FOUR

# FITTING THE ARTIFACT TO THE PERSON

YEARS AGO, WHEN I WAS STUDYING HOW PEOPLE REMEMBER THINGS, I used to ask college students to sit in small, soundproof rooms and listen to long lists of spoken digits. After each list, I would "probe" their memory for the digits by presenting a single digit and asking them to tell me what digit had followed the probe in the list. I got lots of interesting data about the limitations of what is now called "working memory" and what, at the time, my collaborator Nancy Waugh and I called "primary memory." I remember how upset I was when I spied one of my experimental subjects writing down the list of digits, then answering the probe question by reviewing the written list. I was so upset that I immediately ordered her out of the room and out of my experiment: She was cheating! She was upset that I was upset: She was getting them all correct—I should have been pleased, she said, not upset.

Today I tend to agree with the student. After all, I had asked her to do a meaningless task, so she had adopted the sensible, intelligent response. Who but an experimental psychologist would expect anyone to remember anything as silly as unrelated digits without the aid of paper and pencil? Better yet, why would anyone ever have to remember such sorts of things without writing them down? The mind is well equipped to retain large amounts of meaningful material, as long as the material has pattern and structure. It is the meaningless, arbitrary stuff of modern life that gives so much

trouble. Sure, it is often easier to remember something than to carry around and consult written records for everyday tasks, but why must these things be so meaningless, so arbitrary? Most are arbitrary requirements of today's technology. Most, perhaps all, could be dispensed with through appropriate design. It is quite possible to devise a world in which people learn things because they want to, for convenience and privacy, not because they must. Sensible, meaningful things.

Meanwhile, in our technological, machine-centered world, it really does make sense to remember things by writing them down. Why not? Human working memory is limited, so we can extend it by use of the cognitive artifact. But note: Writing something down doesn't really change our memory; rather, it changes the task from one of remembering to one of writing then, later, reading back the information. In general, artifacts don't change our cognitive abilities; they change the tasks we do.

There are two views of a cognitive artifact: the *personal* point of view (the impact the artifact has for the individual person) and the *system* point of view (how the artifact + person, as a system, is different from the cognitive abilities of the person alone). From a person's *personal* point of view, artifacts don't make us smarter or make us have better memories; they change the task. From the *system* point of view, the person + artifact is more powerful than either alone. Performance of the *system* of person + artifact is indeed enhanced, but that of the individual person is not.

The *personal* point of view:
Artifacts change the task.

The *system* point of view:
The person + artifact is smarter than either alone.

An artifact is not a simple aid. That is, you can't just go out and find some cognitive artifact, and there you are, better at something. Nope, most cognitive artifacts present you with yet another thing to be learned, another manual to be read, another course to be taken, or another period of slow, tedious learning to endure.



Reading, writing, and arithmetic are perhaps our most powerful cognitive skills, but these mental artifacts take years to be learned. Not everyone fully masters them. The study of artifacts is also the study of human capabilities. Why is it so hard for some people to learn these skills? Are there better ways of teaching them? And most important to me, what is it that makes some artifacts effective, others not? Could we develop a science of artifact design that would tell us how to make better artifacts, perhaps ones that were easier to learn and use?

## SURFACE AND INTERNAL REPRESENTATION

Once upon a time, before all this electronic and computer stuff came along with its invisible internal representations, we used to be able to see just how our artifacts worked. Everything was physically visible: gears, chains, levers, dials. We could simply move the parts of interest and tell what was going on from their position and motion.

In the modern world of electronic systems, the controls and indicators have almost no physical or spatial relationships to the device itself. As a result, we now have arbitrary or abstract relationships between the controls, the indicators, and the state of the system. This is one reason why these devices are so difficult to learn: Each one uses its own arbitrary choice of operations and methods. The abstraction possible with today's electronic devices means that there doesn't have to be any natural relationship between the appearance of an object and its state.

When a physical file folder is open, it is visibly different from when it is closed. When it is stuffed with paper, it looks different than when it is empty, even when closed. Not so with electronic files. All we can see is whatever the designer thought of providing, which is sometimes a lot, sometimes nothing. The difference is that with the physical folder, the visible properties are an automatic, intrinsic part of its existence, whereas with the electronic folder, any perceivable existence is dependent upon the goodwill and cleverness of its human designer, who provides a perceivable interpretation of the underlying invisible information structures. We

understand our artifacts by what is perceivable. With some artifacts, not enough can be perceived.

The natural visibility of artifacts divides them into two broad categories, *surface* and *internal* artifacts. The distinction has important design implications. With surface artifacts, what we see is all there is: They only have surface representations. Take this book. The only information contained here is that represented by the printed words and images: marks on the white paper. The marks are static and passive: They cannot change, unless you physically erase them. With the book and all surface artifacts, what is perceivable is all that exists.

In contrast to surface artifacts there are internal artifacts, in which part of the information is represented internally within the artifact, invisible to the user. Consider, for example, a calculator. What you see is the surface representations, the information visible on the display and the buttons that allow information and instructions to be entered. Beneath those surface representations, however, lie internal representations for the digits and operations, which are unseen by the user but can be manipulated, transformed, and otherwise modified as needed by the calculations being performed. There are even hidden representations—temporary results of the calculations, internal states used only by the calculator that are not displayed, not visible. With the calculator as with all internal artifacts, there is more than can be perceived.

Memory aids such as paper, books, and chalkboards allow for the display and relatively permanent maintenance of representations. The slide rule and abacus are examples of computational devices that only contain surface representations of their information. These devices are primarily systems for making possible the display and maintenance of symbols. I call these "surface representations" because the symbols are maintained at the visible "surface" of the device—pencil or ink marks on paper; chalk on a board; indentations in sand, clay, or wood; and so on. Some of these representations are passive: Once the information has been added, it cannot be changed by the artifact itself. Thus writing, whether on a chalkboard or printed in a book, can be changed by the user, but not by the artifact. Printed tables of reference information have this property. They are meant to be consulted, not to be changed.



Internal artifacts need interfaces, some means of transforming the information hidden within their internal representations into surface forms that can be used. This poses some important design considerations: On the one hand, it offers unlimited possibilities, for the designer can choose whatever representation makes the operation of the artifact best conform to the needs of the user, unconstrained by the physical limitations of the surface representations. On the other hand, it imposes special requirements on the designer, who must now be an expert in both the technology of the artifact and in human psychology, and for artifacts that are used by groups of people, an expert in social interaction as well. Designers never had to think about these issues before: There are few people who can deal with the broad implications of this challenge.

Artifacts that have only surface representations do not need a special interface: The surface representation itself serves as the interface. But this still does not eliminate the need for careful design. There are always alternative designs that make the artifact more or less successful in the fit between its surface representation and the needs of the person and task. Most surface artifacts do have some hidden parts, and the designer must choose which parts to hide, which to make readily available. Still, with a surface artifact, the very nature of the device guarantees some understanding by its users.

## Properties of Surface Representations

Some artifacts are passive, incapable of changing their representations without activity by their users. Thus chalkboards and pieces of paper are passive artifacts: Their users initiate all actions that change the surface representations. Some artifacts are active, capable of changing their own representations. Clocks, calculators, and computers are active artifacts, capable of changing their representations without any action by users. A mechanical clock is an active surface artifact; a computer is an active internal artifact.

A person is more like an internal artifact than a surface one: What you see is not all there is. When we interact with one another, we have to transform our thoughts into surface representations

so that others can have access to them. This means transforming those ideas into words, facial expressions, gestures, mime, action, sketches, or sounds—exploiting all the sensory capabilities in an attempt to convey our intentions to others. The human surface representations are temporary: Sound fades away, gestures and actions disappear once completed. External surface representations can overcome the limitations of human surface representations. External representations, such as marks and images, can be permanent. This isn't always true, of course: Sounds and video images are transient, lasting only for the duration of the event they signify, but with the proper artifacts, they have the virtue of being forever repeatable.

People and artifacts may have dramatically different internal representations and processing, but the surface representations must be similar or, at least, complementary. A major design problem is to get the surface representations right. Although the surface representations of the artifact must match those of the human, the internal representations need not match and are often most valuable when they differ significantly. But when they differ, the nature of the surface representation is especially critical, for this is where the person gets all the information about the usage and state of the device. Here is where the science of design begins: How shall that information be represented to be of most use?

Some of the properties of simple surface representations are easily seen by asking why a lecturer uses slides or tape recordings. If the lecture is telling of travels to exotic locations, then slides and recordings are necessary because the words of the lecturer cannot convey a complete image of the place being described: The slides and recordings provide better representations—passive, surface, experiential.

What about a lecturer in business or science, where the slides often do not contain any information different from what is also being said? Here slides can serve the cognitive processes of communication in several ways:

- *A shared workspace:* The entire audience can view and reflect upon the same information at once.



- *Cooperative work:* Because there is a shared workspace, everyone can analyze and consider the same points at the same time. Any member of the audience can raise a new question or propose a new insight.
- *Memory permanence:* The slide acts as an external memory, maintaining an accurate record of the words and concepts for as long as the image is projected: The time is under the control of the lecturer as opposed to the vagaries of human memory.
- *Memory quantity:* The slide allows the effective presentation of more information than can be kept active within a person's memory, allowing the viewer to examine different areas selectively, confident that material passed over can be quickly and easily retrieved simply by moving the eye fixation to the appropriate location.
- *Perceptual processing:* The spatial arrangement of ideas helps point out their relationship. The physical presence of the slide helps focus the listener's attention.
- *Individual differences:* Some people prefer auditory information, some visual. Some types of processing are superior when information is auditory, some when it is visual or spatial. The slide provides a redundant communication channel, allowing the listener to select whichever manner of information is easier or preferred.

The appropriate use of a slide assists the audience, for the important concepts remain available for the duration of the slide, not just the duration of the spoken word: The slide acts as an external memory. The speaker can rely on this aspect of the artifact and can refer to critical aspects by pointing to the relevant part of the slide. Even after the slide has been turned off, speakers occasionally point to the place where the information used to be, with the full expectation that the audience will understand what is being referred to, even though nothing is visible on the screen or board.

Slides can also be misused, so much so that some speakers refuse to use them. They can lead to a wandering of attention as the

audience reads the slide instead of listening to the speaker. Poor speakers may fail to synchronize their slide presentations with their speech, and if the slides are presented too quickly, the listeners must sometimes engage in rapid copying of the slide material to their own notes, thus forcing them to miss both what is being said and also the contents and implications of the very points they are trying to copy. Finally, far too many government and industrial speakers delight in giving talks with their entire text on the slides, so they read the slides aloud point by point, despite the fact that their audience can read and follow them by themselves faster than the speakers can talk. The result can be a bored and sleepy audience.

## The Tower of Hanoi, Oranges, and Coffee Cups

One of my former graduate students, Jiajie Zhang, developed a nice demonstration for his Ph.D. dissertation of how the physical characteristics of an artifact can dramatically change the ease of solving a problem. He did this with one of the favorite puzzles of cognitive scientists, the Tower of Hanoi puzzle, but the result applies to many other domains.

The Tower of Hanoi puzzle is an old familiar one to many people. Figure 4.1 shows the version Zhang studied, slightly modified from the usual puzzle.

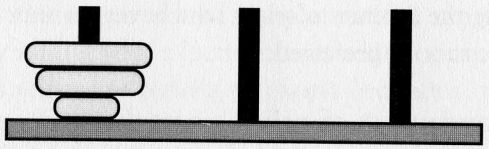


**Figure 4.1** The modified Tower of Hanoi puzzle studied by Jiajie Zhang. Three rings are placed on a peg. The goal is to move all three rings from the leftmost to the rightmost peg. Only one ring may be moved at a time, and a smaller ring may not be placed on top of a larger ring. (This is the opposite of the usual convention—usually, the big ring is on the bottom, the small one on top—but this difference doesn't change anything of importance.)

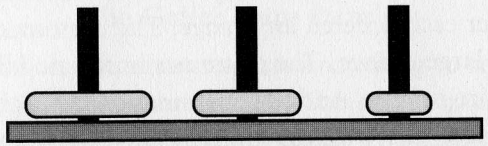


The puzzle consists of three pegs and three rings. At the start, the rings are placed on one peg in order of size (let it be the far-left peg in the drawing). Usually, the rings are arranged so that the smallest is on top, the largest on the bottom. Zhang used the reverse ordering—largest on top, smallest on the bottom—to make this version of the puzzle analogous to "the Coffee Cups puzzle" (to be described shortly). The task is to move the rings from their starting peg to another (the far-right peg in the drawing) following two rules:

*Rule 1:* Only one ring can be transferred at a time.

*Rule 2:* A ring can only be transferred to a peg on which it will be the largest.

Zhang studied a variant of the normal puzzle. First, as already noted, the rings were stacked in the reverse order of normal. Second, the ending state was changed to that of one ring per peg in the order (from left to right) large, medium, small, like this:



Finally, Zhang added a third rule:

*Rule 3:* Only the largest ring on a peg can be transferred to another peg.

The third rule makes no difference for the puzzle: If you follow rule 1, then you will only move a single ring at a time, which, given the physical structure of the pegs, means you can only move the top ring. If rule 2 has been followed, the top ring will always be the largest one on a peg. Then why add the rule? Ah, it is unnecessary in this problem only because of the physical structure of the pegs. What if there were no pegs? What if the physical structure did not provide an advantage? This is what Zhang studied: how the physical properties of the puzzle can aid in the solution.

To get at this point, Zhang invented two other versions of the Tower of Hanoi puzzle. All versions had three objects that had to be moved to three different locations; all used the same three rules; and all were formally equivalent—but they varied considerably in their difficulty. The other puzzles are called "isomorphs" of the Tower of Hanoi. An isomorph, you might remember from Chapter 3, is an equivalent problem (*iso* means "equal"), but described differently.

Here are the other two puzzle isomorphs: the Oranges and the Coffee Cups puzzles. (In the descriptions that follow, I use schematic figures to illustrate the puzzles. In his studies, Zhang used real doughnuts for the rings, real plates, and real cups filled with coffee. However, he used three different sizes of balls instead of oranges.)

### The Oranges Puzzle

A strange, exotic restaurant requires everything to be done in a special manner. Here is an example. Three customers sitting at the counter each ordered an orange. The customer on the left ordered a large orange. The customer in the middle ordered a medium-size orange. And the customer on the right ordered a small orange. The waitress brought all three oranges on one plate. She placed an empty plate in front of two of the customers and the plate with the three oranges in front of the middle customer (as shown in picture 1).

Because of the exotic style of this restaurant, the waitress had to move the oranges to the proper customers following a strange ritual. No orange was allowed to touch the surface of the table. The waitress had to use only one hand to rearrange these three oranges so that each orange would be placed on the correct plate (as shown in picture 2), following these rules:

*Rule 1:* Only one orange can be transferred at a time.

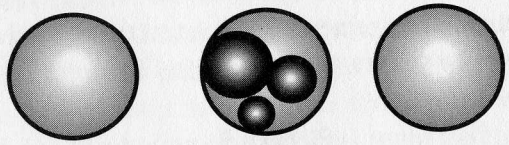
*Rule 2:* An orange can only be transferred to a plate on which it will be the largest.

*Rule 3:* Only the largest orange on a plate can be transferred to another plate.

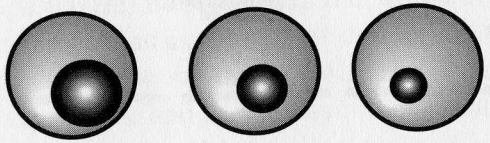


How would the waitress do this? That is, you solve the puzzle and show how the waitress has to move the oranges to go from the arrangement shown in picture 1 to the arrangement shown in picture 2.

Picture 1



Picture 2



### The Coffee Cups Puzzle

A strange, exotic restaurant requires everything to be done in a special manner. Here is an example. Three customers sitting at the counter each ordered a cup of coffee. The customer on the left ordered a large cup of coffee. The customer in the middle ordered a medium-size cup of coffee. And the customer on the right ordered a small cup of coffee. The waiter brought all three cups of coffee on one plate, piled one on top of another. He placed an empty plate in front of two of the customers and the plate with the three cups of coffee in front of the middle customer (as shown in picture 3).

Because of the exotic style of this restaurant, the waiter had to move the cups of coffee to the proper customers following a strange ritual. No cup of coffee was allowed to touch the surface of the table. The waiter had to use only one hand to rearrange these three cups of coffee so that each cup of coffee would be placed on the correct plate (as shown in picture 4), following these rules:

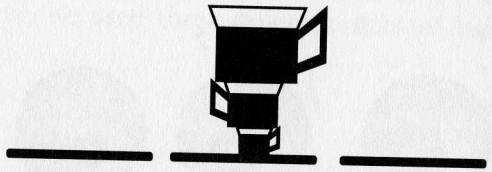
*Rule 1:* Only one cup of coffee can be transferred at a time.

*Rule 2:* A cup of coffee can only be transferred to a plate on which it will be the largest.

*Rule 3:* Only the largest cup of coffee on a plate can be transferred to another plate.

How would the waiter do this? That is, you solve the puzzle and show how the waiter has to move the cups of coffee to go from the arrangement shown in picture 3 to the arrangement shown in picture 4.

Picture 3



Picture 4

One problem involves a weird way to move oranges about, the other a weird way to move coffee cups. Otherwise, they are the same. Why is this interesting? Because these problems differed greatly in difficulty: The oranges puzzle took almost 2½ times as long as the coffee cups puzzle, with almost twice as many moves and with six times as many errors. The difference had to do with the physical constraints.

In the coffee cups puzzle, although three rules were stated, only one was necessary: rule 1, *"Only one cup of coffee can be transferred at a time."* Rules 2 and 3 were unnecessary because the cups imposed these rules by their very construction. The cups were filled with real coffee and so constructed that smaller cups would fit inside of larger ones. Rule 3 was not needed because the size of the plate was such that only one cup could fit on it: When more than



one cup was on a plate, it had to be stacked on top of the other cups on the plate. Violate rule 2, and because the only way to set more than one cup on a plate was to stack them, a small cup wouldn't stack on top of a larger one without spilling coffee. So if you only moved one cup at a time (rule 1), the physical nature of the cups meant that the only one you could move would be the one at the top of the pile, which had to be the largest.

In contrast to the Coffee Cups puzzle, where only one rule was needed, in the Oranges puzzle, all three rules were needed: There were no physical constraints to force compliance with the rules. The original Tower of Hanoi puzzle was rephrased as "the Doughnuts puzzle," with a waiter who had to deliver three different-size donuts to three customers, following exactly the same rules as for the coffee cups puzzle, but changing the words *cup of coffee* to *doughnut*. Each plate had a peg, and the donuts had to be placed on the peg in the manner shown in Figure 4.1. The doughnuts problem only needs two rules, 1 and 2, because the physical constraints of the pegs force compliance with rule 3.

So now we have three problems, all formally identical. One, the Oranges puzzle, requires three written rules. Another, the Doughnuts puzzle, requires two written rules. The third, the Coffee Cups puzzle, only requires one. The less need for rules, the easier the problem: The Coffee Cups puzzle was easiest (done most quickly, with least error), the Doughnuts next, and the Oranges hardest. Why does it matter so much whether the rules were written or were also incorporated into the physical structure of the puzzle?

External representations add power because the physical structures automatically constrain the actions and interpretations, even though all three rules apply to all the puzzles. Someone programming a computer to solve the task would find all three puzzles to be of equal difficulty and would use the same algorithm to solve all of them. This is because the computer would be unable to take advantage of the physical structures.

Why is the physical form so important to people? Zhang pointed out that the problem really had to be represented in three different ways: first, internally within the problem solver's mind; second, externally in the physical puzzle itself, where the physical

constraints play a major role; and finally, once in the mind of the scientists, who study how people solve the problem. It is the scientist who constructs an abstract representation of the problem and the possible moves toward its solution. To the scientist (or computer programmer), all three problems are the same. But to the person who can make use of the physical structures of the puzzle, the more information present in the environment, the less information needs to be maintained within the mind. As a result, for people, the three tasks are very different. In fact, people often don't recognize them as the same problem.

The Coffee Cups and Oranges puzzles may seem peculiar, but they serve as powerful demonstrations of how external representations not only aid in memory and computation but can dramatically affect the way a problem is viewed and the ease with which it can be solved.

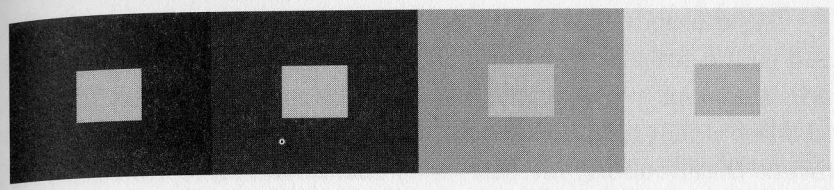
## FITTING THE REPRESENTATION TO THE TASK

Chapter 3 introduced us to the power of representation. There we discovered that the game of 15 is harder for people than ticktacktoe and that it is easier for people to determine that 284 is less than 912 than to determine that it is less than 312. Logically, the game of 15 and ticktacktoe are the same. Logically, one can determine whether 284 is less than the other alternatives simply by comparing their far left digits, so the judgments ought to be of equal difficulty. These examples make it clear that we do not operate by mathematical or symbolic logic: We operate by perceptual routines. We are especially good at making perceptual judgments, not so good at abstract or symbolic ones.

Saying that we are perceptual creatures does not, however, describe how we operate. Our perceptions are complex, not always operating the way that intuition or commonsense, folk psychology would predict. It is tempting to associate physical variables with the psychological experience. After all, making a light or sound more intense increases its brightness or loudness. Changing the frequency of a light or sound changes its hue or pitch. But to make this simple association would be wrong. The relationship between the physical and psychological dimensions is very complex.



How complex? Well, consider how bright something looks. The more light, the brighter the object, right? Wrong. Look at the following picture:



The four inner rectangles all reflect the same amount of light, even though the ones on the left look lighter than the ones on the right. This is because brightness is not the same as light intensity. Brightness depends upon the contrast between an image and surrounding images: A gray patch becomes black when it is next to something bright. Does that sound wrong? Add light to make something darker? Try it. Take this book into a dark closet and then slowly open the door a tiny amount to let some light gradually enter the closet. The rightmost inner square will get darker as you allow more light into the closet.

This principle is exploited in your television set. When the set is off, the screen is gray. Now turn the screen on and note that the image contains black areas and lines. The television set is incapable of placing black on the screen. The screen of a TV set can only emit light: Electrons hit the screen, causing the phosphors on the screen to glow red, green, or blue. A phosphor can only increase the amount of light coming from the screen, not decrease it. The black areas of the screen are created by not sending any electrons to those areas, so the amount of light from a black portion of the screen is exactly the amount that would be emitted if the set were turned off. As we have noted, the screen of the turned-off TV set looks gray, whereas in the picture, we can perceive black. How come? For the same reason that the rightmost inner square looks darker than the leftmost one: It is the contrast that matters.

All this is simply to prove that what you perceive is not necessarily what is there. The *psychology* of perception is very different from the *physics* of perception. Even in cases where more intense lights or sounds are brighter or louder, the relationship is not linear.

Double the amount of light in a room and things do not get very much brighter. This is easy to test. Illuminate a room with a single light bulb and then turn on a second bulb of equal power. You will not notice much difference in the brightness of the room.

Psychologists have determined that perceived intensity of light and sound follows roughly a cube-root law: Brightness and loudness are roughly proportional to the cube root of intensity (the actual law is that brightness or loudness is proportional to $I^{0.3}$).* This means that doubling intensity only makes the perception increase by 20 percent ($2^{0.3} = 1.2$). You have to increase the intensity ten times to make brightness or loudness double.

Most people are completely unaware of this relationship between sound intensity and loudness. Few people realize that there is any difference between what is measured physically and what is perceived psychologically. Even sound engineers, who measure sound logarithmically in decibels (dB), often do not realize that decibels are not an appropriate measure for psychological loudness. A ten times decrease in sound energy means that the number of decibels decreases by ten (−10 dB), but it halves loudness; a one hundred times decrease in sound energy decreases the decibel level by twenty (−20 dB) but makes the sound a quarter as loud; and a thousand times decrease (−30 dB) makes it an eighth as loud.

The complex relationship between sound energy and loudness is often taken advantage of by those who wish to overstate the effects of their manipulation of sound levels. If someone claims "We cut noise levels in half!" be wary: They probably cut noise *intensity* in half, which means that the resulting reduction in *loudness* is barely perceivable to the average listener.

Consider the loudness control of a radio. Here the amount you turn the knob controls sound intensity, which, in turn, affects loudness. Engineers quickly learned that the control couldn't simply control intensity, for if it did, decreasing the control from maximum to the halfway point would hardly make a difference in loudness. Engineers assumed that loudness was proportional to the logarithm of sound intensity, so they made loudness controls logarithmic. That turns out to be wrong, but at least it is better than a linear control. Alas, the engineers who designed light controls



never did learn their lessons, which is why it is sometimes hard to control room lighting or, worse, the light of an alarm clock or a clock radio. What the engineers didn't realize about human vision is that it is sensitive to light energy over a range of more than 100 billion to 1 and, moreover, that after being in the dark for half an hour, the eyes become "dark adapted," considerably increasing their sensitivity to light. This is why many clock radios that use a dim light to show the time at night have trouble getting it right: When you first turn off the lights upon going to bed, the dim light of the clock radio might be barely perceptible, but in the middle of the night, after your eyes have adapted through several hours of darkness, the same light may be bright enough to annoy. Why does all this matter? Because it makes a big difference in how information is presented to us by the surface representations of artifacts.

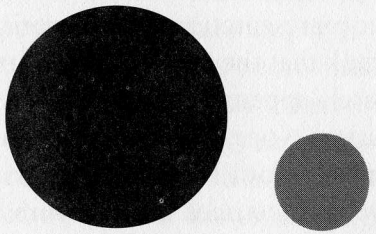
## Graphic Representations

Graphs and graphic representations are a surprisingly recent invention. You would think that the perceptual qualities of graphs would have been immediately apparent as a superior method of presenting numerical information. Nope. Graphic presentations were not used by American businesses until the late 1800s and early 1900s, and even then they were sometimes greeted with skepticism. Today they are widely appreciated for their abilities to present trends and comparisons more effectively than by words or lists of numbers. Of course, after something proves effective, it often becomes overused and applied to everything, even where not appropriate. Worse, there are a wide variety of graphic procedures, and not every procedure is appropriate to every situation.

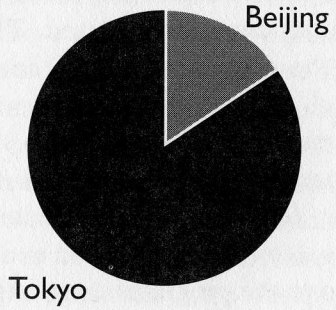
Today it is very easy to create graphs. Too easy. Software programs for the personal computer abound. There are special programs devoted to the construction and analysis of graphically presented data as well as graph-making components in a wide variety of other programs, from spreadsheets to slide-preparation programs to word processors. The problem is that these programs will graph anything, following their own internal rules of logic. Whether those rules also apply to the data and the intended use is a different question, one the programs completely ignore.

My concern is with the violation of psychological principles, whereby graphs are used in inappropriate ways, sometimes deliberately to confuse but more often out of sheer ignorance. My frequent complaints to friends, colleagues, students, and newspapers are commonly met with the excuse "My computer program did that for me automatically; I had no choice." Poor reason: Ignorance of the law is not a valid excuse, whether it be a governmental law or a psychological one.

Let me begin with one of the more common errors in the presentation of information. Suppose that we represent the cities of the world on a map, with the area of the circle that represents each city proportional to population. How well can you judge relative city sizes this way? This is, after all, a typical scheme used in newspaper maps. Well, here are two cities, Beijing and Tokyo, with their areas proportional to their projected size for the year 2000:




Clearly, Tokyo is expected to be much larger than Beijing, but by how much? Let's try again. Here are the same two cities in a "pie graph," where, once again, the area of the graph represents population:






Notice that in the pie graph, the relative dominance of Tokyo over Beijing looks many times larger than it does with the circles. Now compare the same data in bar-graph form, this time where the length of the bar is proportional to the population:


What is the actual ratio of population sizes? In the year 2000, it is estimated that Tokyo will be five times as large as Beijing (30 million versus 6 million estimated population).

Yes, we are perceptual animals, but we excel in seeing patterns, not in forming accurate numerical comparisons from those patterns. Our estimates are most accurate for line lengths—the bar graph—because judgments of line lengths are reasonably accurate, increasing linearly with the physical length of the line. Not so with most things, however, as we have already discussed for loudness and brightness, and as we have experienced with the pie and circle representations.

These distortions of perception are well known by professionals who make graphs. Unfortunately, this knowledge can also be used to mislead the viewer. Inappropriate representations can easily confuse, whether by deliberately making information difficult to discover or by deliberately emphasizing desirable features and obscuring undesirable ones. Figure 4.2 shows how inappropriate use of representational formats can yield misleading results.

The advertisement is intended to show that although the retail price of the Mercedes 300–class automobile was higher than the competition, the cost of ownership over a period of five years was much higher for the competition. How much higher? Well, if you look at the graph, the height of the competitor's bar is 2.5 times that of Mercedes: 250 percent more expensive! Well, no. If you look at the dollar amounts, note that the graph starts out at $42,000: The zero point is far away. The real ratio is 1.15 to 1: only a 15 percent difference in cost. Once the zero point has been taken away from a bar graph, the natural comparison of the line lengths is

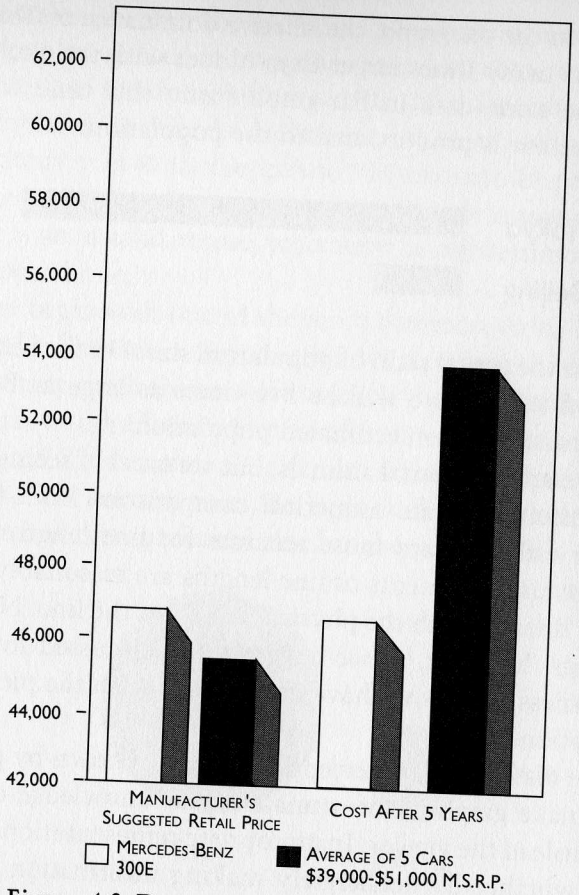



**Figure 4.2 Deceptive graphing.** Here, the black bar in the right part of the diagram is 2.5 times as high as the white one, implying that cost of ownership is 250% higher than that for Mercedes-Benz cars. But, because the zero point has been moved, the bar graphs are only showing interval scale data and comparison of the ratio of their lengths is inappropriate. The correct ratio is 1.15 to 1: only a 15% difference. (From the *Los Angeles Times, San Diego Edition* [May 21, 1991], p. D5.)

meaningless. True, the correct information is available, and in the technical sense, the graph is accurate. But visual information dominates the initial impression, and not every reader will take the time



to do the more difficult, reflective analysis of the numerical information. The dominance of the natural perceptual interpretation is commonly exploited by the advertising industry.

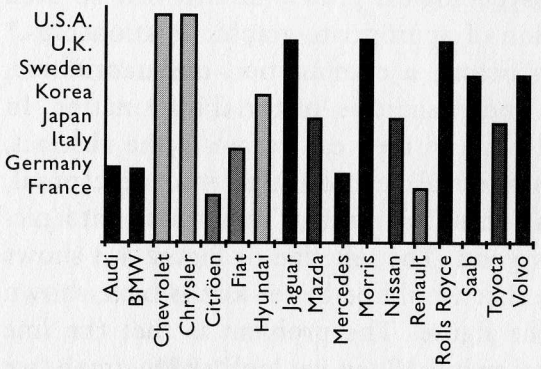
## Psychological Scales and Representation

There are numerous psychological principles that can be used to guide the construction of appropriate graphic relationships.* Graphs are interesting beasts, a combination of quantitative, numerical information and qualitative, pictorial information. In fact, graphs are useful because they can translate the abstract, difficult-to-interpret numerical relationships into perceptual, readily visible pictorial ones. But putting everything into pictures is not necessarily good. The left side of Figure 4.3 shows an example of a graph that would be better as the table shown in the right half of the figure. The problem is that the line lengths imply numerical value. When we look at the graph, we are tempted to judge that a Saab (from Sweden) is three times better than a Mercedes-Benz (from Germany) because its line is three times longer.

These examples illustrate that representations should reflect the appropriate power of the medium.

> **Appropriateness principle:** The representation used by the artifact should provide exactly the information acceptable to the task: neither more nor less.

Note that there is no "correct" way to display any particular relationship, but there are definitely incorrect ways. Remember that a representation should support both organization and search and that what is most appropriate depends upon the task to be done. The tabular representation on the right of Figure 4.3 is organized around country of origin, which makes it work well when the user's task is to start with the country and end up with the automobiles. But if the goal is to start with automobiles and determine the country, this particular tabular representation gives no assistance to the search task. It is not a good representation for looking up the automobile name.




| Country of Origin | Brand of Automobile |
|---|---|
| France | Citröen |
| | Renault |
| Germany | Audi |
| | BMW |
| | Mercedes |
| Italy | Fiat |
| Japan | Mazda |
| | Nissan |
| | Toyota |
| Korea | Hyundai |
| Sweden | Saab |
| | Volvo |
| U.K. | Jaguar |
| | Morris |
| | Rolls Royce |
| U.S.A. | Chevrolet |
| | Chrysler |
| | Ford |

**Figure 4.3 Countries of origin of automobiles.** The graph is perfectly accurate, but the use of ratio-scale line lengths to display nominal-scale information is misleading and disturbing. Most people find this graph absurd. The tabular format, shown on the right, is far more appropriate for this type of information. (The graph on the left was inspired by the work of Mackinlay [1986] who deliberately invented this example to demonstrate its futility.)

How could the display be improved? If it were known that the user would always start with the name of a country, then the table is probably near optimum. If it were known that the user would always start with the name of an automobile, then we would alphabetize the automobile list and display country of origin alongside each name. What if both search directions might be used? Here we need one of several things. One method would be a matrix organization: an alphabetical list of automobiles along one axis, an alphabetical list of countries along the other, with a check mark at relevant intersections. Or we might use a network structure, with two lists—one of automobiles, one of countries—linking together the relevant country-automobile pairs. Even here there are tradeoffs in the display: Some are more aesthetically pleasing than others,



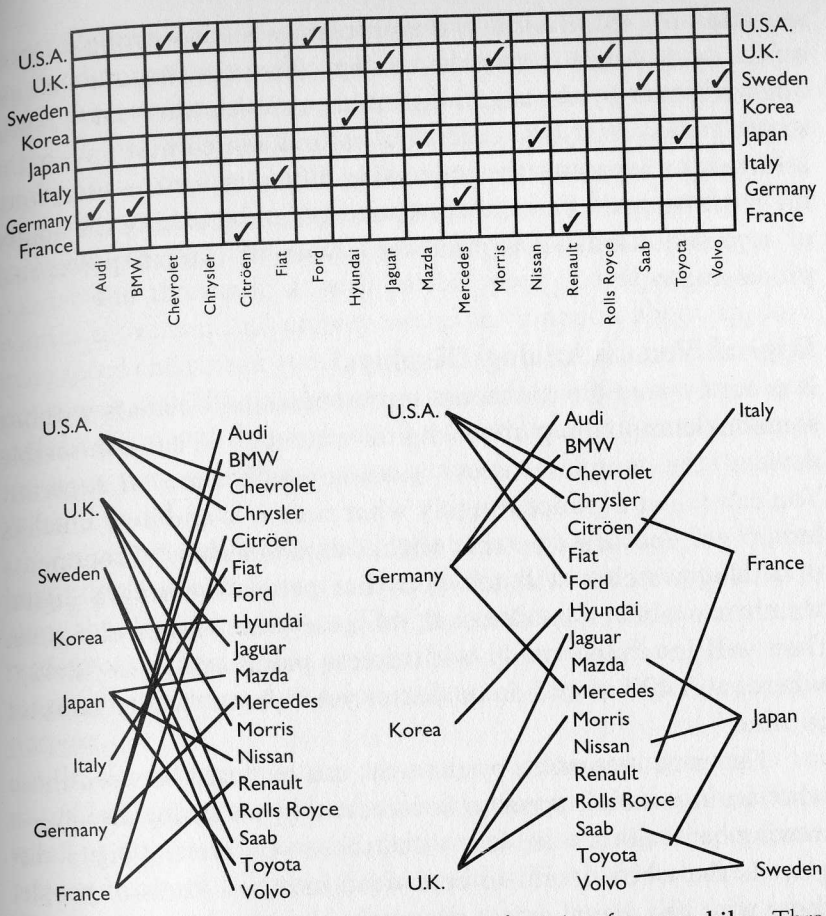



**Figure 4.4 Three ways to show countries of origin of automobiles.** Three ways of listing automobile manufacturer and country of origin in a manner that allows for symmetrical search: just as easy to go from automobile name to country as from country to automobile name. The matrix is modeled on the pill matrix of Chapter 3. However, here it is difficult to follow the long rows and columns. The association diagram in the bottom left is cluttered and hard to follow. The association diagram in the bottom right is easiest to use, except that the country names are no longer in alphabetical order and therefore more difficult to find.

some work more efficiently than others. Compare the several possibilities in Figure 4.4.

Figure 4.3 shows what happens when too powerful a repre-

sentation is used. Basically, the problem with too strong a representational system is that one tends to draw conclusions that are not warranted by the actual information. With too weak a representational format, the lack of a natural representation for the information increases the processing difficulty for the user, forcing more mental effort and computation and reducing the ability of representations to exploit the power of human perceptual processing.

## Digital Versus Analog Displays

It is rarely possible to discuss representational formats without someone complaining about digital watches. "What a miserable device," the complaint goes. "Analog watches are far superior. You can tell at a glance roughly what time it is and how much is left to go. You can't do that with a digital watch." Proponents of analog watches will tell you that people wearing a digital watch can always be detected, because when asked the time, they will invariably reply with excess precision, as in "8:42," whereas "8:40" would do or, better yet, "about twenty minutes to nine."

The same argument, by the way, can be heard between those who like automobile speed to be indicated by an analog display—a moving hand along a dial face—and those who prefer a digital display. Is the whole world to be divided into two kinds of people, those who like digital versus those who don't?

Which representation is superior? Neither: It all depends upon the task. Some representations are superior for some tasks, others for others. Let's consider the automobile speed indicator: the speedometer. Which is better, a digital or an analog display?

Answer: What do you want to know? If you want to know the exact numerical value of the speed, then digital is superior because it presents the answer in precisely the form in which you need to know it: a number. The answer is easy to read rapidly and accurately. With an analog dial, the exact value of speed is more difficult to determine: The pointer has to be interpreted to yield the exact value.

Most of the time, however, we do not need to know the exact



numerical value of our speed: Either we need a rough estimation, or we need to know whether we are above or below some critical speed, such as the legal speed limit. Here the analog speedometer is superior, but only if the dial contains a highly visible reference mark. Then you can simply look to see on which side of the mark the pointer is: Above and you are going too fast, below and you are going slower than the limit. Moreover, the distance between the pointer and the mark, as well as the speed and direction of the pointer movement, all convey useful information about approximate speed and rate of acceleration. In a similar fashion, after some experience with the speedometer, the pointer angles take on meaningful values, and the angle can be determined accurately enough for most purposes with a rapid glance, yielding an approximate value of speed.

In commercial airplanes, pilots are provided with "speed bugs," little reference marks that they can move to critical spots along the airspeed indicator. This makes it easier to transform their reading of the analog dial position into "more than" or "less than" judgments of airspeed. Analog meters are clearly superior for this purpose.

In the automobile, the advantages of the analog readout are mixed because of the lack of a flexible way of marking the critical speed of interest. As a result, neither analog nor digital speedometers seem superior. If the critical speed is 35, then with the analog dial, one must figure out where 35 might lie on the scale, then whether the pointer is above or below it. With a digital speedometer, one must mentally subtract 35 from the indicated number to determine whether one is above or below. Neither method seems particularly virtuous.

What is the best method? Why choose? Why not have both? Why not have an analog speed indicator with a digital display alongside? Today's commercial aircraft use combination analog and digital displays for airspeed and altitude, with easy ways to set "bugs" at critical values. A similar philosophy might be effective in the automobile.

Altitude meters in aircraft used to have three hands, but pilots made so many errors in reading them that today most altimeters

have replaced two of the hands with a digital readout, retaining the analog pointer for only the least significant digits. Thus if the airplane's altitude is 31,255 feet, the digital meter will display the altitude in thousands of feet as "31," and the analog meter will display the hundreds—namely, "2.55" (255 feet). This dual display has several advantages. The most significant reading—for which accurate values are important—is given digitally, in thousands of feet. The pointer shows hundreds of feet, and for this purpose, precise numerical information is seldom required. Moreover, the pilot can readily determine whether the plane is changing altitude upward or downward simply by seeing which way and how fast the hand is spinning. With the digital readout, if the change in altitude is too fast, all that you can see is a blur. (Same problem with digital speedometers.)

What about the clock? Which is best: digital or analog? The answer is the same: It all depends upon the task. We have to bear in mind, however, that watches suffer a major disadvantage when compared to the speedometer. With the speedometer, there is one pointer, one scale. With watches, there are two or three pointers, two different scales. It is not easy to learn to read time on analog watches: Children have considerable difficulty, and even adults who read the time quickly make errors, confusing the hour hand with the minute hand, and vice versa, so that 3:20 might be read as 4:18. With today's watches and clocks we no longer are forced to choose between analog and digital. It is possible to have both.

## FITTING THE REPRESENTATION TO THE PERSON

Did you ever notice how much information is provided for us by the world? Thank goodness, else we would never manage. Look at how many different physical objects we use in our lives: knives, forks, pencils, paper clips, shoes, shoelaces, buttons, zippers—I once estimated that we all are probably familiar with twenty thousand different objects, each small, specialized, and requiring learning. How do we cope? How do we manage to learn about each of these individual items? The answer is that the



physical design of an object makes all the difference. You can often tell just by looking at something what function it serves or, at least, which parts you are supposed to hold, push, or pull; which parts operate upon other devices. A thumbtack has an obvious place to push, an obvious point that can be used to pierce, hold, or even lift objects. Most devices provide enough clues to their operation that even though some instruction might still be needed, the instruction can be accomplished in just a few words, or perhaps just by watching someone else use it once. Physical devices have affordances, mappings, and constraints that greatly aid in figuring out how they can be used. Of course, when the device is poorly designed or constructed, these same factors can greatly hinder its use. This is the story I told in *The Design of Everyday Things.*

Physical artifacts can be designed so that they are easy to learn, easy to use. The same is true for cognitive artifacts, although here some new principles must be provided. We need to consider the nature of the task to be performed and the powers of the human.

We humans seek understanding, causes, and purpose. We are good at remembering experiences, good at stories and events, bad with the minutiae of modern life. We are attentive to our surrounds, remarkably quick to notice changes. And we see patterns and meanings even when they are obscure and hidden. These very same characteristics, however, can conflict with the demands of the modern industrial, technological life. The conflicts are made worse by the technology that is imposed upon us on its terms instead of ours. The conflicts could be minimized or even eliminated if the technology came to us on our terms.

Furthermore, we are social creatures. We communicate and work well in small groups, sharing and cooperating to accomplish tasks beyond the capability of the single individual. The cooperation is aided through the communicative powers of language and body: spoken and written words, gestures, eye contact, and facial expressions. People are biologically predisposed to work in rich, ever-changing, sociocultural environments. We exploit any rela-

tionships we can find and invent interpretations. All this aids us in making sense of an otherwise chaotic world.

Today we live in an information-based technological world. The problem is that this is an invisible technology. Knowledge and information are invisible. They have no natural form. It is up to the conveyer of the information and knowledge to provide shape, substance, and organization. The irony is that too much of our artificial world is oversimplified, overabstract, thereby taking away our most powerful capabilities.

Information media do not necessarily take on a form amenable to humans. They are true internal artifacts, in that information is abstract and invisible. The information tends to be represented internally in the same manner, regardless of its content. This is especially true of digital media, in which everything gets transformed into a numerical representation expressed as binary digits—a long sequence of 0s and 1s (usually encoded as two levels of electrical charge). More and more of our media use a digital format for storage and transmission—television, telephone, radio, books. Digital signals offer a number of advantages: They are relatively easy to process and work with electronically, they offer great immunity to interference by electrical noise, and they are the natural medium of storage and processing used by computer systems.

Digital media have a number of disadvantages. A major problem is that of making access feasible, understandable. The common format for everything doesn't help: The most beautiful painting, the most stirring music, the most profound thoughts are all reduced to the identical format of internal states of the artifact. In fact, one cannot tell whether a given message is music or art, beautiful or ugly, from the internal representation. The medium is completely neutral with respect to the content. Hence the power of the medium: The signals can be transmitted and operated on without much regard for their content. It is only when people get into the picture that the form and content matter. Humans need a meaningful, accessible representation: sounds, sights, touch, organized in meaningful, interpretable ways. The result, however, is that we are ever more dependent upon the design of our devices to make the information visible and to make the artifact usable.



The new-fashioned information artifacts take on arbitrary shape and form. There is no natural mapping, no natural principles of operation. The critical operations all take place invisibly through internal representations. If we are to be able to use these artifacts easily and efficiently, the designers have to provide us with assistance, with an understandable, coherent structure. We are in the hands of the designers, who have the power to make the artifact meaningful, to provide substance and richness, and to make its use support the activities of interest. The best of the artifacts will become invisible, fitting the task so perfectly that they merge with it. They will be a delight to use.

Design should be like telling a story. The design team should start by considering the tasks that the artifact is intended to serve and the people who will use it. To accomplish this, the design team must include expertise in human cognition, in social interaction, in the task that is to be supported, and in the technologies that will be used. Appropriate design is a hard job. But without it, our tools will continue to frustrate, to confuse more than clarify, and to get in the way rather than merge with the task. The power of information artifacts is that they provide an unrivaled opportunity to enhance our lives. The danger is that they can add to the stress of everyday existence.

## TECHNOLOGIES HAVE AFFORDANCES

"You know what bothers me about the difference between television and newspapers?" a friend asked. With television the ads are inescapable. They blast right at you—you can't avoid them, except by leaving the room. It's not that way at all with newspapers. In fact, it's just the opposite problem. Sometimes, when I really want to read the ads, I don't even see them. Why is that?"

The problem lies with the differences between the affordances of television and newspapers. The *affordances* of an object refers to its possible functions: A chair affords support,

whether for standing, sitting, or the placement of objects. A pencil affords lifting, grasping, turning, poking, supporting, tapping, and of course, writing. In design, the critical issue is perceived affordances: what people perceive the object can do. We tend to use objects in ways suggested by the most salient perceived affordances, not in ways that are difficult to discover (hence the fact that many owners of electronic devices often fail to use some of their most powerful features—indeed, often do not even know of their existence).

Affordance also applies to technologies. Different technologies afford different operations. That is, they make some things easy to do, others difficult or impossible. It should come as no surprise that those things that the affordances make easy are apt to get done, those things that the affordances make difficult are not apt to get done.

In the case of my friend's complaint about the inescapability of television ads and the invisibility of newspaper ads, the culprit was the differences between the affordances of the serial, time-paced presentation of television versus the parallel, self-paced presentation of the printed page coupled with the fact that, on the whole, people can only attend to one thing at a time. With the television set, there is only one thing to look at, one sound channel to listen to. One message at a time.* Sure, we can daydream or look away from the television set, but unless we make some active effort to avoid it, the material impinges upon the brain, upon the conscious mind. Couple this with the cleverness of television advertisers to exploit the seductive experiential quality of the medium and it becomes clear how, as viewers, we are readily hooked.

The printed page provides a very different story. Here, as readers, we guide the intake of information. We must actively move our eyes across the page. Multiple articles and advertisements appear on each page, and because people can only read one thing at a time, we have to actively choose which to read: The act of selecting one automatically excludes the others. Moreover, once we start reading an article, our eyes track its location, even across pages, thereby missing any other material on the page.



A television channel presents different information by devoting different time slots to different information contents, presenting everything in the same spatial location. Print media—such as books, magazines, and newspapers—present different information by devoting different spatial locations to different information contents, presenting everything at the same time. The differences in the use of space and time between the television and print media yield different affordances. One result is that we find it difficult to escape television commercials because we focus on spatial locations. Even if we divert our attention as soon as we become aware of the commercial, it is too late. In a newspaper, the advertisements are not in the same spatial location as the stories, so the eyes can miss them, even when we would prefer not to.

Television organizes its information in time, newspapers in space. The result is that television paces the reader (it is event-paced), whereas with the printed page, it is the reader who sets the pace (it is self-paced). This is why the printed page provides better affordances for reflection than does the television show. Because reading is self-paced, there is time to pause and reflect upon what has been read, thus performing a deeper analysis than is possible with the event-paced affordances of television.

When a technology attempts to force a medium into a usage that violates its affordances, then the medium gets in the way. The result makes the difference between a humane technology and an inhumane one. Let me give an example of a reasonable idea made inhumane by the affordances of the medium: voice-messaging systems.

The phrase *voice-messaging systems* refers to several different kinds of things. The goal is to make it easy to provide customers with the specific information they need through selective messaging. These systems have proliferated. Companies find them attractive. As for the callers? Annoyed and upset. Furious even. Why? The concept is reasonable; the failure lies in the affordances. The version that upsets users so much does so because it violates the affordances of the telephone medium.

There are a number of problems from the point of view of the

caller, but most are traceable to the fact that the technology available to the caller is tiny and impoverished: It is a poor vehicle for communication. All the caller has is a keypad with twelve buttons and a telephone instrument that allows voice messages to be spoken and listened to. Somehow, using these minimal tools, the caller has to get access to just the correct information out of all the thousands of possibilities offered by the company. The system gives explanations and alternative choices through voice messages to the caller: The limited affordances of voice in this context are at the heart of the problem.

Voice is a serial medium of relatively low-speed, low-capacity communication. It is transient: The information is available only for the duration of the sound itself. Therefore, if a series of alternatives is to be presented to the listener, they have to be limited in number to the amount that can be retained in working memory: Ten would be too much, five would be acceptable if the caller could pay full attention to the message and not be distracted. Three would be safe. Three alternatives? Where the company might have perhaps five hundred different messages and locations to which the call could be directed? Even limiting the alternatives to five would create problems.

The transient nature of voice is a major problem that fundamentally limits the service. The speed of speaking is another. Each alternative might take one or two seconds to speak. A message with three alternatives would thereby take as long as six seconds; ten alternatives might take twenty seconds. That's a long time. There are two standard solutions to these problems, usually used together.

The first solution is to try to present a larger number of alternatives but encourage the user to punch in the number of the desired one as soon as it is heard. This avoids the working memory problem: Listen to each, make a yes/no decision, and then go on to the next. This, of course, assumes that you can recognize the one you want when you hear it. In fact, often the only way to tell which one you want is to listen to all the alternatives and make your selection by a combination of knowing some are irrelevant, some seem relevant.



The second solution is to make the selection process hierarchical. Each set of choices leads to another set of choices, until the end point is reached. If there are $n$ alternatives the first time, and if each of the $n$ choices leads to a second level, also with $n$ choices, then the two levels combined let you reach $n^2$ alternatives. With three levels, one could reach $n^3$ alternatives, and so on. If there are five alternatives per level and five hundred possible destinations, four levels are required. Four levels, each with five alternatives? It might take ten seconds at each level to listen to the choices and select one, so it would take forty seconds of listening and keypunching to get to the desired location: if, indeed, the end point turned out to be the desired one. Forty seconds is a long time. Try it: Stop right now and do nothing for forty seconds.

The telephone system lacks some critical affordances—in particular, those required for graceful error correction. Suppose you push the wrong number, either by accident or because you had misinterpreted the message. How would you correct it? In the first case, you would probably realize the error as soon as you made it. The system could make one of its alternatives a chance to go back to the previous level, but if it did so, it would either have to increase the number of alternatives at each level (in my example, to six) or replace one of the other alternatives. If the first procedure is used, the time per level increases to twelve seconds, and the total time increases to forty-eight seconds (along with the chance of overloading working memory). If the second procedure is used, then there are only effectively four alternatives per level, so it would take five levels instead of four to reach the destination: The time would increase to fifty seconds.

Let me illustrate the problems with an example. I subscribe to the *Pocket Edition of the Official Airline Guide* (OAG), a monthly publication small enough to carry in a shirt pocket. American Airlines has an automated telephone flight information system ("Dial-AA-Flight"), so I decided to compare how long it would take to find some American Airlines flights using its automated system with the time required using the OAG pocket guide.

The task I set for myself was to see how long it would take to find all American Airlines flights between San Diego, Califor-

nia, and Detroit, Michigan. First, I used the pocket guide: Forty seconds and I had the answer. Then I dialed the phone number of the American Airlines system. I was greeted by some music and voice instructions. One of the possibilities was a training session, so I selected it, and when it was finished, I hung up so as not to bias the timing with training time. Then I called again and this time chose the option for experienced folks. That got me to a second level. The system was well done. American Airlines only presented two or three alternatives at each level in an attempt to keep the information within working-memory span. I had no trouble remembering the alternatives, so after selecting flight-scheduling and fare information, I specified my departure and destination cities by typing the first four letters on the telephone keypad. Each digit on the telephone has three letters, so the digits were ambiguous. (For example, both San Diego and San Francisco are specified by the same digit sequence: 7263.) At the next level, I had to select among the resulting ambiguities. At the eighth level I was asked whether I wanted the flight to leave in AM or PM. That was strange: What if I didn't know when I wanted to leave? What I wanted to do was to compare the alternatives. I had no option but to select one. When going east from California I almost always leave in the morning, so I typed "1*" for AM.

On the ninth level, I was asked to type in the time of the flight that I was interested in. Huh? I didn't know when the flights left, that's why I was calling—so that I might find out. I wasn't sure what number to type, so I decided to use the earliest time I would be willing to leave: 8:00 AM. Alas, I made an error—I typed "7" instead of "8." How could I correct the error? I didn't know. I had not been told. I cleverly decided to type an illegal time, assuming that this would get a message saying something like "That time was not proper, please try again." I entered the time "777*" and waited. Instead of a friendly error message, I got an even friendlier one saying that a human was on the way to help me. I hung up. Total time, not counting training, 128 seconds. Two minutes and eight seconds—over three times longer than with the book—and the



book gave me all the flights, but with the phone system, I had gotten none of them.

I still don't know whether it is possible to get all the flights with the phone system, unless you do it one at a time (typing in each possible departure time?). Yet I never make a reservation by simply saying what day and time I wish to leave: I always compare the range of possibilities and, after some reflection, select the one that best fits my needs for that trip. I think I will stick with my travel agent and the OAG.

No wonder customers are irate. The idea of delivering information by telephone is actually a pretty good one, but the medium does not support it. Voice is too slow, too transient, and the telephone keypad too limited. Voice-messaging systems will only work painlessly if the medium is transformed to have better affordances, probably by changing the equipment available to the caller. Notice how the printed medium of the pocket OAG is superior for rapid scanning and efficient presentation of information. Voice simply won't do: It has the wrong affordances. Voice is serial. Vision is parallel. Voice is transient. Printed or displayed images are relatively permanent. If my telephone had a high-quality visual display with high resolution and contrast, then the voice-message system could be replaced with a visual-message system. I would always see a page of information with multiple alternatives, including simple ways to get to human assistance at any point. It would not be hard to present twenty to fifty alternatives on a properly configured visual display, so five hundred alternatives could be reached in two to three levels. Reading is faster than listening, and if the displays were properly designed to allow rapid scanning and easy error correction (a very big "if," given the poor track record of existing technology), the affordances of visual presentation would transform a clumsy, inefficient scheme into a workable one.

Suppose that the visually presented system were in use, would this solve the problems? Here I am, foisting yet another technological solution on us, yet another way to avoid having people talk to people. My answer here is the same as always: There is no standard

answer, it all depends upon the situation. Sometimes it is better for all concerned to get rapid, efficient access to the answer, even if technologically presented. Sometimes we need human interaction. A well-designed system will provide both.

The affordances of the medium do make a difference. My simple analyses and test show some of the reasons these voice-message systems are so disliked. But this raises another question: If they are so universally disliked, why are these systems used? Why are their numbers increasing? Why this effort to force the technology to do something for which it is so ill suited? The answer is that the systems appear to provide major benefits to the company. They relieve employees from a continual barrage of phone callers who have standard questions, and they do save the cost of numerous telephone and information operators. This attitude, of course, neglects the cost to the callers who are frustrated and angered by the system. One company officer who ordered the machines taken out of his company described it to his employees this way: "What you're saying is your time is worth more than their time."

Voice-mail and voice-messaging systems do have situations where they work well. They often provide a superior means of delivering personal messages. One person told me of a system for providing information about films and theater schedules that he considered a superior use of the technology—better than any other existing method. (I have not had a chance to try it.) The technology can work well in appropriate situations for appropriate tasks.

Technology usually provides a series of tradeoffs. Each asset is offset by a deficit. It is always necessary to decide whether the assets outweigh the deficits. Frequently, the tradeoffs fall differently upon different people. A major problem occurs when those who suffer from technology's deficits and those who benefit are not the same people. In the telephone system, the benefits are to the company, the burdens fall upon the users. This kind of tradeoff I have come to call Grudin's law, after Jonathan Grudin, who first proposed it:



**Grudin's law:** When those who benefit are not those who do the work, then the technology is likely to fail or, at least, be subverted.

Grudin's law very definitely applies to the voice-messaging systems. May they die a rapid death.