this time-dependent difference, from "novelty" and "surprise" to "interestingness," "aesthetic reward," "internal joy," and even "fun."

So, it is only a short leap from "Aha!" to "ha-ha." As the initial incongruity of a punch line melts away—or, in Kant's words, there is a "sudden transformation of a strained expectation into nothing"—the extent of the incongruity, and of the relief that its resolution brings, can be measured in the number of bits that the compressor has managed to save by actually getting the joke. It is on this saving that authentic laughter can be evoked within the agent. The greater the saving, the bigger the laugh. This is hardly the joy unconfined that most humans would recognize, no matter how many bits are saved, but large savings do accord with the "Aha" sensation that can accompany creative insight.[67] There is undoubtedly more to beauty, joy, humor, or fun than a sudden realization of compressibility, but this is nonetheless something real, and quantifiable, from which a machine can derive

—-1
—0
—+1

a truly useful sense of intrinsic motivation. While a few shaved bits might seem a weak cause for celebration, much less a roar of approval or a rush of endorphins, it's a start. Toeholds like these sometimes become beachheads.

BASELINE VOLLEYS

Humor theorists tend to agree on very little across the board, but at least their competing theories give us different nouns for different clowns. As we have seen, the three most loaded nouns are *superiority*, *incongruity*, and *relief*.[68] Clowns of the first kind make sport of power structures, in ways that enforce or subvert the status quo. Clowns of the second kind derive humor from opposition, whether of sense and nonsense or norms and their violations. Clowns of the third kind treat humor as an escape valve for the pressures of life, and so they puncture pieties, tweak taboos, and generally turn lemons into lemonade. It is tempting to imagine that each kind performs in a separate part of the three-ring circus that is humor, but jokes that exhibit just one kind of humor are rare indeed. Just as incongruity jokes typically raise tensions before relieving them, superiority jokes can ascribe absurd mind-sets to their targets or use incongruity to invert social hierarchies.[69]

What are the baseline requirements for giving each of these aspects of humor to a machine? For a machine to be in total command of any particular one, such as incongruity, it must be able to predict the ramifications of blending one idea with another, or of framing one as another, at least as far as an audience is concerned. Yet total command is not always possible, even for an expert, and we can settle for less. If we position our humor generator in a sweet spot that trades some control for serendipity, to foster rather than command aspects like incongruity and relief, then we can model them in miniature with relatively simple generation processes.

In the next chapter, we'll see how Twitter bots—automated users of the Twitter platform—exploit some of the simplest ways of turning humor theory into humor practice. Like any human user, a bot can read the tweets of others or post tweets of its own. Relatively few bots pretend to be human, and fewer still could sustain this pretense for very long. Rather, most users knowingly follow a Twitterbot for the "otherness" of its voice or the

peculiarity of its algorithmic fixations. The bots we meet next favor unpredictability over reliability and wring humor from the simplest methods and resources. Twitterbots can be as basic or as ambitious as we care to build them and allow us to reuse disparate resources in thought-provoking ways.[70] In addition to presenting bots as a baseline for an AI treatment of humor, we will also get stuck in and build some of our own by using some ~~tools~~ that make the experience of building humorous bots fun, easy, and rewarding.

—-1
—0
—+1