3

Mentalese

The year 1984 has come and gone, and it is losing its connotation of the totalitarian nightmare of George Orwell's 1949 novel. But relief may **be** premature. In an appendix to *Nineteen Eighty-four,* Orwell wrote of an even more ominous date. In 1984, the infidel Winston Smith had to be converted with imprisonment, degradation, drugs, and torture; by 2050, there would be no Winston Smiths. For in that year the ultimate technology for thought control would be in place: the language Newspeak.

> The purpose of Newspeak was not only to provide a medium of expression for the world-view and mental habits proper to the devotees of Ingsoc [English Socialism], but to make all other modes of thought impossible. It was intended that when Newspeak had been adopted once and for all and Oldspeak forgotten, a heretical thought—that is, a thought diverging from the principles of Ingsoc—should be literally unthinkable, at least so far as thought is dependent on words. Its vocabulary was so constructed as to give exact and often very subtle expression to every meaning that a Party member could properly wish to express, while excluding all other meanings and also the possibility of arriving at them by indirect methods. This was done partly by the invention of new words, but chiefly by eliminating undesirable words and by stripping such words as remained of unorthodox meanings, and so far as possible of all secondary meanings whatever. To give a single example. The word *free* still existed in Newspeak, but it could only be used in

> such statements as "This dog is free from lice" or "This field is free from weeds." It could not be used in its old sense of "politically free" or "intellectually free," since political and intellectual freedom no longer existed even as concepts, and were therefore of necessity nameless.
>
> ... A person growing up with Newspeak as his sole language would no more know that *equal* had once had the secondary meaning of "politically equal," or that *free* had once meant "intellectually free," than, for instance, a person who had never heard of chess would be aware of the secondary meanings attaching to *queen* and *rook*. There would be many crimes and errors which it would be beyond his power to commit, simply because they were nameless and therefore unimaginable.

But there is a straw of hope for human freedom: Orwell's caveat "at least so far as thought is dependent on words." Note his equivocation: at the end of the first paragraph, a concept is unimaginable and therefore nameless; at the end of the second, a concept is nameless and therefore unimaginable. *Is* thought dependent on words? Do people literally think in English, Cherokee, Kivunjo, or, by 2050, Newspeak? Or are our thoughts couched in some silent medium of the brain—a language of thought, or "mentalese"—and merely clothed in words whenever we need to communicate them to a listener? No question could be more central to understanding the language instinct.

In much of our social and political discourse, people simply assume that words determine thoughts. Inspired by Orwell's essay "Politics and the English Language," pundits accuse governments of manipulating our minds with euphemisms like *pacification* (bombing), *revenue enhancement* (taxes), and *nonretention* (firing). Philosophers argue that since animals lack language, they must also lack consciousness—Wittgenstein wrote, "A dog could not have the thought 'perhaps it will rain tomorrow' "—and therefore they do not possess the rights of conscious beings. Some feminists blame sexist thinking on sexist language, like the use of *he* to refer to a generic person. Inevitably, reform movements have sprung up. Many replacements for *he* have been suggested over the years, including *E, hesh, po, tey, co, jhe, ve, xe, he'er, thon,* and *na*. The most extreme of these movements is General Semantics, begun in 1933 by the engineer Count Alfred

Korzybski and popularized in long-time best-sellers by his disciples Stuart Chase and S. I. Hayakawa. (This is the same Hayakawa who later achieved notoriety as the protest-defying college president and snoozing U.S. senator.) General Semantics lays the blame for human folly on insidious "semantic damage" to thought perpetrated by the structure of language. Keeping a forty-year-old in prison for a theft he committed as a teenager assumes that the forty-year-old John and the eighteen-year-old John are "the same person," a cruel logical error that would be avoided if we referred to them not as *John* but as *John*$_{1972}$ and *John*$_{1994}$, respectively. The verb *to be* is a particular source of illogic, because it identifies individuals with abstractions, as in *Mary is a woman,* and licenses evasions of responsibility, like Ronald Reagan's famous nonconfession *Mistakes were made*. One faction seeks to eradicate the verb altogether.

And supposedly there is a scientific basis for these assumptions: the famous Sapir-Whorf hypothesis of linguistic determinism, stating that people's thoughts are determined by the categories made available by their language, and its weaker version, linguistic relativity, stating that differences among languages cause differences in the thoughts of their speakers. People who remember little else from their college education can rattle off the factoids: the languages that carve the spectrum into color words at different places, the fundamentally different Hopi concept of time, the dozens of Eskimo words for snow. The implication is heavy: the foundational categories of reality are not "in" the world but are imposed by one's culture (and hence can be challenged, perhaps accounting for the perennial appeal of the hypothesis to undergraduate sensibilities).

But it is wrong, all wrong. The idea that thought is the same thing as language is an example of what can be called a conventional absurdity: a statement that goes against all common sense but that everyone believes because they dimly recall having heard it somewhere and because it is so pregnant with implications. (The "fact" that we use only five percent of our brains, that lemmings commit mass suicide, that the *Boy Scout Manual* annually outsells all other books, and that we can be coerced into buying by subliminal messages are other examples.) Think about it. We have all had the experience of uttering or writing a sentence, then stopping and realizing that it wasn't exactly what we meant to say. To have that feeling, there has to be a "what we meant to say" that is different from what we said.

Sometimes it is not easy to find *any* words that properly convey a thought. When we hear or read, we usually remember the gist, not the exact words, so there has to be such a thing as a gist that is not the same as a bunch of words. And if thoughts depended on words, how could a new word ever be coined? How could a child learn a word to begin with? How could translation from one language to another be possible?

The discussions that assume that language determines thought carry on only by a collective suspension of disbelief. A dog, Bertrand Russell noted, may not be able to tell you that its parents were honest though poor, but can anyone really conclude from this that the dog is *unconscious?* (Out cold? A zombie?) A graduate student once argued with me using the following deliciously backwards logic: language must affect thought, because if it didn't, we would have no reason to fight sexist usage (apparently, the fact that it is offensive is not reason enough). As for government euphemism, it is contemptible not because it is a form of mind control but because it is a form of lying. (Orwell was quite clear about this in his masterpiece essay.) For example, "revenue enhancement" has a much broader meaning than "taxes," and listeners naturally assume that if a politician had meant "taxes" he would have said "taxes." Once a euphemism is pointed out, people are not so brainwashed that they have trouble understanding the deception. The National Council of Teachers of English annually lampoons government doublespeak in a widely reproduced press release, and calling attention to euphemism is a popular form of humor, like the speech from the irate pet store customer in *Monty Python's Flying Circus:*

> This parrot is no more. It has ceased to be. It's expired and gone to meet its maker. This is a late parrot. It's a stiff. Bereft of life, it rests in peace. If you hadn't nailed it to the perch, it would be pushing up the daisies. It's rung down the curtain and joined the choir invisible. This is an ex-parrot.

As we shall see in this chapter, there is no scientific evidence that languages dramatically shape their speakers' ways of thinking. But I want to do more than review the unintentionally comical history of attempts to prove that they do. The idea that language shapes thinking seemed plausible when scientists were in the dark about how thinking works or even how to study it. Now that cognitive scientists know how to think about thinking, there is less of a temptation to equate it with language just because words are more palpable than thoughts. By understanding *why* linguistic determinism is wrong, we will be in a better position to understand how language itself works when we turn to it in the next chapters.

The linguistic determinism hypothesis is closely linked to the names Edward Sapir and Benjamin Lee Whorf. Sapir, a brilliant linguist, was a student of the anthropologist Franz Boas. Boas and his students (who also include Ruth Benedict and Margaret Mead) were important intellectual figures in this century, because they argued that nonindustrial peoples were not primitive savages but had systems of language, knowledge, and culture as complex and valid in their world view as our own. In his study of Native American languages Sapir noted that speakers of different languages have to pay attention to different aspects of reality simply to put words together into grammatical sentences. For example, when English speakers decide whether or not to put *-ed* onto the end of a verb, they must pay attention to tense, the relative time of occurrence of the event they are referring to and the moment of speaking. Wintu speakers need not bother with tense, but when they decide which suffix to put on their verbs, they must pay attention to whether the knowledge they are conveying was learned through direct observation or by hearsay.

Sapir's interesting observation was soon taken much farther. Whorf was an inspector for the Hartford Fire Insurance Company and an amateur scholar of Native American languages, which led him to take courses from Sapir at Yale. In a much-quoted passage, he wrote:

> We dissect nature along lines laid down by our native languages. The categories and types that we isolate from the world of phenomena we do not find there because they stare every observer in the face; on the contrary, the world is presented in a kaleidoscopic flux of impressions which has to be organized by our minds—and this means largely by the linguistic systems in our minds. We cut nature up, organize it into concepts, and ascribe significances as we do, largely because we are parties to an agreement to organize it in this way—an agreement that holds throughout our speech community

> and is codified in the patterns of our language. The agreement is, of course, an implicit and unstated one, *but its terms are absolutely obligatory;* we cannot talk at all except by subscribing to the organization and classification of data which the agreement decrees.

What led Whorf to this radical position? He wrote that the idea first occurred to him in his work as a fire prevention engineer when he was struck by how language led workers to misconstrue dangerous situations. For example, one worker caused a serious explosion by tossing a cigarette into an "empty" drum that in fact was full of gasoline vapor. Another lit a blowtorch near a "pool of water" that was really a basin of decomposing tannery waste, which, far from being "watery," was releasing inflammable gases. Whorf's studies of American languages strengthened his conviction. For example, in Apache, *It is a dripping spring* must be expressed "As water, or springs, whiteness moves downward." "How utterly unlike our way of thinking!" he wrote.

But the more you examine Whorf's arguments, the less sense they make. Take the story about the worker and the "empty" drum. The seeds of disaster supposedly lay in the semantics of *empty,* which, Whorf claimed, means both "without its usual contents" and "null and void, empty, inert." The hapless worker, his conception of reality molded by his linguistic categories, did not distinguish between the "drained" and "inert" senses, hence, flick . . . boom! But wait. Gasoline vapor is invisible. A drum with nothing but vapor in it looks just like a drum with nothing in it at all. Surely this walking catastrophe was fooled by his eyes, not by the English language.

The example of whiteness moving downward is supposed to show that the Apache mind does not cut up events into distinct objects and actions. Whorf presented many such examples from Native American languages. The Apache equivalent of *The boat is grounded on the beach* is "It is on the beach pointwise as an event of canoe motion." *He invites people to a feast* becomes "He, or somebody, goes for eaters of cooked food." *He cleans a gun with a ramrod* is translated as "He directs a hollow moving dry spot by movement of tool." All this, to be sure, is utterly unlike our way of talking. But do we know that it is utterly unlike our way of thinking?

As soon as Whorf's articles appeared, the psycholinguists Eric Lenneberg and Roger Brown pointed out two non sequiturs in his argument. First, Whorf did not actually study any Apaches; it is not clear that he ever met one. His assertions about Apache psychology are based entirely on Apache grammar—making his argument circular. Apaches speak differently, so they must think differently. How do we know that they think differently? Just listen to the way they speak!

Second, Whorf rendered the sentences as clumsy, word-for-word translations, designed to make the literal meanings seem as odd as possible. But looking at the actual glosses that Whorf provided, I could, with equal grammatical justification, render the first sentence as the mundane "Clear stuff—water—is falling." Turning the tables, I could take the English sentence "He walks" and render it "As solitary masculinity, leggedness proceeds." Brown illustrates how strange the German mind must be, according to Whorf's logic, by reproducing Mark Twain's own translation of a speech he delivered in flawless German to the Vienna Press Club:

> I am indeed the truest friend of the German language—and not only now, but from long since—yes, before twenty years already. . . . I would only some changes effect. I would only the language method—the luxurious, elaborate construction compress, the eternal parenthesis suppress, do away with, annihilate; the introduction of more than thirteen subjects in one sentence forbid; the verb so far to the front pull that one it without a telescope discover can. With one word, my gentlemen, I would your beloved language simplify so that, my gentlemen, when you her for prayer need, One her yonder-up understands.
>
> . . . I might gladly the separable verb also a little bit reform. I might none do let what Schiller did: he has the whole history of the Thirty Years' War between the two members of a separate verb in-pushed. That has even Germany itself aroused, and one has Schiller the permission refused the History of the Hundred Years' War to compose—God be it thanked! After all these reforms established be will, will the German language the noblest and the prettiest on the world be.

Among Whorf's "kaleidoscopic flux of impressions," color is surely the most eye-catching. He noted that we see objects in different hues, depending on the wavelengths of the light they reflect, but that physicists tell us that wavelength is a continuous dimension with nothing delineating red, yellow, green, blue, and so on. Languages differ in their inventory of color words: Latin lacks generic "gray" and "brown"; Navajo collapses blue and green into one word; Russian has distinct words for dark blue and sky blue; Shona speakers use one word for the yellower greens and the greener yellows, and a different one for the bluer greens and the nonpurplish blues. You can fill in the rest of the argument. It is language that puts the frets in the spectrum; Julius Caesar would not know shale from Shinola.

But although physicists see no basis for color boundaries, physiologists do. Eyes do not register wavelength the way a thermometer registers temperature. They contain three kinds of cones, each with a different pigment, and the cones are wired to neurons in a way that makes the neurons respond best to red patches against a green background or vice versa, blue against yellow, black against white. No matter how influential language might be, it would seem preposterous to a physiologist that it could reach down into the retina and rewire the ganglion cells.

Indeed, humans the world over (and babies and monkeys, for that matter) color their perceptual worlds using the same palette, and this constrains the vocabularies they develop. Although languages may disagree about the wrappers in the sixty-four crayon box—the burnt umbers, the turquoises, the fuchsias—they agree much more on the wrappers in the eight-crayon box—the fire-engine reds, grass greens, lemon yellows. Speakers of different languages unanimously pick these shades as the best examples of their color words, as long as the language has a color word in that general part of the spectrum. And where languages do differ in their color words, they differ predictably, not according to the idiosyncratic tastes of some word-coiner. Languages are organized a bit like the Crayola product line, the fancier ones adding colors to the more basic ones. If a language has only two color words, they are for black and white (usually encompassing dark and light, respectively). If it has three, they are for black, white, and red; if four, black, white, red, and either yellow or green. Five adds in both yellow and green; six, blue; seven, brown; more than seven, purple, pink, orange, or gray. But the clinching experiment was carried out in the New Guinea highlands with the Grand Valley Dani, a people speaking one of the black-and-white languages. The psychologist Eleanor Rosch found that the Dani were quicker at learning a new color category that was based on fire-engine red than a category based on an off-red. The way we see colors determines how we learn words for them, not vice versa.

The fundamentally different Hopi concept of time is one of the more startling claims about how minds can vary. Whorf wrote that the Hopi language contains "no words, grammatical forms, constructions, or expressions that refer directly to what we call 'time,' or to past, or future, or to enduring or lasting." He suggested, too, that the Hopi had "no general notion or intuition of TIME as a smooth flowing continuum in which everything in the universe proceeds at an equal rate, out of a future, through a present, into a past." According to Whorf, they did not conceptualize events as being like points, or lengths of time like days as countable things. Rather, they seemed to focus on change and process itself, and on psychological distinctions between presently known, mythical, and conjecturally distant. The Hopi also had little interest in "exact sequences, dating, calendars, chronology."

What, then, are we to make of the following sentence translated from Hopi?

> Then indeed, the following day, quite early in the morning at the hour when people pray to the sun, around that time then he woke up the girl again.

Perhaps the Hopi are not as oblivious to time as Whorf made them out to be. In his extensive study of the Hopi, the anthropologist Ekkehart Malotki, who reported this sentence, also showed that Hopi speech contains tense, metaphors for time, units of time (including days, numbers of days, parts of the day, yesterday and tomorrow, days of the week, weeks, months, lunar phases, seasons, and the year), ways to quantify units of time, and words like "ancient," "quick," "long time," and "finished." Their culture keeps records with sophisticated methods of dating, including a horizon-based sun calendar, exact ceremonial day sequences, knotted calendar strings, notched calendar sticks, and several devices for timekeeping using the principle of the sundial. No one is really sure how Whorf came up with his outlandish claims, but his limited, badly analyzed sample of Hopi speech and his long-time leanings toward mysticism must have contributed.

Speaking of anthropological canards, no discussion of language and thought would be complete without the Great Eskimo Vocabulary Hoax. Contrary to popular belief, the Eskimos do not have more words for snow than do speakers of English. They do not have four hundred words for snow, as it has been claimed in print, or two hundred, or one hundred, or forty-eight, or even nine. One dictionary puts the figure at two. Counting generously, experts can come up with about a dozen, but by such standards English would not be far behind, with *snow, sleet, slush, blizzard, avalanche, hail, hardpack, powder, flurry, dusting,* and a coinage of Boston's WBZ-TV meteorologist Bruce Schwoegler, *snizzling*.

Where did the myth come from? Not from anyone who has actually studied the Yupik and Inuit-Inupiaq families of polysynthetic languages spoken from Siberia to Greenland. The anthropologist Laura Martin has documented how the story grew like an urban legend, exaggerated with each retelling. In 1911 Boas casually mentioned that Eskimos used four unrelated word roots for snow. Whorf embellished the count to seven and implied that there were more. His article was widely reprinted, then cited in textbooks and popular books on language, which led to successively inflated estimates in other textbooks, articles, and newspaper columns of Amazing Facts.

The linguist Geoffrey Pullum, who popularized Martin's article in his essay "The Great Eskimo Vocabulary Hoax," speculates about why the story got so out of control: "The alleged lexical extravagance of the Eskimos comports so well with the many other facets of their polysynthetic perversity: rubbing noses; lending their wives to strangers; eating raw seal blubber; throwing Grandma out to be eaten by polar bears." It is an ironic twist. Linguistic relativity came out of the Boas school, as part of a campaign to show that nonliterate cultures were as complex and sophisticated as European ones. But the supposedly mind-broadening anecdotes owe their appeal to a patronizing willingness to treat other cultures' psychologies as weird and exotic compared to our own. As Pullum notes,

> Among the many depressing things about this credulous transmission and elaboration of a false claim is that even if there *were* a large number of roots for different snow types in some Arctic language, this would *not,* objectively, be intellectually interesting; it would be a most mundane and unremarkable fact. Horsebreeders have various

> names for breeds, sizes, and ages of horses; botanists have names for leaf shapes; interior decorators have names for shades of mauve; printers have many different names for fonts (Carlson, Garamond, Helvetica, Times Roman, and so on), naturally enough. . . . Would anyone think of writing about printers the same kind of slop we find written about Eskimos in bad linguistics textbooks? Take [the following] random textbook . . ., with its earnest assertion "It is quite obvious that in the culture of the Eskimos . . . snow is of great enough importance to split up the conceptual sphere that corresponds to one word and one thought in English into several distinct classes . . ." Imagine reading: "It is quite obvious that in the culture of printers . . . fonts are of great enough importance to split up the conceptual sphere that corresponds to one word and one thought among non-printers into several distinct classes . . ." Utterly boring, even if true. Only the link to those legendary, promiscuous, blubber-gnawing hunters of the ice-packs could permit something this trite to be presented to us for contemplation.

If the anthropological anecdotes are bunk, what about controlled studies? The thirty-five years of research from the psychology laboratory is distinguished by how little it has shown. Most of the experiments have tested banal "weak" versions of the Whorfian hypothesis, namely that words can have some effect on memory or categorization. Some of these experiments have actually worked, but that is hardly surprising. In a typical experiment, subjects have to commit paint chips to memory and are tested with a multiple-choice procedure. In some of these studies, the subjects show slightly better memory for colors that have readily available names in their language. But even colors without names are remembered fairly well, so the experiment does not show that the colors are remembered by verbal labels alone. All it shows is that subjects remembered the chips in two forms, a nonverbal visual image and a verbal label, presumably because two kinds of memory, each one fallible, are better than one. In another type of experiment subjects have to say which two out of three color chips go together; they often put the ones together that have the same name in their language. Again, no surprise. I can imagine the subjects thinking to themselves, "Now how on earth does this guy expect me to pick two chips to put together? He didn't give me any hints, and they're all pretty similar. Well, I'd probably call those two 'green' and that one 'blue,' and that seems as good a reason to put them together as any." In these experiments, language is, technically speaking, influencing a form of thought in some way, but so what? It is hardly an example of incommensurable world views, or of concepts that are nameless and therefore unimaginable, or of dissecting nature along lines laid down by our native languages according to terms that are absolutely obligatory.

The only really dramatic finding comes from the linguist and now Swarthmore College president Alfred Bloom in his book *The Linguistic Shaping of Thought.* English grammar, says Bloom, provides its speakers with the subjunctive construction: *If John were to go to the hospital, he would meet Mary.* The subjunctive is used to express "counterfactual" situations, events that are known to be false but entertained as hypotheticals. (Anyone familiar with Yiddish knows a better example, the ultimate riposte to someone reasoning from improbable premises: *Az di bobe volt gehat beytsim volt zi geven mayn zeyde,* "If my grandmother had balls, she'd be my grandfather.") Chinese, in contrast, lacks a subjunctive and any other simple grammatical construction that directly expresses a counterfactual. The thought must be expressed circuitously, something like "If John is going to the hospital . . . but he is not going to the hospital . . . but if he is going, he meets Mary."

Bloom wrote stories containing sequences of implications from a counterfactual premise and gave them to Chinese and American students. For example, one story said, in outline, "Bier was an eighteenth-century European philosopher. There was some contact between the West arid China at that time, but very few works of Chinese philosophy had been translated. Bier could not read Chinese, but if he had been able to read Chinese, he would have discovered B; what would have most influenced him would have been C; once influenced by that Chinese perspective, Bier would then have done D," and so on. The subjects were then asked to check off whether B, C, and D actually occurred. The American students gave the correct answer, no, ninety-eight percent of the time; the Chinese students gave the correct answer only seven percent of the time! Bloom concluded that the Chinese language renders its speakers unable to entertain hypothetical false worlds without great mental effort. (As far as I know, no one has tested the converse prediction on speakers of Yiddish.)

The cognitive psychologists Terry Au, Yohtaro Takano, and Lisa Liu were not exactly enchanted by these tales of the concreteness of the Oriental mind. Each one identified serious flaws in Bloom's experiments. One problem was that his stories were written in stilted Chinese. Another was that some of the science stories turned out, upon careful rereading, to be genuinely ambiguous. Chinese college students tend to have more science training than American students, and thus they were *better* at detecting the ambiguities that Bloom himself missed. When these flaws were fixed, the differences vanished.

People can be forgiven for overrating language. Words make noise, or sit on a page, for all to hear and see. Thoughts are trapped inside the head of the thinker. To know what someone else is thinking, or to talk to each other about the nature of thinking, we have to use—what else, words! It is no wonder that many commentators have trouble even conceiving of thought without words—or is it that they just don't have the language to talk about it?

As a cognitive scientist I can afford to be smug about common sense being true (thought is different from language) and linguistic determinism being a conventional absurdity. For two sets of tools now make it easier to think clearly about the whole problem. One is a body of experimental studies that break the word barrier and assess many kinds of nonverbal thought. The other is a theory of how thinking might work that formulates the questions in a satisfyingly precise way.

We have already seen an example of thinking without language: Mr. Ford, the fully intelligent aphasic discussed in Chapter 2. (One could, however, argue that his thinking abilities had been constructed before his stroke on the scaffolding of the language he then possessed.) We have also met deaf children who lack a language and soon invent one. Even more pertinent are the deaf adults occasionally discovered who lack any form of language whatsoever—no sign language, no writing, no lip reading, no speech. In her recent book *A Man Without Words,* Susan Schaller tells the story of Ildefonso, a twenty-seven-year-old illegal immigrant from a small Mexican village whom she met while working as a sign language interpreter in Los Angeles. Ildefonso's animated eyes conveyed an unmistakable intelligence and curiosity, and Schaller became his volunteer teacher and companion. He soon showed her that he had a full grasp of number: he learned to do addition on paper in three minutes and had little trouble understanding the base-ten logic behind two-digit numbers. In an epiphany reminiscent of the story of Helen Keller, Ildefonso grasped the principle of naming when Schaller tried to teach him the sign for "cat." A dam burst, and he demanded to be shown the signs for all the objects he was familiar with. Soon he was able to convey to Schaller parts of his life story: how as a child he had begged his desperately poor parents to send him to school, the kinds of crops he had picked in different states, his evasions of immigration authorities. He led Schaller to other languageless adults in forgotten corners of society. Despite their isolation from the verbal world, they displayed many abstract forms of thinking, like rebuilding broken locks, handling money, playing card games, and entertaining each other with long pantomimed narratives.

Our knowledge of the mental life of Ildefonso and other languageless adults must remain impressionistic for ethical reasons: when they surface, the first priority is to teach them language, not to study how they manage without it. But there are other languageless beings who have been studied experimentally, and volumes have been written about how they reason about space, time, objects, number, rate, causality, and categories. Let me recount three ingenious examples. One involves babies, who cannot think in words because they have not yet learned any. One involves monkeys, who cannot think in words because they are incapable of learning them. The third involves human adults, who, whether or not they think in words, claim their best thinking is done without them.

The developmental psychologist Karen Wynn has recently shown that five-month-old babies can do a simple form of mental arithmetic. She used a technique common in infant perception research. Show a baby a bunch of objects long enough, and the baby gets bored and looks away; change the scene, and if the baby notices the difference, he or she will regain interest. The methodology has shown that babies as young as five days old are sensitive to number. In one experiment, an experimenter bores a baby with an object, then occludes the object with an opaque screen. When the screen is removed, if the same object is present, the babies look for a little while, then get bored again. But if, through invisible subterfuge, two or three objects have ended up there, the surprised babies stare longer.

In Wynn's experiment, the babies were shown a rubber Mickey Mouse doll on a stage until their little eyes wandered. Then a screen came up, and a prancing hand visibly reached out from behind a curtain and placed a second Mickey Mouse behind the screen. When the screen was removed, if there were two Mickey Mouses visible (something the babies had never actually seen), the babies looked for only a few moments. But if there was only one doll, the babies were captivated—even though this was exactly the scene that had bored them before the screen was put in place. Wynn also tested a second group of babies, and this time, after the screen came up to obscure a *pair* of dolls, a hand visibly reached behind the screen and removed one of them. If the screen fell to reveal a single Mickey, the babies looked briefly; if it revealed the old scene with two, the babies had more trouble tearing themselves away. The babies must have been keeping track of how many dolls were behind the screen, updating their counts as dolls were added or subtracted. If the number inexplicably departed from what they expected, they scrutinized the scene, as if searching for some explanation.

Vervet monkeys live in stable groups of adult males and females and their offspring. The primatologists Dorothy Cheney and Robert Seyfarth have noticed that extended families form alliances like the Montagues and Capulets. In a typical interaction they observed in Kenya, one juvenile monkey wrestled another to the ground screaming. Twenty minutes later the victim's sister approached the perpetrator's sister and without provocation bit her on the tail. For the retaliator to have identified the proper target, she would have had to solve the following analogy problem: A (victim) is to B (myself) as C (perpetrator) is to X, using the correct relationship "sister of (or perhaps merely "relative of; there were not enough vervets in the park for Cheney and Seyfarth to tell).

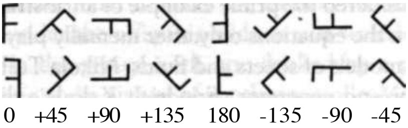
But do monkeys really know how their groupmates are related to each other, and, more impressively, do they realize that different pairs of individuals like brothers and sisters can be related in the same way? Cheney and Seyfarth hid a loudspeaker behind a bush and played tapes of a two-year-old monkey screaming. The females in the area reacted by looking at the mother of the infant who had been recorded—showing that they not only recognized the infant by its scream but recalled who its mother was. Similar abilities have been shown in the longtailed macaques that Verena Dasser coaxed into a laboratory adjoining a large outdoor enclosure. Three slides were projected: a mother at the center, one of her offspring on one side, and an unrelated juvenile of the same age and sex on the other. Each screen had a button under it. After the monkey had been trained to press a button under the offspring slide, it was tested on pictures of other mothers in the group, each one flanked by a picture of that mother's offspring and a picture of another juvenile. More than ninety percent of the time the monkey picked the offspring. In another test, the monkey was shown two slides, each showing a pair of monkeys, and was trained to press a button beneath the slide showing a particular mother and her juvenile daughter. When presented with slides of new monkeys in the group, the subject monkey always picked the mother-and-offspring pair, whether the offspring was male, female, infant, juvenile, or adult. Moreover, the monkeys appeared to be relying not only on physical resemblance between a given pair of monkeys, or on the sheer number of hours they had previously spent together, as the basis for recognizing they were kin, but on something more subtle in the history of their interaction. Cheney and Seyfarth, who work hard at keeping track of who is related to whom in what way in the groups of animals they study, note that monkeys would make excellent primatologists.

Many creative people insist that in their most inspired moments they think not in words but in mental images. Samuel Taylor Coleridge wrote that visual images of scenes and words once appeared involuntarily before him in a dreamlike state (perhaps opium-induced). He managed to copy the first forty lines onto paper, resulting in the poem we know as "Kubla Khan," before a knock on the door shattered the images and obliterated forever what would have been the rest of the poem. Many contemporary novelists, like Joan Didion, report that their acts of creation begin not with any notion of a character or a plot but with vivid mental pictures that dictate their choice of words. The modern sculptor James Surls plans his projects lying on a couch listening to music; he manipulates the sculptures in his mind's eye, he says, putting an arm on, taking an arm off, watching the images roll and tumble.

Physical scientists are even more adamant that their thinking is geometrical, not verbal. Michael Faraday, the originator of our modern conception of electric and magnetic fields, had no training in mathematics but arrived at his insights by visualizing lines of force as narrow tubes curving through space. James Clerk Maxwell formalized the concepts of electromagnetic fields in a set of mathematical equations and is considered the prime example of an abstract theoretician, but he set down the equations only after mentally playing with elaborate imaginary models of sheets and fluids. Nikola Tesla's idea for the electrical motor and generator, Friedrich Kekule's discovery of the benzene ring that kicked off modern organic chemistry, Ernest Lawrence's conception of the cyclotron, James Watson and Francis Crick's discovery of the DNA double helix—all came to them in images. The most famous self-described visual thinker is Albert Einstein, who arrived at some of his insights by imagining himself riding a beam of light and looking back at a clock, or dropping a coin while standing in a plummeting elevator. He wrote:

> The psychical entities which seem to serve as elements in thought are certain signs and more or less clear images which can be "voluntarily" reproduced and combined. . . . This combinatory play seems to be the essential feature in productive thought—before there is any connection with logical construction in words or other kinds of signs which can be communicated to others. The above-mentioned elements are, in my case, of visual and some muscular type. Conventional words or other signs have to be sought for laboriously only in a secondary state, when the mentioned associative play is sufficiently established and can be reproduced at will.

Another creative scientist, the cognitive psychologist Roger Shepard, had his own moment of sudden visual inspiration, and it led to a classic laboratory demonstration of mental imagery in mere mortals. Early one morning, suspended between sleep and awakening in a state of lucid consciousness, Shepard experienced "a spontaneous kinetic image of three-dimensional structures majestically turning in space." Within moments and before fully awakening, Shepard had a clear idea for the design of an experiment. A simple variant of his idea was later carried out with his then-student Lynn Cooper. Cooper and Shepard flashed thousands of slides, each showing a single letter of the alphabet, to their long-suffering student volunteers. Sometimes the letter was upright, but sometimes it was tilted or mirror-reversed or both. As an example, here are the sixteen versions of the letter F:



The subjects were asked to press one button if the letter was normal (that is, like one of the letters in the top row of the diagram), another if it was a mirror image (like one of the letters in the bottom row). To do the task, the subjects had to compare the letter in the slide against some memory record of what the normal version of the letter looks like right-side up. Obviously, the right-side-up slide (0 degrees) is the quickest, because it matches the letter in memory exactly, but for the other orientations, some mental transformation to the upright is necessary first. Many subjects reported that they, like the famous sculptors and scientists, "mentally rotated" an image of the letter to the upright. By looking at the reaction times, Shepard and Cooper showed that this introspection was accurate. The upright letters were fastest, followed by the 45 degree letters, the 90 degree letters, and the 135 degree letters, with the 180 degree (upside-down) letters the slowest. In other words, the farther the subjects had to mentally rotate the letter, the longer they took. From the data, Cooper and Shepard estimated that letters revolve in the mind at a rate of 56 RPM.

Note that if the subjects had been manipulating something resembling *verbal descriptions* of the letters, such as "an upright spine with one horizontal segment that extends rightwards from the top and another horizontal segment that extends rightwards from the middle," the results would have been very different. Among all the topsy-turvy letters, the upside-down versions (180 degrees) should be fastest: one simply switches all the "top"s to "bottom"s and vice versa, and the "left"s to "right"s and vice versa, and one has a new description of the shape as it would appear right-side up, suitable for matching against memory. Sideways letters (90 degrees) should be slower, because "top" gets changed either to "right" or to "left," depending on whether it lies clockwise (+ 90 degrees) or counterclockwise (— 90 degrees) from the upright. Diagonal letters (45 and 135 degrees) should be slowest, because every word in the description has to be replaced: "top" has to be replaced with either "top right" or "top left," and so on. So the order of difficulty should be 0, 180, 90, 45, 135, not the majestic rotation of 0, 45, 90, 135, 180 that Cooper and Shepard saw in the data. Many other experiments have corroborated the idea that visual thinking uses not language but a mental graphics system, with operations that rotate, scan, zoom, pan, displace, and fill in patterns of contours.

What sense, then, can we make of the suggestion that images, numbers, kinship relations, or logic can be represented in the brain without being couched in words? In the first half of this century, philosophers had an answer: none. Reifying thoughts as things in the head was a logical error, they said. A picture or family tree or number in the head would require a little man, a homunculus, to look at it. And what would be inside *his* head—even smaller pictures, with an even smaller man looking at them? But the argument was unsound. It took Alan Turing, the brilliant British mathematician and philosopher, to make the idea of a mental representation scientifically respectable. Turing described a hypothetical machine that could be said to engage in reasoning. In fact this simple device, named a Turing Machine in his honor, is powerful enough to solve any problem that any computer, past, present, or future, can solve. And it clearly uses an internal symbolic representation—a kind of mentalese—without requiring a little man or any occult processes. By looking at how a Turing machine works, we can get a grasp of what it would mean for a human mind to think in mentalese as opposed to English.

In essence, to reason is to deduce new pieces of knowledge from old ones. A simple example is the old chestnut from introductory logic: if you know that Socrates is a man and that all men are mortal, you can figure out that Socrates is mortal. But how could a hunk of matter like a brain accomplish this feat? The first key idea is a *representation:* a physical object whose parts and arrangement correspond piece for piece to some set of ideas or facts. For example, the pattern of ink on this page

Socrates isa man

is a representation of the idea that Socrates is a man. The shape of one group of ink marks, Socrates, is a symbol that stands for the concept of Socrates. The shape of another set of ink marks, isa, stands for the concept of being an instance of, and the shape of the third, man, stands for the concept of man. Now, it is crucial to keep one thing in mind. I have put these ink marks in the shape of English words as a courtesy to you, the reader, so that you can keep them straight as we work through the example. But all that really matters is that they have different shapes. I could have used a star of David, a smiley face, and the Mercedes-Benz logo, as long as I used them consistently.

Similarly, the fact that the Socrates ink marks are to the left of the isa ink marks on the page, and the man ink marks are to the right, stands for the idea that Socrates is a man. If I change any part of the representation, like replacing isa with isasonofa, or flipping the positions of Socrates and man, we would have a representation of a different idea. Again, the left-to-right English order is just a mnemonic device for your convenience. I could have done it right-to-left or up-and-down, as long as I used that order consistently.

Keeping these conventions in mind, now imagine that the page has a second set of ink marks, representing the proposition that every man is mortal:

```
Socrates isa man
Every man ismortal
```

To get reasoning to happen, we now need a *processor*. A processor is not a little man (so one needn't worry about an infinite regress of homunculi inside homunculi) but something much stupider: a gadget with a fixed number of reflexes. A processor can react to different pieces of a representation and do something in response, including altering the representation or making new ones. For example, imagine a machine that can move around on a printed page. It has a cutout in the shape of the letter sequence isa, and a light sensor that can tell when the cutout is superimposed on a set of ink marks in the exact shape of the cutout. The sensor is hooked up to a little pocket copier, which can duplicate any set of ink marks, either by printing identical ink marks somewhere else on the page or by burning them into a new cutout.

Now imagine that this sensor-copier-creeper machine is wired up with four reflexes. First, it rolls down the page, and whenever it detects some isa ink marks, it moves to the left, and copies the ink marks it finds there onto the bottom left corner of the page. Let loose on our page, it would create the following:

```
Socrates isa man
Every man ismortal



Socrates
```

Its second reflex, also in response to finding an isa, is to get itself to the right of that isa and copy any ink marks it finds there into the holes of a new cutout. In our case, this forces the processor to make a cutout in the shape of man. Its third reflex is to scan down the page checking for ink marks shaped like Every, and if it finds some, seeing if the ink marks to the right align with its new cutout. In our example, it finds one: the man in the middle of the second line. Its fourth reflex, upon finding such a match, is to move to the right and copy the ink marks it finds there onto the bottom center of the page. In our example, those are the ink marks ismortal. If you are following me, you'll see that our page now looks like this:

```
Socrates isa man
Every man ismortal





Socrates  ismortal
```

A primitive kind of reasoning has taken place. Crucially, although the gadget and the page it sits on collectively display a kind of intelligence, there is nothing in either of them that is itself intelligent. Gadget and page are just a bunch of ink marks, cutouts, photocells, lasers, and wires. What makes the whole device smart is the exact *correspondence* between the logician's rule "If X is a Y and all Y's are Z, then X is Z" and the way the device scans, moves, and prints. Logically speaking, "X is a Y" means that what is true of Y is also true of X, and mechanically speaking, X isa Y causes what is printed next to the Y to be also printed next to the X. The machine, blindly following the laws of physics, just responds to the shape of the ink marks isa (without understanding what it means to us) and copies other ink marks in a way that ends up mimicking the operation of the logical rule. What makes it "intelligent" is that the sequence of sensing and moving and copying results in its printing a representation of a conclusion that is true if and only if the page contains representations of premises that are true. If one gives the device as much paper as it needs, Turing showed, the machine can do anything that any computer can do—and perhaps, he conjectured, anything that any physically embodied mind can do.

Now, this example uses ink marks on paper as its representation and a copying-creeping-sensing machine as its processor. But the representation can be in any physical medium at all, as long as the patterns are used consistently. In the brain, there might be three groups of neurons, one used to represent the individual that the proposition is about (Socrates, Aristotle, Rod Stewart, and so on), one to represent the logical relationship in the proposition (is a, is not, is like, and so on), and one to represent the class or type that the individual is being categorized as (men, dogs, chickens, and so on). Each concept would correspond to the firing of a particular neuron; for example, in the first group of neurons, the fifth neuron might fire to represent Socrates and the seventeenth might fire to represent Aristotle; in the third group, the eighth neuron might fire to represent men, the twelfth neuron might fire to represent dogs. The processor might be a network of other neurons feeding into these groups, connected together in such a way that it reproduces the firing pattern in one group of neurons in some other group (for example, if the eighth neuron is firing in group 3, the processor network would turn on the eighth neuron in some fourth group, elsewhere in the brain). Or the whole thing could be done in silicon chips. But in all three cases the principles are the same. The way the elements in the processor are wired up would cause them to sense and copy pieces of a representation, and to produce new representations, in a way that mimics the rules of reasoning. With many thousands of representations and a set of somewhat more sophisticated processors (perhaps different kinds of representations and processors for different kinds of thinking), you might have a genuinely intelligent brain or computer. Add an eye that can detect certain contours in the world and turn on representations that symbolize them, and muscles that can act on the world whenever certain representations symbolizing goals are turned on, and you have a behaving organism (or add a TV camera and set of levers and wheels, and you have a robot).

This, in a nutshell, is the theory of thinking called "the physical symbol system hypothesis" or the "computational" or "representational" theory of mind. It is as fundamental to cognitive science as the cell doctrine is to biology and plate tectonics is to geology. Cognitive psychologists and neuroscientists are trying to figure out what kinds of representations and processors the brain has. But there are ground rules that must be followed at all times: no little men inside, and no peeking. The representations that one posits in the mind have to be arrangements of symbols, and the processor has to be a device with a fixed set of reflexes, period. The combination, acting all by itself, has to produce the intelligent conclusions. The theorist is forbidden to peer inside and "read" the symbols, "make sense" of them, and poke around to nudge the device in smart directions like some deus ex machina.

Now we are in a position to pose the Whorfian question in a precise way. Remember that a representation does not have to look like English or any other language; it just has to use symbols to represent concepts, and arrangements of symbols to represent the logical relations among them, according to some consistent scheme. But though internal representations in an English speaker's mind don't *have* to look like English, they *could,* in principle, look like English—or like whatever language the person happens to speak. So here is the question: Do they in fact? For example, if we know that Socrates is a man, is it because we have neural patterns that correspond one-to-one to the English words *Socrates, is, a,* and *man,* and groups of neurons in the brain that correspond to the subject of an English sentence, the verb, and the object, laid out in that order? Or do we use some other code for representing concepts and their relations in our heads, a language of thought or mentalese that is not the same as any of the world's languages? We can answer this question by seeing whether English sentences embody the information that a processor would need to perform valid sequences of reasoning—without requiring any fully intelligent homunculus inside doing the "understanding."

The answer is a clear no. English (or any other language people speak) is hopelessly unsuited to serve as our internal medium of computation. Consider some of the problems.

The first is ambiguity. These headlines actually appeared in newspapers:

Child's Stool Great for Use in Garden
Stud Tires Out
Stiff Opposition Expected to Casketless Funeral Plan
Drunk Gets Nine Months in Violin Case
Iraqi Head Seeks Arms
Queen Mary Having Bottom Scraped
Columnist Gets Urologist in Trouble with His Peers

Each headline contains a word that is ambiguous. But surely the thought underlying the word is *not* ambiguous; the writers of the headlines surely knew which of the two senses of the words *stool, stud,* and *stiff* they themselves had in mind. And if there can be two thoughts corresponding to one word, thoughts can't be words.

The second problem with English is its lack of logical explicitness. Consider the following example, devised by the computer scientist Drew McDermott:

Ralph is an elephant.
Elephants live in Africa.
Elephants have tusks.

Our inference-making device, with some minor modifications to handle the English grammar of the sentences, would deduce "Ralph lives in Africa" and "Ralph has tusks." This sounds fine but isn't. Intelligent you, the reader, knows that the Africa that Ralph lives in is the same Africa that all the other elephants live in, but that Ralph's tusks are his own. But the symbol-copier-creeper-sensor that is supposed to be a model of you *doesn't* know that, because the distinction is nowhere to be found in any of the statements. If you object that this is just common sense, you would be right—but it's common sense that we're trying to account for, and English sentences do not embody the information that a processor needs to carry out common sense.

A third problem is called "co-reference." Say you start talking about an individual by referring to him as *the tall blond man with one black shoe*. The second time you refer to him in the conversation you are likely to call him *the man;* the third time, just *him*. But the three expressions do not refer to three people or even to three ways of thinking about a single person; the second and third are just ways of saving breath. Something in the brain must treat them as the same thing; English isn't doing it.

A fourth, related problem comes from those aspects of language that can only be interpreted in the context of a conversation or text—what linguists call "deixis." Consider articles like *a* and *the*. What is the difference between *killed a policeman* and *killed the policeman?* Only that in the second sentence, it is assumed that some specific policeman was mentioned earlier or is salient in the context. Thus in isolation the two phrases are synonymous, but in the following contexts (the first from an actual newspaper article) their meanings are completely different:

> A policeman's 14-year-old son, apparently enraged after being disciplined for a bad grade, opened fire from his house, *killing a policeman* and wounding three people before he was shot dead.
>
> A policeman's 14-year-old son, apparently enraged after being disciplined for a bad grade, opened fire from his house, *killing the policeman* and wounding three people before he was shot dead.

Outside of a particular conversation or text, then, the words *a* and *the* are quite meaningless. They have no place in one's permanent mental database. Other conversation-specific words like *here, there, this, that, now, then, I, me, my, her, we,* and *you* pose the same problems, as the following old joke illustrates:

> First guy: I didn't sleep with my wife before we were married, did you?
>
> Second guy: I don't know. What was her maiden name?

A fifth problem is synonymy. The sentences

> Sam sprayed paint onto the wall.
> Sam sprayed the wall with paint.
> Paint was sprayed onto the wall by Sam.
> The wall was sprayed with paint by Sam.

refer to the same event and therefore license many of the same inferences. For example, in all four cases, one may conclude that the wall has paint on it. But they are four distinct arrangements of words. You know that they mean the same thing, but no simple processor, crawling over them as marks, would know that. Something else that is not one of those arrangements of words must be representing the single event that you know is common to all four. For example, the event might be represented as something like

(Sam spray paint$_i$) cause (paint$_i$ go to (on wall))

—which, assuming we don't take the English words seriously, is not too far from one of the leading proposals about what mentalese looks like.

These examples (and there are many more) illustrate a single important point. The representations underlying thinking, on the one hand, and the sentences in a language, on the other, are in many ways at cross-purposes. Any particular thought in our head embraces a vast amount of information. But when it comes to communicating a thought to someone else, attention spans are short and mouths are slow. To get information into a listener's head in a reasonable amount of time, a speaker can encode only a fraction of the message into words and must count on the listener to fill in the rest. But *inside a single bead,* the demands are different. Air time is not a limited resource: different parts of the brain are connected to one another directly with thick cables that can transfer huge amounts of information quickly. Nothing can be left to the imagination, though, because the internal representations *are* the imagination.

We end up with the following picture. People do not think in English or Chinese or Apache; they think in a language of thought. This language of thought probably looks a bit like all these languages; presumably it has symbols for concepts, and arrangements of symbols that correspond to who did what to whom, as in the paint-spraying representation shown above. But compared with any given language, mentalese must be richer in some ways and simpler in others. It must be richer, for example, in that several concept symbols must correspond to a given English word like *stool* or *stud.* There must be extra paraphernalia that differentiate logically distinct kinds of concepts, like Ralph's tusks versus tusks in general, and that link different symbols that refer to the same thing, like *the tall blond man with one black shoe* and *the man.* On the other hand, mentalese must be simpler than spoken languages; conversation-specific words and constructions (like *a* and *the)* are absent, and information about pronouncing words, or even ordering them, is unnecessary. Now, it could be that English speakers think in some kind of simplified and annotated quasi-English, with the design I have just described, and that Apache speakers think in a simplified and annotated quasi-Apache. But to get these languages of thought to subserve reasoning properly, they would have to look much more like each other than either one does to its spoken counterpart, and it is likely that they are the same: a universal mentalese.

Knowing a language, then, is knowing how to translate mentalese into strings of words and vice versa. People without a language would still have mentalese, and babies and many nonhuman animals presumably have simpler dialects. Indeed, if babies did not have a mentalese to translate to and from English, it is not clear how learning English could take place, or even what learning English would mean.

So where does all this leave Newspeak? Here are my predictions for the year 2050. First, since mental life goes on independently of particular languages, concepts of freedom and equality will be thinkable even if they are nameless. Second, since there are far more concepts than there are words, and listeners must always charitably fill in what the speaker leaves unsaid, existing words will quickly gain new senses, perhaps even regain their original senses. Third, since children are not content to reproduce any old input from adults but create a complex grammar that can go beyond it, they would creolize Newspeak into a natural language, possibly in a single generation. The twenty-first-century toddler may be Winston Smith's revenge.