CHAPTER ONE

THE PSYCHOPATHOLOGY OF EVERYDAY THINGS

If I were placed in the cockpit of a modern jet airliner, my inability to perform well would neither surprise nor bother me. But why should I have trouble with doors and light switches, water faucets and stoves? “Doors?” I can hear the reader saying. “You have trouble opening doors?” Yes. I push doors that are meant to be pulled, pull doors that should be pushed, and walk into doors that neither pull nor push, but slide. Moreover, I see others having the same troubles—unnecessary troubles. My problems with doors have become so well known that confusing doors are often called “Norman doors.” Imagine becoming famous for doors that don’t work right. I’m pretty sure that’s not what my parents planned for me. (Put “Norman doors” into your favorite search engine—be sure to include the quote marks: it makes for fascinating reading.)

How can such a simple thing as a door be so confusing? A door would seem to be about as simple a device as possible. There is not much you can do to a door: you can open it or shut it. Suppose you are in an office building, walking down a corridor. You come to a door. How does it open? Should you push or pull, on the left or the right? Maybe the door slides. If so, in which direction? I have seen doors that slide to the left, to the right, and even up into the ceiling. The design of the door should indicate how to work it without any need for signs, certainly without any need for trial and error.

FIGURE 1.1. Coffeepot for Masochists. The French artist Jacques Carelman in his series of books *Catalogue d'objets introuvables* (Catalog of unfindable objects) provides delightful examples of everyday things that are deliberately unworkable, outrageous, or otherwise ill-formed. One of my favorite items is what he calls "coffeepot for masochists." The photograph shows a copy given to me by collegues at the University of California, San Diego. It is one of my treasured art objects. (Photograph by Aymin Shamma for the author.)

A friend told me of the time he got trapped in the doorway of a post office in a European city. The entrance was an imposing row of six glass swinging doors, followed immediately by a second, identical row. That's a standard design: it helps reduce the airflow and thus maintain the indoor temperature of the building. There was no visible hardware: obviously the doors could swing in either direction: all a person had to do was push the side of the door and enter.

My friend pushed on one of the outer doors. It swung inward, and he entered the building. Then, before he could get to the next row of doors, he was distracted and turned around for an instant. He didn't realize it at the time, but he had moved slightly to the right. So when he came to the next door and pushed it, nothing happened. "Hmm," he thought, "must be locked." So he pushed the side of the adjacent door. Nothing. Puzzled, my friend decided to go outside again. He turned around and pushed against the side of a door. Nothing. He pushed the adjacent door. Nothing. The door he had just entered no longer worked. He turned around once more and tried the inside doors again. Nothing. Concern, then mild panic. He was trapped! Just then, a group of people on the other side of the entranceway (to my friend's right) passed easily through both sets of doors. My friend hurried over to follow their path.

How could such a thing happen? A swinging door has two sides. One contains the supporting pillar and the hinge, the other is unsupported. To open the door, you must push or pull on the unsupported edge. If you push on the hinge side, nothing happens. In my friend's case, he was in a building where the designer aimed for beauty, not utility. No distracting lines, no visible pillars, no visible hinges. So how can the ordinary user know which side to push on? While distracted, my friend had moved toward the (invisible) supporting pillar, so he was pushing the doors on the hinged side. No wonder nothing happened. Attractive doors. Stylish. Probably won a design prize.

Two of the most important characteristics of good design are *discoverability* and *understanding*. Discoverability: Is it possible to even figure out what actions are possible and where and how to perform them? Understanding: What does it all mean? How is the product supposed to be used? What do all the different controls and settings mean?

The doors in the story illustrate what happens when discoverability fails. Whether the device is a door or a stove, a mobile phone or a nuclear power plant, the relevant components must be visible, and they must communicate the correct message: What actions are possible? Where and how should they be done? With doors that push, the designer must provide signals that naturally indicate where to push. These need not destroy the aesthetics. Put a vertical plate on the side to be pushed. Or make the supporting pillars visible. The vertical plate and supporting pillars are natural signals, naturally interpreted, making it easy to know just what to do: no labels needed.

With complex devices, discoverability and understanding require the aid of manuals or personal instruction. We accept this if the device is indeed complex, but it should be unnecessary for simple things. Many products defy understanding simply because they have too many functions and controls. I don't think that simple home appliances—stoves, washing machines, audio and television sets—should look like Hollywood's idea of a spaceship control room. They already do, much to our consternation. Faced with a bewildering array of controls and displays, we simply memorize one or two fixed settings to approximate what is desired.

In England I visited a home with a fancy new Italian washer-dryer combination, with super-duper multisymbol controls, all to do everything anyone could imagine doing with the washing and drying of clothes. The husband (an engineering psychologist) said he refused to go near it. The wife (a physician) said she had simply memorized one setting and tried to ignore the rest. I asked to see the manual: it was just as confusing as the device. The whole purpose of the design is lost.

The Complexity of Modern Devices

All artificial things are designed. Whether it is the layout of furniture in a room, the paths through a garden or forest, or the intricacies of an electronic device, some person or group of people had to decide upon the layout, operation, and mechanisms. Not all designed things involve physical structures. Services, lectures, rules and procedures, and the organizational structures of businesses and governments do not have physical mechanisms, but their rules of operation have to be designed, sometimes informally, sometimes precisely recorded and specified.

But even though people have designed things since prehistoric times, the field of design is relatively new, divided into many areas of specialty. Because everything is designed, the number of areas is enormous, ranging from clothes and furniture to complex control rooms and bridges. This book covers everyday things, focusing on the interplay between technology and people to ensure that the products actually fulfill human needs while being understandable and usable. In the best of cases, the products should also be delightful and enjoyable, which means that not only must the requirements of engineering, manufacturing, and ergonomics be satisfied, but attention must be paid to the entire experience, which means the aesthetics of form and the quality of interaction. The major areas of design relevant to this book are industrial design, interaction design, and experience design. None of the fields is well defined, but the focus of the efforts does vary, with industrial designers emphasizing form and material, interactive designers emphasizing understandability and usability, and experience designers emphasizing the emotional impact. Thus:

Industrial design: The professional service of creating and developing concepts and specifications that optimize the function, value, and appearance of products and systems for the mutual benefit of both user and manufacturer (from the *Industrial Design Society of America's* website).

Interaction design: The focus is upon how people interact with technology. The goal is to enhance people's understanding of what can be done, what is happening, and what has just occurred. Interaction design draws upon principles of psychology, design, art, and emotion to ensure a positive, enjoyable experience.

Experience design: The practice of designing products, processes, services, events, and environments with a focus placed on the quality and enjoyment of the total experience.

Design is concerned with how things work, how they are controlled, and the nature of the interaction between people and technology. When done well, the results are brilliant, pleasurable products. When done badly, the products are unusable, leading to great frustration and irritation. Or they might be usable, but force us to behave the way the product wishes rather than as we wish.

Machines, after all, are conceived, designed, and constructed by people. By human standards, machines are pretty limited. They do not maintain the same kind of rich history of experiences that people have in common with one another, experiences that enable us to interact with others because of this shared understanding. Instead, machines usually follow rather simple, rigid rules of behavior. If we get the rules wrong even slightly, the machine does what it is told, no matter how insensible and illogical. People are imaginative and creative, filled with common sense; that is, a lot of valuable knowledge built up over years of experience. But instead of capitalizing on these strengths, machines require us to be precise and accurate, things we are not very good at. Machines have no leeway or common sense. Moreover, many of the rules followed by a machine are known only by the machine and its designers.

When people fail to follow these bizarre, secret rules, and the machine does the wrong thing, its operators are blamed for not understanding the machine, for not following its rigid specifications. With everyday objects, the result is frustration. With complex devices and commercial and industrial processes, the resulting difficulties can lead to accidents, injuries, and even deaths. It is time to reverse the situation: to cast the blame upon the machines and their design. It is the machine and its design that are at fault. It is the duty of machines and those who design them to understand people. It is not our duty to understand the arbitrary, meaningless dictates of machines.

The reasons for the deficiencies in human-machine interaction are numerous. Some come from the limitations of today's technology. Some come from self-imposed restrictions by the designers, often to hold down cost. But most of the problems come from a complete lack of understanding of the design principles necessary for effective human-machine interaction. Why this deficiency? Because much of the design is done by engineers who are experts in technology but limited in their understanding of people. "We are people ourselves," they think, "so we understand people." But in fact, we humans are amazingly complex. Those who have not studied human behavior often think it is pretty simple. Engineers, moreover, make the mistake of thinking that logical explanation is sufficient: "If only people would read the instructions," they say, "everything would be all right."

Engineers are trained to think logically. As a result, they come to believe that all people must think this way, and they design their machines accordingly. When people have trouble, the engineers are upset, but often for the wrong reason. "What are these people doing?" they will wonder. "Why are they doing that?" The problem with the designs of most engineers is that they are too logical. We have to accept human behavior the way it is, not the way we would wish it to be.

I used to be an engineer, focused upon technical requirements, quite ignorant of people. Even after I switched into psychology and cognitive science, I still maintained my engineering emphasis upon logic and mechanism. It took a long time for me to realize that my understanding of human behavior was relevant to my interest in the design of technology. As I watched people struggle with technology, it became clear that the difficulties were caused by the technology, not the people.

I was called upon to help analyze the American nuclear power plant accident at Three Mile Island (the island name comes from the fact that it is located on a river, three miles south of Middle-town in the state of Pennsylvania). In this incident, a rather simple mechanical failure was misdiagnosed. This led to several days of difficulties and confusion, total destruction of the reactor, and a very close call to a severe radiation release, all of which brought the American nuclear power industry to a complete halt. The operators were blamed for these failures: "human error" was the immediate analysis. But the committee I was on discovered that the plant's control rooms were so poorly designed that error was inevitable: design was at fault, not the operators. The moral was simple: we were designing things for people, so we needed to understand both technology and people. But that's a difficult step for many engineers: machines are so logical, so orderly. If we didn't have people, everything would work so much better. Yup, that's how I used to think.

My work with that committee changed my view of design. Today, I realize that design presents a fascinating interplay of technology and psychology, that the designers must understand both. Engineers still tend to believe in logic. They often explain to me in great, logical detail, why their designs are good, powerful, and wonderful. "Why are people having problems?" they wonder. "You are being too logical," I say. "You are designing for people the way you would like them to be, not for the way they really are."

When the engineers object, I ask whether they have ever made an error, perhaps turning on or off the wrong light, or the wrong stove burner. "Oh yes," they say, "but those were errors." That's the point: even experts make errors. So we must design our machines on the assumption that people will make errors. (Chapter 5 provides a detailed analysis of human error.)

Human-Centered Design

People are frustrated with everyday things. From the ever-increasing complexity of the automobile dashboard, to the increasing automation in the home with its internal networks, complex music, video, and game systems for entertainment and communication, and the increasing automation in the kitchen, everyday life sometimes seems like a never-ending fight against confusion, continued errors, frustration, and a continual cycle of updating and maintaining our belongings.

In the multiple decades that have elapsed since the first edition of this book was published, design has gotten better. There are now many books and courses on the topic. But even though much has improved, the rapid rate of technology change outpaces the advances in design. New technologies, new applications, and new methods of interaction are continually arising and evolving. New industries spring up. Each new development seems to repeat the mistakes of the earlier ones; each new field requires time before it, too, adopts the principles of good design. And each new invention of technology or interaction technique requires experimentation and study before the principles of good design can be fully integrated into practice. So, yes, things are getting better, but as a result, the challenges are ever present.

The solution is human-centered design (HCD), an approach that puts human needs, capabilities, and behavior first, then designs to accommodate those needs, capabilities, and ways of behaving. Good design starts with an understanding of psychology and technology. Good design requires good communication, especially from machine to person, indicating what actions are possible, what is happening, and what is about to happen. Communication is especially important when things go wrong. It is relatively easy to design things that work smoothly and harmoniously as long as things go right. But as soon as there is a problem or a misunderstanding, the problems arise. This is where good design is essential. Designers need to focus their attention on the cases where things go wrong, not just on when things work as planned. Actually, this is where the most satisfaction can arise: when something goes wrong but the machine highlights the problems, then the person understands the issue, takes the proper actions, and the problem is solved. When this happens smoothly, the collaboration of person and device feels wonderful.

TABLE 1.1. The Role of HCD and Design Specializations

Experience design	
Industrial design	These are areas of focus
Interaction design	
Human-centered design	The process that ensures that the designs match the needs and capabilities of the people for whom they are intended

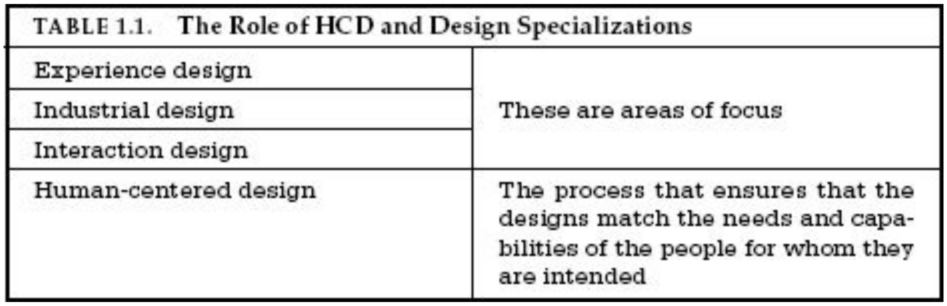


TABLE 1.1. The Role of HCD and Design Specializations

Experience design	These are areas of focus
Industrial design	
Interaction design	
Human-centered design	The process that ensures that the designs match the needs and capabilities of the people for whom they are intended

Human-centered design is a design philosophy. It means starting with a good understanding of people and the needs that the design is intended to meet. This understanding comes about primarily through observation, for people themselves are often unaware of their true needs, even unaware of the difficulties they are encountering. Getting the specification of the thing to be defined is one of the most difficult parts of the design, so much so that the HCD principle is to avoid specifying the problem as long as possible but instead to iterate upon repeated approximations. This is done through rapid tests of ideas, and after each test modifying the approach and the problem definition. The results can be products that truly meet the needs of people. Doing HCD within the rigid time, budget, and other constraints of industry can be a challenge: Chapter 6 examines these issues.

Where does HCD fit into the earlier discussion of the several different forms of design, especially the areas called industrial, interaction, and experience design? These are all compatible. HCD is a philosophy and a set of procedures, whereas the others are areas of focus (see Table 1.1). The philosophy and procedures of HCD add deep consideration and study of human needs to the design process, whatever the product or service, whatever the major focus.

Fundamental Principles of Interaction

Great designers produce pleasurable experiences. *Experience*: note the word. Engineers tend not to like it; it is too subjective. But when I ask them about their favorite automobile or test equipment, they will smile delightedly as they discuss the fit and finish, the sensation of power during acceleration, their ease of control while shifting or steering, or the wonderful feel of the knobs and switches on the instrument. Those are experiences.

Experience is critical, for it determines how fondly people remember their interactions. Was the overall experience positive, or was it frustrating and confusing? When our home technology behaves in an uninterpretable fashion we can become confused, frustrated, and even angry—all strong negative emotions. When there is understanding it can lead to a feeling of control, of mastery, and of satisfaction or even pride—all strong positive emotions. Cognition and emotion are tightly intertwined, which means that the designers must design with both in mind.

When we interact with a product, we need to figure out how to work it. This means discovering what it does, how it works, and what operations are possible: discoverability. Discoverability results from appropriate application of five fundamental psychological concepts covered in the next few chapters: *affordances, signifiers, constraints, mappings*, and *feedback*. But there is a sixth principle, perhaps most important of all: the *conceptual model* of the system. It is the conceptual model that provides true understanding. So I now turn to these fundamental principles, starting with affordances, signifiers, mappings, and feedback, then moving to conceptual models. Constraints are covered in Chapters 3 and 4.

AFFORDANCES

We live in a world filled with objects, many natural, the rest artificial. Every day we encounter thousands of objects, many of them new to us. Many of the new objects are similar to ones we already know, but many are unique, yet we manage quite well. How do we do this? Why is it that when we encounter many unusual natural objects, we know how to interact with them? Why is this true with many of the artificial, human-made objects we encounter? The answer lies with a few basic principles. Some of the most important of these principles come from a consideration of affordances.

The term *affordance* refers to the relationship between a physical object and a person (or for that matter, any interacting agent, whether animal or human, or even machines and robots). An affordance is a relationship between the properties of an object and the capabilities of the agent that determine just how the object could possibly be used. A chair affords ("is for") support and, therefore, affords sitting. Most chairs can also be carried by a single person (they afford lifting), but some can only be lifted by a strong person or by a team of people. If young or relatively weak people cannot lift a chair, then for these people, the chair does not have that affordance, it does not afford lifting.

The presence of an affordance is jointly determined by the qualities of the object and the abilities of the agent that is interacting. This relational definition of affordance gives considerable difficulty to many people. We are used to thinking that properties are associated with objects. But affordance is not a property. An affordance is a relationship. Whether an affordance exists depends upon the properties of both the object and the agent.

Glass affords transparency. At the same time, its physical structure blocks the passage of most physical objects. As a result, glass affords seeing through and support, but not the passage of air or most physical objects (atomic particles can pass through glass). The blockage of passage can be considered an anti-affordance—the prevention of interaction. To be effective, affordances and anti-affordances have to be discoverable—perceivable. This poses a difficulty with glass. The reason we like glass is its relative invisibility, but this aspect, so useful in the normal window, also hides its anti-affordance property of blocking passage. As a result, birds often try to fly through windows. And every year, numerous people injure themselves when they walk (or run) through closed glass doors or large picture windows. If an affordance or anti-affordance cannot be perceived, some means of signaling its presence is required: I call this property a *signifier* (discussed in the next section).

The notion of affordance and the insights it provides originated with J. J. Gibson, an eminent psychologist who provided many advances to our understanding of human perception. I had interacted with him over many years, sometimes in formal conferences and seminars, but most fruitfully over many bottles of beer, late at night, just talking. We disagreed about almost everything. I was an engineer who became a cognitive psychologist, trying to understand how the mind works. He started off as a Gestalt psychologist, but then developed an approach that is today named after him: Gibsonian psychology, an ecological approach to perception. He argued that the world contained the clues and that people simply picked them up through "direct perception." I argued that nothing could be direct: the brain had to process the information arriving at the sense organs to put together a coherent interpretation. "Nonsense," he loudly proclaimed; "it requires no interpretation: it is directly perceived." And then he would put his hand to his ears, and with a triumphant flourish, turn off his hearing aids: my counterarguments would fall upon deaf ears—literally.

When I pondered my question—how do people know how to act when confronted with a novel situation—I realized that a large part of the answer lay in Gibson's work. He pointed out that all the senses work together, that we pick up information about the world by the combined result of all of them. "Information pickup" was one of his favorite phrases, and Gibson believed that the combined information picked up by all of our sensory apparatus—sight, sound, smell, touch, balance, kinesthetic, acceleration, body position— determines our perceptions without the need for internal processing or cognition. Although he and I disagreed about the role played by the brain's internal processing, his brilliance was in focusing attention on the rich amount of information present in the world. Moreover, the physical objects conveyed important information about how people could interact with them, a property he named "affordance."

Affordances exist even if they are not visible. For designers, their visibility is critical: visible affordances provide strong clues to the operations of things. A flat plate mounted on a door affords pushing. Knobs afford turning, pushing, and pulling. Slots are for inserting things into. Balls are for throwing or bouncing. Perceived affordances help people figure out what actions are possible without the need for labels or instructions. I call the signaling component of affordances *signifiers*.

SIGNIFIERS

Are affordances important to designers? The first edition of this book introduced the term *affordances* to the world of design. The design community loved the concept and affordances soon propagated into the instruction and writing about design. I soon found mention of the term everywhere. Alas, the term became used in ways that had nothing to do with the original.

Many people find affordances difficult to understand because they are relationships, not properties. Designers deal with fixed properties, so there is a temptation to say that the property is an affordance. But that is not the only problem with the concept of affordances.

Designers have practical problems. They need to know how to design things to make them understandable. They soon discovered that when working with the graphical designs for electronic displays, they needed a way to designate which parts could be touched, slid upward, downward, or sideways, or tapped upon. The actions could be done with a mouse, stylus, or fingers. Some systems responded to body motions, gestures, and spoken words, with no touching of any physical device. How could designers describe what they were doing? There was no word that fit, so they took the closest existing word—*affordance*.

Soon designers were saying such things as, “I put an affordance there,” to describe why they displayed a circle on a screen to indicate where the person should touch, whether by mouse or by finger. “No,” I said, “that is not an affordance. That is a way of communicating where the touch should be. You are communicating where to do the touching: the affordance of touching exists on the entire screen: you are trying to signify *where* the touch should take place. That’s not the same thing as saying *what* action is possible.”

Not only did my explanation fail to satisfy the design community, but I myself was unhappy. Eventually I gave up: designers needed a word to describe what they were doing, so they chose *affordance*. What alternative did they have? I decided to provide a better answer: *signifiers*. Affordances determine what actions are possible. Signifiers communicate where the action should take place. We need both.

People need some way of understanding the product or service they wish to use, some sign of what it is for, what is happening, and what the alternative actions are. People search for clues, for any sign that might help them cope and understand. It is the sign that is important, anything that might signify meaningful information. Designers need to provide these clues. What people need, and what designers must provide, are signifiers. Good design requires, among other things, good communication of the purpose, structure, and operation of the device to the people who use it. That is the role of the signifier.

The term *signifier* has had a long and illustrious career in the exotic field of semiotics, the study of signs and symbols. But just as I appropriated *affordance* to use in design in a manner somewhat different than its inventor had intended, I use *signifier* in a somewhat different way than it is used in semiotics. For me, the term *signifier* refers to any mark or sound, any perceivable indicator that communicates appropriate behavior to a person.

Signifiers can be deliberate and intentional, such as the sign PUSH on a door, but they may also be accidental and unintentional, such as our use of the visible trail made by previous people walking through a field or over a snow-covered terrain to determine the best path. Or how we might use the presence or absence of people waiting at a train station to determine whether we have missed the train. (I explain these ideas in more detail in my book *Living with Complexity*.)

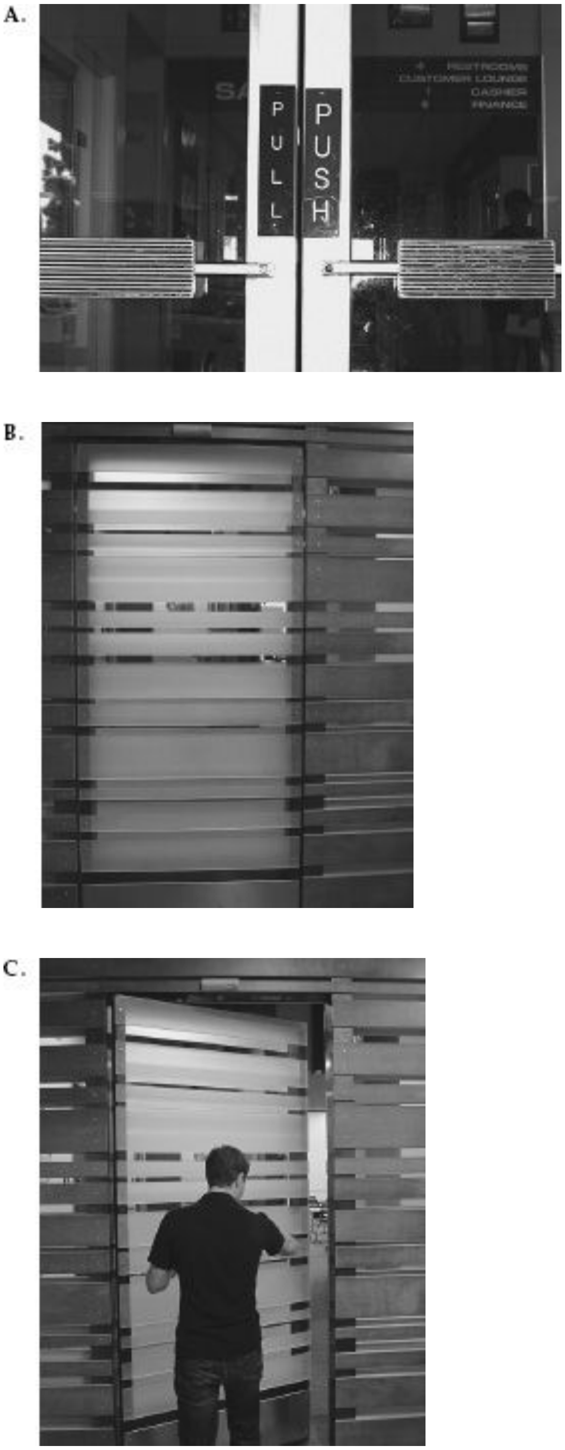


FIGURE 1.2. Problem Doors: Signifiers Are Needed. Door hardware can signal whether to push or pull without signs, but the

hardware of the two doors in the upper photo, A, are identical even though one should be pushed, the other pulled. The flat, ribbed horizontal bar has the obvious perceived affordance of pushing, but as the signs indicate, the door on the left is to be pulled, the one on the right is to be pushed. In the bottom pair of photos, B and C, there are no visible signifiers or affordances. How does one know which side to push? Trial and error. When external signifiers—signs— have to be added to something as simple as a door, it indicates bad design. (Photographs by the author.)

The signifier is an important communication device to the recipient, whether or not communication was intended. It doesn't matter whether the useful signal was deliberately placed or whether it is incidental: there is no necessary distinction. Why should it matter whether a flag was placed as a deliberate clue to wind direction (as is done at airports or on the masts of sailboats) or was there as an advertisement or symbol of pride in one's country (as is done on public buildings). Once I interpret a flag's motion to indicate wind direction, it does not matter why it was placed there.

Consider a bookmark, a deliberately placed signifier of one's place in reading a book. But the physical nature of books also makes a bookmark an accidental signifier, for its placement also indicates how much of the book remains. Most readers have learned to use this accidental signifier to aid in their enjoyment of the reading. With few pages left, we know the end is near. And if the reading is torturous, as in a school assignment, one can always console oneself by knowing there are "only a few more pages to get through." Electronic book readers do not have the physical structure of paper books, so unless the software designer deliberately provides a clue, they do not convey any signal about the amount of text remaining.

A.

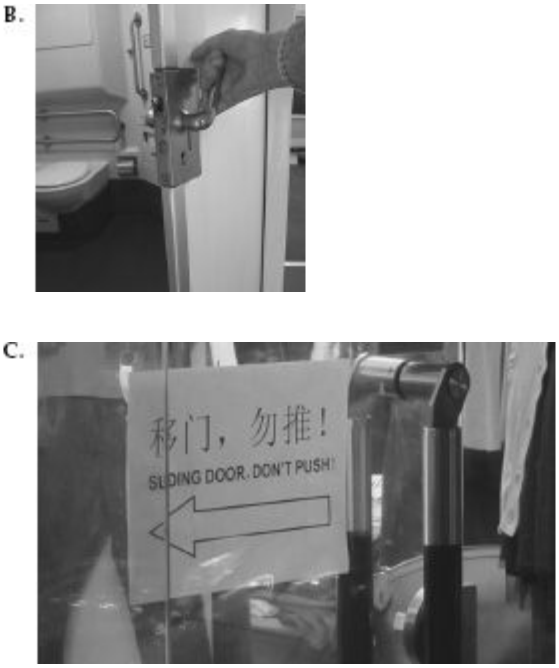


FIGURE 1.3. Sliding Doors: Seldom Done Well. Sliding doors are seldom signified properly. The top two photographs show the sliding door to the toilet on an Amtrak train in the United States. The handle clearly signifies “pull,” but in fact, it needs to be rotated and the door slid to the right. The owner of the store in Shanghai, China, Photo C, solved the problem with a sign. “DON’T PUSH!” it says, in both English and Chinese. Amtrak’s toilet door could have used a similar kind of sign. (Photographs by the author.)

Whatever their nature, planned or accidental, signifiers provide valuable clues as to the nature of the world and of social activities. For us to function in this social, technological world, we need to develop internal models of what things mean, of how they operate. We seek all the clues we can find to help in this enterprise, and in this way, we are detectives, searching for whatever guidance we might find. If we are fortunate, thoughtful designers provide the clues for us. Otherwise, we must use our own creativity and imagination.

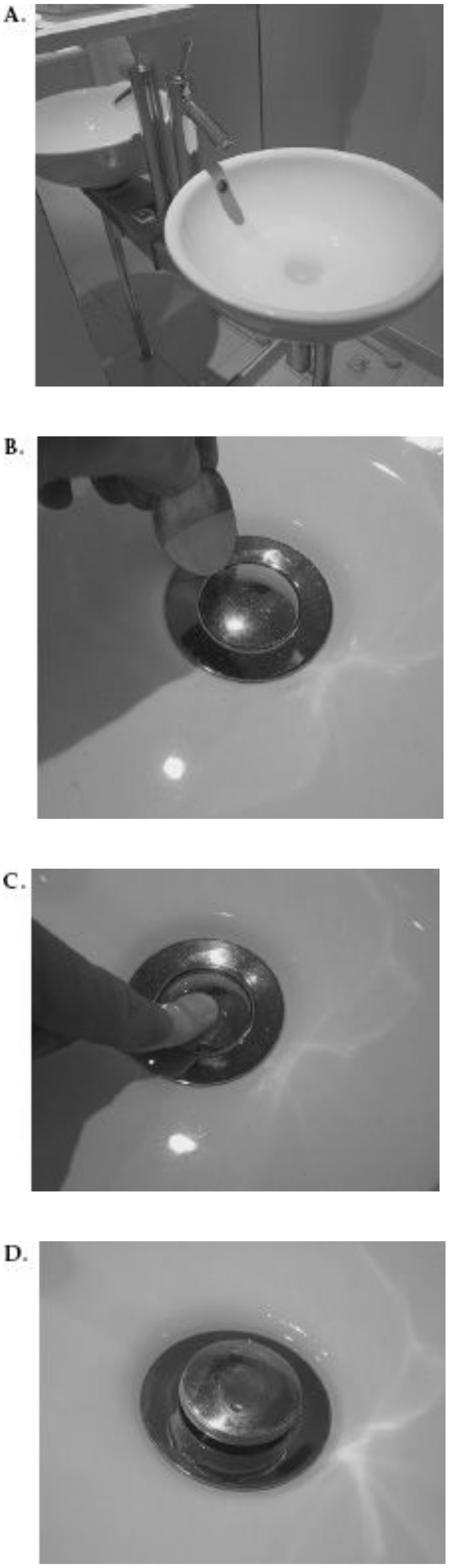


FIGURE 1.4. The Sink That Would Not Drain: Where Signifiers Fail. I washed my hands in my hotel sink in London, but then, as shown

in Photo A, was left with the question of how to empty the sink of the dirty water. I searched all over for a control: none. I tried prying open the sink stopper with a spoon (Photo B): failure. I finally left my hotel room and went to the front desk to ask for instructions. (Yes, I actually did.) “Push down on the stopper,” I was told. Yes, it worked (Photos C and D). But how was anyone to ever discover this? And why should I have to put my clean hands back into the dirty water to empty the sink? The problem here is not just the lack of signifier, it is the faulty decision to produce a stopper that requires people to dirty their clean hands to use it. (Photographs by the author.)

Affordances, perceived affordances, and signifiers have much in common, so let me pause to ensure that the distinctions are clear.

Affordances represent the possibilities in the world for how an agent (a person, animal, or machine) can interact with something. Some affordances are perceivable, others are invisible. Signifiers are signals. Some signifiers are signs, labels, and drawings placed in the world, such as the signs labeled “push,” “pull,” or “exit” on doors, or arrows and diagrams indicating what is to be acted upon or in which direction to gesture, or other instructions. Some signifiers are simply the perceived affordances, such as the handle of a door or the physical structure of a switch. Note that some perceived affordances may not be real: they may look like doors or places to push, or an impediment to entry, when in fact they are not. These are misleading signifiers, oftentimes accidental but sometimes purposeful, as when trying to keep people from doing actions for which they are not qualified, or in games, where one of the challenges is to figure out what is real and what is not.

A.

FIGURE 1.5. Accidental Affordances Can Become Strong Signifiers. This wall, at the Industrial Design department of KAIST, in Korea, provides an anti-affordance, preventing people from falling down the stair shaft. Its top is flat, an accidental by-product of the design. But flat surfaces afford support, and as soon as one person discovers it can be used to dispose of empty drink containers, the discarded container becomes a signifier, telling others that it is permissible to discard their items there. (Photographs by the author.)

My favorite example of a misleading signifier is a row of vertical pipes across a service road that I once saw in a public park. The pipes obviously blocked cars and trucks from driving on that road: they were good examples of anti-affordances. But to my great surprise, I saw a park vehicle simply go through the pipes. Huh? I walked over and examined them: the pipes were made of rubber, so vehicles could simply drive right over them. A very clever signifier, signaling a blocked road (via an apparent anti-affordance) to the average person, but permitting passage for those who knew.

To summarize:

- Affordances are the possible interactions between people and the environment. Some affordances are perceivable, others are not.

- Perceived affordances often act as signifiers, but they can be ambiguous.

- Signifiers signal things, in particular what actions are possible and how they should be done. Signifiers must be perceivable, else they fail to function.

In design, signifiers are more important than affordances, for they communicate how to use the design. A signifier can be words, a graphical illustration, or just a device whose perceived affordances are unambiguous. Creative designers incorporate the signifying part of the design into a cohesive experience. For the most part, designers can focus upon signifiers.

Because affordances and signifiers are fundamentally important principles of good design, they show up frequently in the pages of this book. Whenever you see hand-lettered signs pasted on doors, switches, or products, trying to explain how to work them, what to do and what not to do, you are also looking at poor design.

AFFORDANCES AND SIGNIFIERS: A CONVERSATION

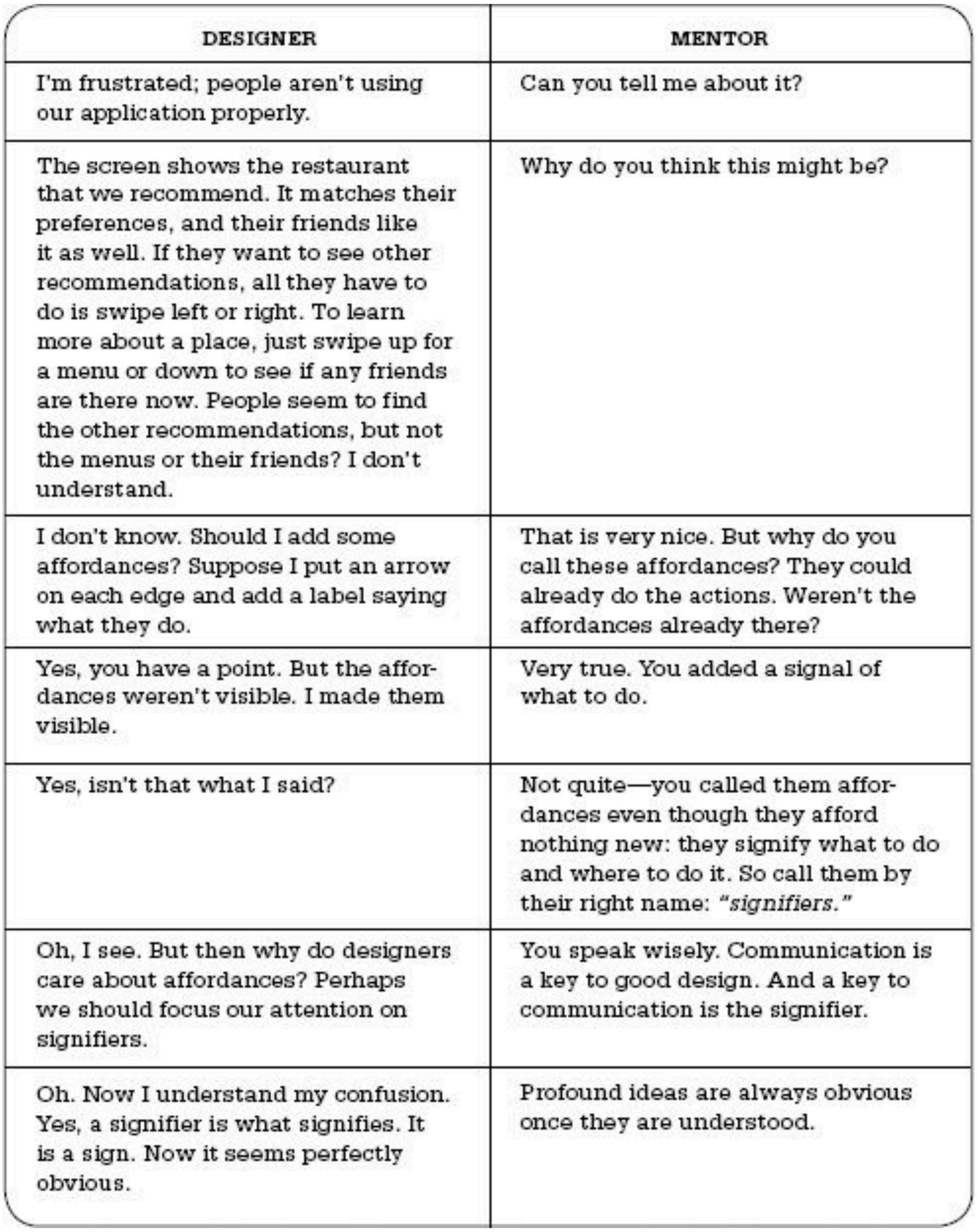
A designer approaches his mentor. He is working on a system that recommends restaurants to people, based upon their preferences and those of their friends. But in his tests, he discovered that people never used all of the features. “Why not?” he asks his mentor.

(With apologies to Socrates.)

DESIGNER	**MENTOR**

I’m frustrated; people aren’t using our application properly.	Can you tell me about it?
The screen shows the restaurant that we recommend. It matches their preferences, and their friends like it as well. If they want to see other recommendations, all they have to do is swipe left or right. To learn more about a place, just swipe up for a menu or down to see if any friends are there now. People seem to find the other recommendations, but not the menus or their friends? I don’t understand.	Why do you think this might be?
I don’t know. Should I add some affordances? Suppose I put an arrow on each edge and add a label saying what they do.	That is very nice. But why do you call these affordances? They could already do the actions. Weren’t the affordances already there?
Yes, you have a point. But the affordances weren’t visible. I made them visible.	Very true. You added a signal of what to do.
Yes, isn’t that what I said?	Not quite—you called them affordances even though they afford nothing new: they signify what to do and where to do it. So call them by their right name: *“signifiers.”*

Oh, I see. But then why do designers care about affordances? Perhaps we should focus our attention on signifiers.	You speak wisely. Communication is a key to good design. And a key to communication is the signifier.
Oh. Now I understand my confusion. Yes, a signifier is what signifies. It is a sign. Now it seems perfectly obvious.	Profound ideas are always obvious once they are understood.



DESIGNER	MENTOR
I'm frustrated; people aren't using our application properly.	Can you tell me about it?
The screen shows the restaurant that we recommend. It matches their preferences, and their friends like it as well. If they want to see other recommendations, all they have to do is swipe left or right. To learn more about a place, just swipe up for a menu or down to see if any friends are there now. People seem to find the other recommendations, but not the menus or their friends? I don't understand.	Why do you think this might be?
I don't know. Should I add some affordances? Suppose I put an arrow on each edge and add a label saying what they do.	That is very nice. But why do you call these affordances? They could already do the actions. Weren't the affordances already there?
Yes, you have a point. But the affordances weren't visible. I made them visible.	Very true. You added a signal of what to do.
Yes, isn't that what I said?	Not quite—you called them affordances even though they afford nothing new: they signify what to do and where to do it. So call them by their right name: *"signifiers."*
Oh, I see. But then why do designers care about affordances? Perhaps we should focus our attention on signifiers.	You speak wisely. Communication is a key to good design. And a key to communication is the signifier.
Oh. Now I understand my confusion. Yes, a signifier is what signifies. It is a sign. Now it seems perfectly obvious.	Profound ideas are always obvious once they are understood.

MAPPING

Mapping is a technical term, borrowed from mathematics, meaning the relationship between the elements of two sets of things. Suppose there are many lights in the ceiling of a classroom or auditorium and a row of light switches on the wall at the front of the room. The mapping of switches to lights specifies which switch controls which light.

FIGURE 1.6. Signifiers on a Touch Screen. The arrows and icons are signifiers: they provide signals about the permissible operations for this restaurant guide. Swiping left or right brings up new restaurant recommendations. Swiping up reveals the menu for the restaurant being displayed; swiping down, friends who recommend the restaurant.

Mapping is an important concept in the design and layout of controls and displays. When the mapping uses spatial correspondence between the layout of the controls and the devices being controlled, it is easy to determine how to use them. In steering a car, we rotate the steering wheel clockwise to cause the car to turn right: the top of the wheel moves in the same direction as the car. Note that other choices could have been made. In early cars, steering was controlled by a variety of devices, including tillers, handlebars, and reins. Today, some vehicles use joysticks, much as in a computer game. In cars that used tillers, steering was done much as one steers a boat: move the tiller to the left to turn to the right. Tractors, construction equipment such as bulldozers and cranes, and military tanks that have tracks instead of wheels use separate controls for the speed and direction of each track: to turn right, the left track is increased in speed, while the right track is slowed or even reversed. This is also how a wheelchair is steered.

All of these mappings for the control of vehicles work because each has a compelling conceptual model of how the operation of the control affects the vehicle. Thus, if we speed up the left wheel of a wheelchair while stopping the right wheel, it is easy to imagine the chair's pivoting on the right wheel, circling to the right. In a small boat, we can understand the tiller by realizing that pushing the tiller to the left causes the ship's rudder to move to the right and the resulting force of the water on the rudder slows down the right side of the boat, so that the boat rotates to the right. It doesn't matter whether these conceptual models are accurate: what matters is that they provide a clear way of remembering and understanding the mappings. The relationship between a control and its results is easiest to learn wherever there is an understandable mapping between the controls, the actions, and the intended result.

Natural mapping, by which I mean taking advantage of spatial analogies, leads to immediate understanding. For example, to move an object up, move the control up. To make it easy to determine which control works which light in a large room or auditorium, arrange the controls in the same pattern as the lights. Some natural mappings are cultural or biological, as in the universal standard that moving the hand up signifies more, moving it down signifies less, which is why it is appropriate to use vertical position to represent intensity or amount. Other natural mappings follow from the principles of perception and allow for the natural grouping or patterning of controls and feedback. Groupings and proximity are important principles from Gestalt psychology that can be used to map controls to function: related controls should be grouped together. Controls should be close to the item being controlled.

FIGURE 1.7. Good Mapping: Automobile Seat Adjustment Control. This is an excellent example of natural mapping. The control is in the shape of the seat itself: the mapping is straightforward. To move the front edge of the seat higher, lift up on the front part of the button. To make the seat back recline, move the button back. The

same principle could be applied to much more common objects. This particular control is from Mercedes-Benz, but this form of mapping is now used by many automobile companies. (Photograph by the author.)

Note that there are many mappings that feel “natural” but in fact are specific to a particular culture: what is natural for one culture is not necessarily natural for another. In Chapter 3, I discuss how different cultures view time, which has important implications for some kinds of mappings.

A device is easy to use when the set of possible actions is visible, when the controls and displays exploit natural mappings. The principles are simple but rarely incorporated into design. Good design takes care, planning, thought, and an understanding of how people behave.

FEEDBACK

Ever watch people at an elevator repeatedly push the Up button, or repeatedly push the pedestrian button at a street crossing? Ever drive to a traffic intersection and wait an inordinate amount of time for the signals to change, wondering all the time whether the detection circuits noticed your vehicle (a common problem with bicycles)? What is missing in all these cases is feedback: some way of letting you know that the system is working on your request.

Feedback—communicating the results of an action—is a well-known concept from the science of control and information theory. Imagine trying to hit a target with a ball when you cannot see the target. Even as simple a task as picking up a glass with the hand requires feedback to aim the hand properly, to grasp the glass, and to lift it. A misplaced hand will spill the contents, too hard a grip will break the glass, and too weak a grip will allow it to fall. The human nervous system is equipped with numerous feedback mechanisms, including visual, auditory, and touch sensors, as well as vestibular and proprioceptive systems that monitor body position and muscle and limb movements. Given the importance of feedback, it is amazing how many products ignore it.

Feedback must be immediate: even a delay of a tenth of a second can be disconcerting. If the delay is too long, people often give up, going off to do other activities. This is annoying to the people, but it can also be wasteful of resources when the system spends considerable time and effort to satisfy the request, only to find that the intended recipient is no longer there. Feedback must also be informative. Many companies try to save money by using inexpensive lights or sound generators for feedback. These simple light flashes or beeps are usually more annoying than useful. They tell us that something has happened, but convey very little information about what has happened, and then nothing about what we should do about it. When the signal is auditory, in many cases we cannot even be certain which device has created the sound. If the signal is a light, we may miss it unless our eyes are on the correct spot at the correct time. Poor feedback can be worse than no feedback at all, because it is distracting, uninformative, and in many cases irritating and anxiety-provoking.

Too much feedback can be even more annoying than too little. My dishwasher likes to beep at three a.m. to tell me that the wash is done, defeating my goal of having it work in the middle of the night so as not to disturb anyone (and to use less expensive electricity). But worst of all is inappropriate, uninterpretable feedback. The irritation caused by a “backseat driver” is well enough known that it is the staple of numerous jokes. Backseat drivers are often correct, but their remarks and comments can be so numerous and continuous that instead of helping, they become an irritating distraction. Machines that give too much feedback are like backseat drivers. Not only is it distracting to be subjected to continual flashing lights, text announcements, spoken voices, or beeps and boops, but it can be dangerous. Too many announcements cause people to ignore all of them, or wherever possible, disable all of them, which means that critical and important ones are apt to be missed. Feedback is essential, but not when it gets in the way of other things, including a calm and relaxing environment.

Poor design of feedback can be the result of decisions aimed at reducing costs, even if they make life more difficult for people. Rather than use multiple signal lights, informative displays, or rich, musical sounds with varying patterns, the focus upon cost reduction forces the design to use a single light or sound to convey multiple types of information. If the choice is to use a light, then one flash might mean one thing; two rapid flashes, something else. A long flash might signal yet another state; and a long flash followed by a brief one, yet another. If the choice is to use a sound, quite often the least expensive sound device is selected, one that can only produce a high-frequency beep. Just as with the lights, the only way to signal different states of the machine is by beeping different patterns. What do all these different patterns mean? How can we possibly learn and remember them? It doesn't help that every different machine uses a different pattern of lights or beeps, sometimes with the same patterns meaning contradictory things for different machines. All the beeps sound alike, so it often isn't even possible to know which machine is talking to us.

Feedback has to be planned. All actions need to be confirmed, but in a manner that is unobtrusive. Feedback must also be prioritized, so that unimportant information is presented in an unobtrusive fashion, but important signals are presented in a way that does capture attention. When there are major emergencies, then even important signals have to be prioritized. When every device is signaling a major emergency, nothing is gained by the resulting cacophony. The continual beeps and alarms of equipment can be dangerous. In many emergencies, workers have to spend valuable time turning off all the alarms because the sounds interfere with the concentration required to solve the problem. Hospital operating rooms, emergency wards. Nuclear power control plants. Airplane cockpits. All can become confusing, irritating, and life-endangering places because of excessive feedback, excessive alarms, and incompatible message coding. Feedback is essential, but it has to be done correctly. Appropriately.

CONCEPTUAL MODELS

A conceptual model is an explanation, usually highly simplified, of how something works. It doesn't have to be complete or even accurate as long as it is useful. The files, folders, and icons you see displayed on a computer screen help people create the conceptual model of documents and folders inside the computer, or of apps or applications residing on the screen, waiting to be summoned. In fact, there are no folders inside the computer—those are effective conceptualizations designed to make them easier to use. Sometimes these depictions can add to the confusion, however. When reading e-mail or visiting a website, the material appears to be on the device, for that is where it is displayed and manipulated. But in fact, in many cases the actual material is “in the cloud,” located on some distant machine. The conceptual model is of one, coherent image, whereas it may actually consist of parts, each located on different machines that could be almost anywhere in the world. This simplified model is helpful for normal usage, but if the network connection to the cloud services is interrupted, the result can be confusing. Information is still on their screen, but users can no longer save it or retrieve new things: their conceptual model offers no explanation. Simplified models are valuable only as long as the assumptions that support them hold true.

There are often multiple conceptual models of a product or device. People’s conceptual models for the way that regenerative braking in a hybrid or electrically powered automobile works are quite different for average drivers than for technically sophisticated drivers, different again for whoever must service the system, and yet different again for those who designed the system.

Conceptual models found in technical manuals and books for technical use can be detailed and complex. The ones we are concerned with here are simpler: they reside in the minds of the people who are using the product, so they are also “mental models.” Mental models, as the name implies, are the conceptual models in people’s minds that represent their understanding of how things work. Different people may hold different mental models of the same item. Indeed, a single person might have multiple models of the same item, each dealing with a different aspect of its operation: the models can even be in conflict.

Conceptual models are often inferred from the device itself. Some models are passed on from person to person. Some come from manuals. Usually the device itself offers very little assistance, so the model is constructed by experience. Quite often these models are erroneous, and therefore lead to difficulties in using the device.

The major clues to how things work come from their perceived structure—in particular from signifiers, affordances, constraints, and mappings. Hand tools for the shop, gardening, and the house tend to make their critical parts sufficiently visible that conceptual models of their operation and function are readily derived. Consider a pair of scissors: you can see that the number of possible actions is limited. The holes are clearly there to put something into, and the only logical things that will fit are fingers. The holes are both affordances—they allow the fingers to be inserted—and signifiers—they indicate where the fingers are to go. The sizes of the holes provide constraints to limit the possible fingers: a big hole suggests several fingers; a small hole, only one. The mapping between holes and fingers—the set of possible operations—is signified and constrained by the holes. Moreover, the operation is not sensitive to finger placement: if you use the wrong fingers (or the wrong hand), the scissors still work, although not as comfortably. You can figure out the scissors because their operating parts are visible and the implications clear. The conceptual model is obvious, and there is effective use of signifiers, affordances, and constraints.

FIGURE 1.8. Junghans Mega 1000 Digital Radio Controlled Watch. There is no good conceptual model for understanding the operation of my watch. It has five buttons with no hints as to what each one does. And yes, the buttons do different things in their different modes. But it is a very nice-looking watch, and always has the exact time because it checks official radio time stations. (The top row of the display is the date: Wednesday, February 20, the eighth week of the year.) (Photograph by the author.)

What happens when the device does not suggest a good conceptual model? Consider my digital watch with five buttons: two along the top, two along the bottom, and one on the left side (Figure 1.8). What is each button for? How would you set the time? There is no way to tell—no evident relationship between the operating controls and the functions, no constraints, no apparent mappings. Moreover, the buttons have multiple ways of being used. Two of the buttons do different things when pushed quickly or when kept depressed for several seconds.

Some operations require simultaneous depression of several of the buttons. The only way to tell how to work the watch is to read the manual, over and over again. With the scissors, moving the handle makes the blades move. The watch provides no visible relationship between the buttons and the possible actions, no discernible relationship between the actions and the end results. I really like the watch: too bad I can't remember all the functions.

Conceptual models are valuable in providing understanding, in predicting how things will behave, and in figuring out what to do when things do not go as planned. A good conceptual model allows us to predict the effects of our actions. Without a good model, we operate by rote, blindly; we do operations as we were told to do them; we can't fully appreciate why, what effects to expect, or what to do if things go wrong. As long as things work properly, we can manage. When things go wrong, however, or when we come upon a novel situation, then we need a deeper understanding, a good model.

For everyday things, conceptual models need not be very complex. After all, scissors, pens, and light switches are pretty simple devices. There is no need to understand the underlying physics or chemistry of each device we own, just the relationship between the controls and the outcomes. When the model presented to us is inadequate or wrong (or, worse, nonexistent), we can have difficulties. Let me tell you about my refrigerator.

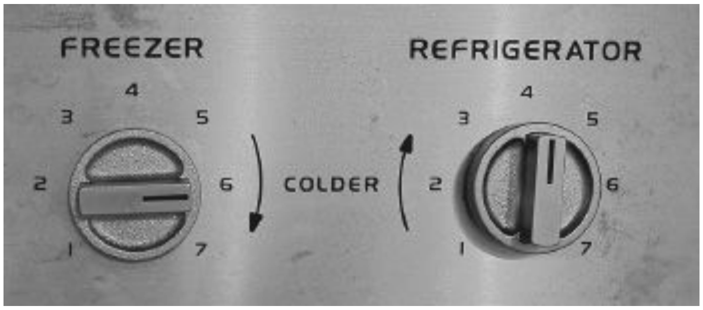


FIGURE 1.9. Refrigerator Controls. Two compartments— fresh food and freezer—and two controls (in the fresh food unit). Your task: Suppose the freezer is too cold, the fresh food section just right. How would you adjust the controls so as to make the freezer warmer and keep the fresh food the same? (Photograph by the author.)

I used to own an ordinary, two-compartment refrigerator—nothing very fancy about it. The problem was that I couldn't set the temperature properly. There were only two things to do: adjust the temperature of the freezer compartment and adjust the temperature of the fresh food compartment. And there were two controls, one labeled "freezer," the other "refrigerator." What's the problem?

Oh, perhaps I'd better warn you. The two controls are not independent. The freezer control also affects the fresh food temperature, and the fresh food control also affects the freezer. Moreover, the manual warns that one should "always allow twenty-four (24) hours for the temperature to stabilize whether setting the controls for the first time or making an adjustment."

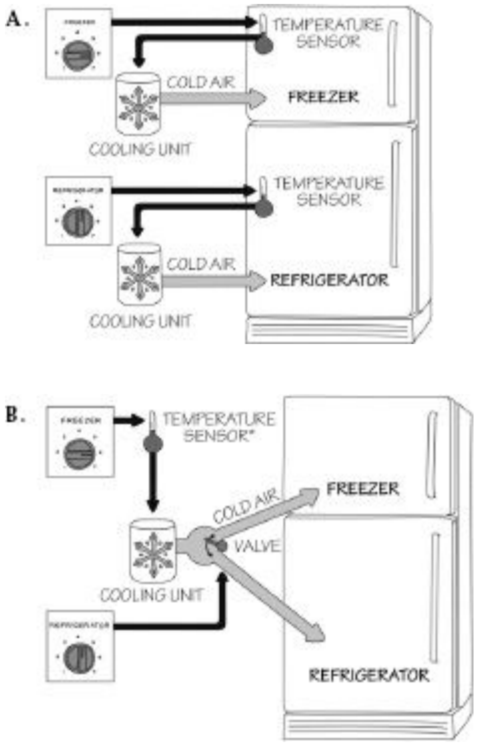


FIGURE 1.10. Two Conceptual Models for a Refrigerator. The conceptual model A is provided by the system image of the refrigerator as gleaned from the controls. Each control determines the temperature of the named part of the refrigerator. This means that each compartment has its own temperature sensor and cooling unit. This is wrong. The correct conceptual model is shown in B. There is no way of knowing where the temperature sensor is located so it is shown outside

the refrigerator. The freezer control determines the freezer temperature (so is this where the sensor is located?). The refrigerator control determines how much of the cold air goes to the freezer and how much to the refrigerator.

It was extremely difficult to regulate the temperature of my old refrigerator. Why? Because the controls suggest a false conceptual model. Two compartments, two controls, which implies that each control is responsible for the temperature of the compartment that carries its name: this conceptual model is shown in Figure 1.10A. It is wrong. In fact, there is only one thermostat and only one cooling mechanism. One control adjusts the thermostat setting, the other the relative proportion of cold air sent to each of the two compartments of the refrigerator. This is why the two controls interact: this conceptual model is shown in Figure 1.10B. In addition, there must be a temperature sensor, but there is no way of knowing where it is located. With the conceptual model suggested by the controls, adjusting the temperatures is almost impossible and always frustrating. Given the correct model, life would be much easier.

Why did the manufacturer suggest the wrong conceptual model? We will never know. In the twenty-five years since the publication of the first edition of this book, I have had many letters from people thanking me for explaining their confusing refrigerator, but never any communication from the manufacturer (General Electric). Perhaps the designers thought the correct model was too complex, that the model they were giving was easier to understand. But with the wrong conceptual model, it was impossible to set the controls. And even though I am convinced I knew the correct model, I still couldn’t accurately adjust the temperatures because the refrigerator design made it impossible to discover which control was for the temperature sensor, which for the relative proportion of cold air, and in which compartment the sensor was located. The lack of immediate feedback for the actions did not help: it took twenty-four hours to see whether the new setting was appropriate. I shouldn’t have to keep a laboratory notebook and do controlled experiments just to set the temperature of my refrigerator.

I am happy to say that I no longer own that refrigerator. Instead I have one that has two separate controls, one in the fresh food compartment, one in the freezer compartment. Each control is nicely calibrated in degrees and labeled with the name of the compartment it controls. The two compartments are independent: setting the temperature in one has no effect on the temperature in the other. This solution, although ideal, does cost more. But far less expensive solutions are possible. With today's inexpensive sensors and motors, it should be possible to have a single cooling unit with a motor-controlled valve controlling the relative proportion of cold air diverted to each compartment. A simple, inexpensive computer chip could regulate the cooling unit and valve position so that the temperatures in the two compartments match their targets. A bit more work for the engineering design team? Yes, but the results would be worth it. Alas, General Electric is still selling refrigerators with the very same controls and mechanisms that cause so much confusion. The photograph in Figure 1.9 is from a contemporary refrigerator, photographed in a store while preparing this book.

The System Image

People create mental models of themselves, others, the environment, and the things with which they interact. These are conceptual models formed through experience, training, and instruction. These models serve as guides to help achieve our goals and in understanding the world.

How do we form an appropriate conceptual model for the devices we interact with? We cannot talk to the designer, so we rely upon whatever information is available to us: what the device looks like, what we know from using similar things in the past, what was told to us in the sales literature, by salespeople and advertisements, by articles we may have read, by the product website and instruction manuals. I call the combined information available to us the *system image*. When the system image is incoherent or inappropriate, as in the case of the refrigerator, then the user cannot easily use the device. If it is incomplete or contradictory, there will be trouble.

As illustrated in Figure 1.11, the designer of the product and the person using the product form somewhat disconnected vertices of a triangle. The designer's conceptual model is the designer's conception of the product, occupying one vertex of the triangle. The product itself is no longer with the designer, so it is isolated as a second vertex, perhaps sitting on the user's kitchen counter. The system image is what can be perceived from the physical structure that has been built (including documentation, instructions, signifiers, and any information available from websites and help lines). The user's conceptual model comes from the system image, through interaction with the product, reading, searching for online information, and from whatever manuals are provided. The designer expects the user's model to be identical to the design model, but because designers cannot communicate directly with users, the entire burden of communication is on the system image.

FIGURE 1.11. The Designer's Model, the User's Model, and the System Image. The designer's conceptual model is the designer's conception of the look, feel, and operation of a product. The system image is what can be derived from the physical structure that has been built (including documentation). The user's mental model is developed through interaction with the product and the system image. Designers expect the user's model to be identical to their own, but because they cannot communicate directly with the user, the burden of communication is with the system image.

Figure 1.11 indicates why communication is such an important aspect of good design. No matter how brilliant the product, if people cannot use it, it will receive poor reviews. It is up to the designer to provide the appropriate information to make the product understandable and usable. Most important is the provision of a good conceptual model that guides the user when thing go wrong. With a good conceptual model, people can figure out what has happened and correct the things that went wrong. Without a good model, they struggle, often making matters worse.

Good conceptual models are the key to understandable, enjoyable products: good communication is the key to good conceptual models.

The Paradox of Technology

Technology offers the potential to make life easier and more enjoyable; each new technology provides increased benefits. At the same time, added complexities increase our difficulty and frustration with technology. The design problem posed by technological advances is enormous. Consider the wristwatch. A few decades ago, watches were simple. All you had to do was set the time and keep the watch wound. The standard control was the stem: a knob at the side of the watch. Turning the knob would wind the spring that provided power to the watch movement. Pulling out the knob and turning it rotated the hands. The operations were easy to learn and easy to do. There was a reasonable relationship between the turning of the knob and the resulting turning of the hands. The design even took into account human error. In its normal position, turning the stem wound the mainspring of the clock. The stem had to be pulled before it would engage the gears for setting the time. Accidental turns of the stem did no harm.

Watches in olden times were expensive instruments, manufactured by hand. They were sold in jewelry stores. Over time, with the introduction of digital technology, the cost of watches decreased rapidly, while their accuracy and reliability increased. Watches became tools, available in a wide variety of styles and shapes and with an ever-increasing number of functions. Watches were sold everywhere, from local shops to sporting goods stores to electronic stores. Moreover, accurate clocks were incorporated in many appliances, from phones to musical keyboards: many people no longer felt the need to wear a watch. Watches became inexpensive enough that the average person could own multiple watches. They became fashion accessories, where one changed the watch with each change in activity and each change of clothes.

In the modern digital watch, instead of winding the spring, we change the battery, or in the case of a solar-powered watch, ensure that it gets its weekly dose of light. The technology has allowed more functions: the watch can give the day of the week, the month, and the year; it can act as a stopwatch (which itself has several functions), a countdown timer, and an alarm clock (or two); it has the ability to show the time for different time zones; it can act as a counter and even as a calculator. My watch, shown in Figure 1.8, has many functions. It even has a radio receiver to allow it to set its time with official time stations around the world. Even so, it is far less complex than many that are available. Some watches have built-in compasses and barometers, accelerometers, and temperature gauges. Some have GPS and Internet receivers so they can display the weather and news, e-mail messages, and the latest from social networks. Some have built-in cameras. Some work with buttons, knobs, motion, or speech. Some detect gestures. The watch is no longer just an instrument for telling time: it has become a platform for enhancing multiple activities and lifestyles.

The added functions cause problems: How can all these functions fit into a small, wearable size? There are no easy answers. Many people have solved the problem by not using a watch. They use their phone instead. A cell phone performs all the functions much better than the tiny watch, while also displaying the time.

Now imagine a future where instead of the phone replacing the watch, the two will merge, perhaps worn on the wrist, perhaps on the head like glasses, complete with display screen. The phone, watch, and components of a computer will all form one unit. We will have flexible displays that show only a tiny amount of information in their normal state, but that can unroll to considerable size. Projectors will be so small and light that they can be built into watches or phones (or perhaps rings and other jewelry), projecting their images onto any convenient surface. Or perhaps our devices won’t have displays, but will quietly whisper the results into our ears, or simply use whatever display happens to be available: the display in the seatback of cars or airplanes, hotel room televisions, whatever is nearby. The devices will be able to do many useful things, but I fear they will also frustrate: so many things to control, so little space for controls or signifiers. The obvious solution is to use exotic gestures or spoken commands, but how will we learn, and then remember, them? As I discuss later, the best solution is for there to be agreed upon standards, so we need learn the controls only once. But as I also discuss, agreeing upon these is a complex process, with many competing forces hindering rapid resolution. We will see.

The same technology that simplifies life by providing more functions in each device also complicates life by making the device harder to learn, harder to use. This is the paradox of technology and the challenge for the designer.

The Design Challenge

Design requires the cooperative efforts of multiple disciplines. The number of different disciplines required to produce a successful product is staggering. Great design requires great designers, but that isn't enough: it also requires great management, because the hardest part of producing a product is coordinating all the many, separate disciplines, each with different goals and priorities. Each discipline has a different perspective of the relative importance of the many factors that make up a product. One discipline argues that it must be usable and understandable, another that it must be attractive, yet another that it has to be affordable. Moreover, the device has to be reliable, be able to be manufactured and serviced. It must be distinguishable from competing products and superior in critical dimensions such as price, reliability, appearance, and the functions it provides. Finally, people have to actually purchase it. It doesn't matter how good a product is if, in the end, nobody uses it.

Quite often each discipline believes its distinct contribution to be most important: "Price," argues the marketing representative, "price plus these features." "Reliable," insist the engineers. "We have to be able to manufacture it in our existing plants," say the manufacturing representatives. "We keep getting service calls," say the support people; "we need to solve those problems in the design." "You can't put all that together and still have a reasonable product," says the design team. Who is right? Everyone is right. The successful product has to satisfy all these requirements.

The hard part is to convince people to understand the viewpoints of the others, to abandon their disciplinary viewpoint and to think of the design from the viewpoints of the person who buys the product and those who use it, often different people. The viewpoint of the business is also important, because it does not matter how wonderful the product is if not enough people buy it. If a product does not sell, the company must often stop producing it, even if it is a great product. Few companies can sustain the huge cost of keeping an unprofitable product alive long enough for its sales to reach profitability—with new products, this period is usually measured in years, and sometimes, as with the adoption of high-definition television, decades.

Designing well is not easy. The manufacturer wants something that can be produced economically. The store wants something that will be attractive to its customers. The purchaser has several demands. In the store, the purchaser focuses on price and appearance, and perhaps on prestige value. At home, the same person will pay more attention to functionality and usability. The repair service cares about maintainability: how easy is the device to take apart, diagnose, and service? The needs of those concerned are different and often conflict. Nonetheless, if the design team has representatives from all the constituencies present at the same time, it is often possible to reach satisfactory solutions for all the needs. It is when the disciplines operate independently of one another that major clashes and deficiencies occur. The challenge is to use the principles of human-centered design to produce positive results, products that enhance lives and add to our pleasure and enjoyment. The goal is to produce a great product, one that is successful, and that customers love. It can be done.

CHAPTER TWO

THE PSYCHOLOGY OF EVERYDAY ACTIONS

During my family's stay in England, we rented a furnished house while the owners were away. One day, our landlady returned to the house to get some personal papers. She walked over to the old, metal filing cabinet and attempted to open the top drawer. It wouldn't open. She pushed it forward and backward, right and left, up and down, without success. I offered to help. I wiggled the drawer. Then I twisted the front panel, pushed down hard, and banged the front with the palm of one hand. The cabinet drawer slid open. "Oh," she said, "I'm sorry. I am so bad at mechanical things." No, she had it backward. It is the mechanical thing that should be apologizing, perhaps saying, "I'm sorry. I am so bad with people."

My landlady had two problems. First, although she had a clear goal (retrieve some personal papers) and even a plan for achieving that goal (open the top drawer of the filing cabinet, where those papers are kept), once that plan failed, she had no idea of what to do. But she also had a second problem: she thought the problem lay in her own lack of ability: she blamed herself, falsely.

How was I able to help? First, I refused to accept the false accusation that it was the fault of the landlady: to me, it was clearly a fault in the mechanics of the old filing cabinet that prevented the drawer from opening. Second, I had a conceptual model of how the cabinet worked, with an internal mechanism that held the door shut in normal usage, and the belief that the drawer mechanism was probably out of alignment. This conceptual model gave me a plan: wiggle the drawer. That failed. That caused me to modify my plan: wiggling may have been appropriate but not forceful enough, so I resorted to brute force to try to twist the cabinet back into its proper alignment. This felt good to me—the cabinet drawer moved slightly—but it still didn't open. So I resorted to the most powerful tool employed by experts the world around—I banged on the cabinet. And yes, it opened. In my mind, I decided (without any evidence) that my hit had jarred the mechanism sufficiently to allow the drawer to open.

This example highlights the themes of this chapter. First, how do people do things? It is easy to learn a few basic steps to perform operations with our technologies (and yes, even filing cabinets are technology). But what happens when things go wrong? How do we detect that they aren't working, and then how do we know what to do? To help understand this, I first delve into human psychology and a simple conceptual model of how people select and then evaluate their actions. This leads the discussion to the role of understanding (via a conceptual model) and of emotions: pleasure when things work smoothly and frustration when our plans are thwarted. Finally, I conclude with a summary of how the lessons of this chapter translate into principles of design.

How People Do Things: The Gulfs of Execution and Evaluation

When people use something, they face two gulfs: the Gulf of Execution, where they try to figure out how it operates, and the Gulf of Evaluation, where they try to figure out what happened (Figure 2.1). The role of the designer is to help people bridge the two gulfs.

In the case of the filing cabinet, there were visible elements that helped bridge the Gulf of Execution when everything was working perfectly. The drawer handle clearly signified that it should be pulled and the slider on the handle indicated how to release the catch that normally held the drawer in place. But when these operations failed, there then loomed a big gulf: what other operations could be done to open the drawer?

The Gulf of Evaluation was easily bridged, at first. That is, the catch was released, the drawer handle pulled, yet nothing happened. The lack of action signified a failure to reach the goal. But when other operations were tried, such as my twisting and pulling, the filing cabinet provided no more information about whether I was getting closer to the goal.

The Gulf of Evaluation reflects the amount of effort that the person must make to interpret the physical state of the device and to determine how well the expectations and intentions have been met. The gulf is small when the device provides information about its state in a form that is easy to get, is easy to interpret, and matches the way the person thinks about the system. What are the major design elements that help bridge the Gulf of Evaluation? Feedback and a good conceptual model.

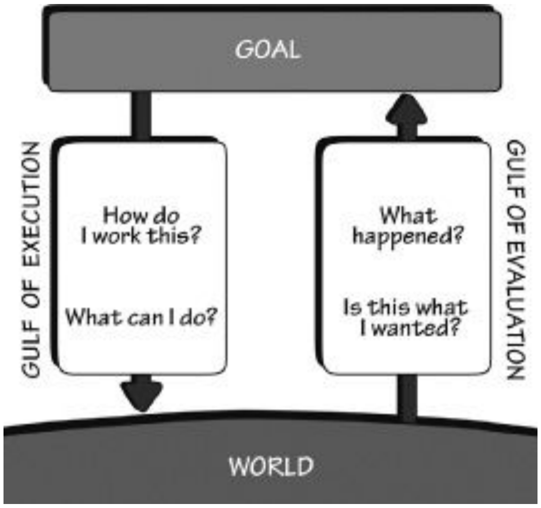


FIGURE 2.1. The Gulfs of Execution and Evaluation. When people encounter a device, they face two gulfs: the Gulf of Execution, where they try to figure out how to use it, and the Gulf of Evaluation, where they try to figure out what state it is in and whether their actions got them to their goal.

The gulfs are present for many devices. Interestingly, many people do experience difficulties, but explain them away by blaming themselves. In the case of things they believe they should be capable of using—water faucets, refrigerator temperature controls, stove tops—they simply think, “I’m being stupid.” Alternatively, for complicated-looking devices—sewing machines, washing machines, digital watches, or almost any digital controls—they simply give up, deciding that they are incapable of understanding them. Both explanations are wrong. These are the things of everyday household use. None of them has a complex underlying structure. The difficulties reside in their design, not in the people attempting to use them.

How can the designer help bridge the two gulfs? To answer that question, we need to delve more deeply into the psychology of human action. But the basic tools have already been discussed: We bridge the Gulf of Execution through the use of signifiers, constraints, mappings, and a conceptual model. We bridge the Gulf of Evaluation through the use of feedback and a conceptual model.

The Seven Stages of Action

There are two parts to an action: executing the action and then evaluating the results: doing and interpreting. Both execution and evaluation require understanding: how the item works and what results it produces. Both execution and evaluation can affect our emotional state.

Suppose I am sitting in my armchair, reading a book. It is dusk, and the light is getting dimmer and dimmer. My current activity is reading, but that goal is starting to fail because of the decreasing illumination. This realization triggers a new goal: get more light. How do I do that? I have many choices. I could open the curtains, move so that I sit where there is more light, or perhaps turn on a nearby light. This is the planning stage, determining which of the many possible plans of action to follow. But even when I decide to turn on the nearby light, I still have to determine how to get it done. I could ask someone to do it for me, I could use my left hand or my right. Even after I have decided upon a plan, I still have to specify how I will do it. Finally, I must execute—do—the action. When I am doing a frequent act, one for which I am quite experienced and skilled, most of these stages are subconscious. When I am still learning how to do it, determining the plan, specifying the sequence, and interpreting the result are conscious.

Suppose I am driving in my car and my action plan requires me to make a left turn at a street intersection. If I am a skilled driver, I don’t have to give much conscious attention to specify or perform the action sequence. I think “left” and smoothly execute the required action sequence. But if I am just learning to drive, I have to think about each separate component of the action. I must apply the brakes and check for cars behind and around me, cars and pedestrians in front of me, and whether there are traffic signs or signals that I have to obey. I must move my feet back and forth between pedals and my hands to the turn signals and back to the steering wheel (while I try to remember just how my instructor told me I should position my hands while making a turn), and my visual attention is divided among all the activity around me, sometimes looking directly, sometimes rotating my head, and sometimes using the rear- and side-view mirrors. To the skilled driver, it is all easy and straightforward. To the beginning driver, the task seems impossible.

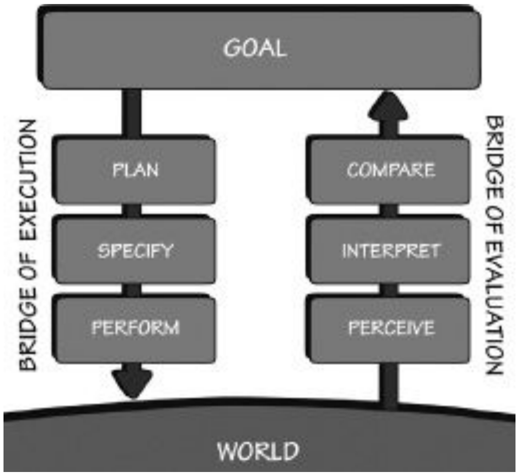


FIGURE 2.2. The Seven Stages of the Action Cycle. Putting all the stages together yields the three stages of execution (plan, specify, and perform), three stages of evaluation (perceive, interpret, and compare), and, of course, the goal: seven stages in all.

The specific actions bridge the gap between what we would like to have done (our goals) and all possible physical actions to achieve those goals. After we specify what actions to make, we must actually do them—the stages of execution. There are three stages of execution that follow from the goal: plan, specify, and perform (the left side of Figure 2.2). Evaluating what happened has three stages: first, perceiving what happened in the world; second, trying to make sense of it (interpreting it); and, finally, comparing what happened with what was wanted (the right side of Figure 2.2).

There we have it. Seven stages of action: one for goals, three for execution, and three for evaluation (Figure 2.2).

1. **Goal** (form the goal)
2. **Plan** (the action)
3. **Specify** (an action sequence)
4. **Perform** (the action sequence)
5. **Perceive** (the state of the world)
6. **Interpret** (the perception)
7. **Compare** (the outcome with the goal)

The seven-stage action cycle is simplified, but it provides a useful framework for understanding human action and for guiding design. It has proven to be helpful in designing interaction. Not all of the activity in the stages is conscious. Goals tend to be, but even they may be subconscious. We can do many actions, repeatedly cycling through the stages while being blissfully unaware that we are doing so. It is only when we come across something new or reach some impasse, some problem that disrupts the normal flow of activity, that conscious attention is required.

Most behavior does not require going through all stages in sequence; however, most activities will not be satisfied by single actions. There must be numerous sequences, and the whole activity may last hours or even days. There are multiple feedback loops in which the results of one activity are used to direct further ones, in which goals lead to subgoals, and plans lead to subplans. There are activities in which goals are forgotten, discarded, or reformulated.

Let’s go back to my act of turning on the light. This is a case of event-driven behavior: the sequence starts with the world, causing evaluation of the state and the formulation of a goal. The trigger was an environmental event: the lack of light, which made reading difficult. This led to a violation of the goal of reading, so it led to a subgoal—get more light. But reading was not the high-level goal. For each goal, one has to ask, “Why is that the goal?” Why was I reading? I was trying to prepare a meal using a new recipe, so I needed to reread it before I started. Reading was thus a subgoal. But cooking was itself a subgoal. I was cooking in order to eat, which had the goal of satisfying my hunger. So the hierarchy of goals is roughly: satisfy hunger; eat; cook; read cookbook; get more light. This is called a root cause analysis: asking “Why?” until the ultimate, fundamental cause of the activity is reached.

The action cycle can start from the top, by establishing a new goal, in which case we call it goal-driven behavior. In this situation, the cycle starts with the goal and then goes through the three stages of execution. But the action cycle can also start from the bottom, triggered by some event in the world, in which case we call it either data-driven or event-driven behavior. In this situation, the cycle starts with the environment, the world, and then goes through the three stages of evaluation.

For many everyday tasks, goals and intentions are not well specified: they are opportunistic rather than planned. Opportunistic actions are those in which the behavior takes advantage of circumstances. Rather than engage in extensive planning and analysis, we go about the day’s activities and do things as opportunities arise. Thus, we may not have planned to try a new café or to ask a question of a friend. Rather, we go through the day’s activities, and if we find ourselves near the café or encountering the friend, then we allow the opportunity to trigger the appropriate activity. Otherwise, we might never get to that café or ask our friend the question. For crucial tasks we make special efforts to ensure that they get done. Opportunistic actions are less precise and certain than specified goals and intentions, but they result in less mental effort, less inconvenience, and perhaps more interest. Some of us adjust our lives around the expectation of opportunities. And sometimes, even for goal-driven behavior, we try to create world events that will ensure that the sequence gets completed. For example, sometimes when I must do an important task, I ask someone to set a deadline for me. I use the approach of that deadline to trigger the work. It may only be a few hours before the deadline that I actually get to work and do the job, but the important point is that it does get done.

This self-triggering of external drivers is fully compatible with the seven-stage analysis.

The seven stages provide a guideline for developing new products or services. The gulfs are obvious places to start, for either gulf, whether of execution or evaluation, is an opportunity for product enhancement. The trick is to develop observational skills to detect them. Most innovation is done as an incremental enhancement of existing products. What about radical ideas, ones that introduce new product categories to the marketplace? These come about by reconsidering the goals, and always asking what the real goal is: what is called the *root cause* analysis.

Harvard Business School marketing professor Theodore Levitt once pointed out, “People don’t want to buy a quarter-inch drill. They want a quarter-inch hole!” Levitt’s example of the drill implying that the goal is really a hole is only partially correct, however. When people go to a store to buy a drill, that is not their real goal. But why would anyone want a quarter-inch hole? Clearly that is an intermediate goal. Perhaps they wanted to hang shelves on the wall. Levitt stopped too soon.

Once you realize that they don’t really want the drill, you realize that perhaps they don’t really want the hole, either: they want to install their bookshelves. Why not develop methods that don’t require holes? Or perhaps books that don’t require bookshelves. (Yes, I know: electronic books, e-books.)

Human Thought: Mostly Subconscious

Why do we need to know about the human mind? Because things are designed to be used by people, and without a deep understanding of people, the designs are apt to be faulty, difficult to use, difficult to understand. That is why it is useful to consider the seven stages of action. The mind is more difficult to comprehend than actions. Most of us start by believing we already understand both human behavior and the human mind. After all, we are all human: we have all lived with ourselves all of our lives, and we like to think we understand ourselves. But the truth is, we don’t. Most of human behavior is a result of subconscious processes. We are unaware of them. As a result, many of our beliefs about how people behave—including beliefs about ourselves—are wrong. That is why we have the multiple social and behavioral sciences, with a good dash of mathematics, economics, computer science, information science, and neuroscience.

Consider the following simple experiment. Do all three steps:

1. Wiggle the second finger of your hand.
2. Wiggle the third finger of the same hand.
3. Describe what you did differently those two times.

On the surface, the answer seems simple: I thought about moving my fingers and they moved. The difference is that I thought about a different finger each time. Yes, that's true. But how did that thought get transmitted into action, into the commands that caused different muscles in the arm to control the tendons that wiggled the fingers? This is completely hidden from consciousness.

The human mind is immensely complex, having evolved over a long period with many specialized structures. The study of the mind is the subject of multiple disciplines, including the behavioral and social sciences, cognitive science, neuroscience, philosophy, and the information and computer sciences. Despite many advances in our understanding, much still remains mysterious, yet to be learned. One of the mysteries concerns the nature of and distinction between those activities that are conscious and those that are not. Most of the brain's operations are subconscious, hidden beneath our awareness. It is only the highest level, what I call *reflective*, that is conscious.

Conscious attention is necessary to learn most things, but after the initial learning, continued practice and study, sometimes for thousands of hours over a period of years, produces what psychologists call "overlearning," Once skills have been overlearned, performance appears to be effortless, done automatically, with little or no awareness. For example, answer these questions:

What is the phone number of a friend?

What is Beethoven’s phone number?

What is the capital of:

- Brazil?
- Wales?
- The United States?
- Estonia?

Think about how you answered these questions. The answers you knew come immediately to mind, but with no awareness of how that happened. You simply “know” the answer. Even the ones you got wrong came to mind without any awareness. You might have been aware of some doubt, but not of how the name entered your consciousness. As for the countries for which you didn’t know the answer, you probably knew you didn’t know those immediately, without effort. Even if you knew you knew, but couldn’t quite recall it, you didn’t know how you knew that, or what was happening as you tried to remember.

You might have had trouble with the phone number of a friend because most of us have turned over to our technology the job of remembering phone numbers. I don’t know anybody’s phone number—I barely remember my own. When I wish to call someone, I just do a quick search in my contact list and have the telephone place the call. Or I just push the “2” button on the phone for a few seconds, which autodials my home. Or in my auto, I can simply speak: “Call home.” What’s the number? I don’t know: my technology knows. Do we count our technology as an extension of our memory systems? Of our thought processes? Of our mind?

What about Beethoven’s phone number? If I asked my computer, it would take a long time, because it would have to search all the people I know to see whether any one of them was Beethoven. But you immediately discarded the question as nonsensical. You don’t personally know Beethoven. And anyway, he is dead. Besides, he died in the early 1800s and the phone wasn’t invented until the late 1800s.

How do we know what we do not know so rapidly? Yet some things that we do know can take a long time to retrieve. For example, answer this:

In the house you lived in three houses ago, as you entered the front door, was the doorknob on the left or right?

Now you have to engage in conscious, reflective problem solving, first to retrieve just which house is being talked about, and then what the correct answer is. Most people can determine the house, but have difficulty answering the question because they can readily imagine the doorknob on both sides of the door. The way to solve this problem is to imagine doing some activity, such as walking up to the front door while carrying heavy packages with both hands: how do you open the door? Alternatively, visualize yourself inside the house, rushing to the front door to open it for a visitor. Usually one of these imagined scenarios provides the answer. But note how different the memory retrieval for this question was from the retrieval for the others. All these questions involved long-term memory, but in very different ways. The earlier questions were memory for factual information, what is called *declarative memory*. The last question could have been answered factually, but is usually most easily answered by recalling the activities performed to open the door. This is called *procedural memory*. I return to a discussion of human memory in Chapter 3.

Walking, talking, reading. Riding a bicycle or driving a car. Singing. All of these skills take considerable time and practice to master, but once mastered, they are often done quite automatically. For experts, only especially difficult or unexpected situations require conscious attention.

Because we are only aware of the reflective level of conscious processing, we tend to believe that all human thought is conscious. But it isn’t. We also tend to believe that thought can be separated from emotion. This is also false. Cognition and emotion cannot be separated. Cognitive thoughts lead to emotions: emotions drive cognitive thoughts. The brain is structured to act upon the world, and every action carries with it expectations, and these expectations drive emotions. That is why much of language is based on physical metaphors, why the body and its interaction with the environment are essential components of human thought.

Emotion is highly underrated. In fact, the emotional system is a powerful information processing system that works in tandem with cognition. Cognition attempts to make sense of the world: emotion assigns value. It is the emotional system that determines whether a situation is safe or threatening, whether something that is happening is good or bad, desirable or not. Cognition provides understanding: emotion provides value judgments. A human without a working emotional system has difficulty making choices. A human without a cognitive system is dysfunctional.

Because much human behavior is subconscious—that is, it occurs without conscious awareness—we often don't know what we are about to do, say, or think until after we have done it. It's as if we had two minds: the subconscious and the conscious, which don't always talk to each other. Not what you have been taught? True, nonetheless. More and more evidence is accumulating that we use logic and reason after the fact, to justify our decisions to ourselves (to our conscious minds) and to others. Bizarre? Yes, but don't protest: enjoy it.

Subconscious thought matches patterns, finding the best possible match of one's past experience to the current one. It proceeds rapidly and automatically, without effort. Subconscious processing is one of our strengths. It is good at detecting general trends, at recognizing the relationship between what we now experience and what has happened in the past. And it is good at generalizing, at making predictions about the general trend, based on few examples. But subconscious thought can find matches that are inappropriate or wrong, and it may not distinguish the common from the rare. Subconscious thought is biased toward regularity and structure, and it is limited in formal power. It may not be capable of symbolic manipulation, of careful reasoning through a sequence of steps.

Conscious thought is quite different. It is slow and labored. Here is where we slowly ponder decisions, think through alternatives, compare different choices. Conscious thought considers first this approach, then that—comparing, rationalizing, finding explanations. Formal logic, mathematics, decision theory: these are the tools of conscious thought. Both conscious and subconscious modes of thought are powerful and essential aspects of human life. Both can provide insightful leaps and creative moments. And both are subject to errors, misconceptions, and failures.

Emotion interacts with cognition biochemically, bathing the brain with hormones, transmitted either through the bloodstream or through ducts in the brain, modifying the behavior of brain cells. Hormones exert powerful biases on brain operation. Thus, in tense, threatening situations, the emotional system triggers the release of hormones that bias the brain to focus upon relevant parts of the environment. The muscles tense in preparation for action. In calm, nonthreatening situations, the emotional system triggers the release of hormones that relax the muscles and bias the brain toward exploration and creativity. Now the brain is more apt to notice changes in the environment, to be distracted by events, and to piece together events and knowledge that might have seemed unrelated earlier.

TABLE 2.1. Subconscious and Conscious Systems of Cognition

Subconscious	**Conscious**
Fast	Slow
Automatic	Controlled
Multiple resources	Limited resources
Controls skilled behavior	Invoked for novel situations: when learning, when in danger, when things go wrong

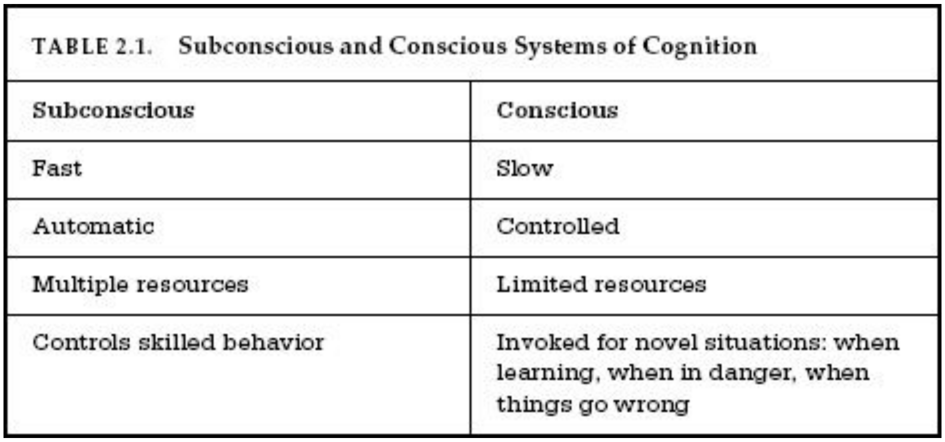


TABLE 2.1. Subconscious and Conscious Systems of Cognition

Subconscious	Conscious
Fast	Slow
Automatic	Controlled
Multiple resources	Limited resources
Controls skilled behavior	Invoked for novel situations: when learning, when in danger, when things go wrong

A positive emotional state is ideal for creative thought, but it is not very well suited for getting things done. Too much, and we call the person scatterbrained, flitting from one topic to another, unable to finish one thought before another comes to mind. A brain in a negative emotional state provides focus: precisely what is needed to maintain attention on a task and finish it. Too much, however, and we get tunnel vision, where people are unable to look beyond their narrow point of view. Both the positive, relaxed state and the anxious, negative, and tense state are valuable and powerful tools for human creativity and action. The extremes of both states, however, can be dangerous.

Human Cognition and Emotion

The mind and brain are complex entities, still the topic of considerable scientific research. One valuable explanation of the levels of processing within the brain, applicable to both cognitive and emotional processing, is to think of three different levels of processing, each quite different from the other, but all working together in concert. Although this is a gross oversimplification of the actual processing, it is a good enough approximation to provide guidance in understanding human behavior. The approach I use here comes from my book *Emotional Design*. There, I suggested that a useful approximate model of human cognition and emotion is to consider three levels of processing: visceral, behavioral, and reflective.

THE VISCERAL LEVEL

The most basic level of processing is called *visceral*. This is sometimes referred to as “the lizard brain.” All people have the same basic visceral responses. These are part of the basic protective mechanisms of the human affective system, making quick judgments about the environment: good or bad, safe or dangerous. The visceral system allows us to respond quickly and subconsciously, without conscious awareness or control. The basic biology of the visceral system minimizes its ability to learn. Visceral learning takes place primarily by sensitization or desensitization through such mechanisms as adaptation and classical conditioning. Visceral responses are fast and automatic. They give rise to the startle reflex for novel, unexpected events; for such genetically programmed behavior as fear of heights, dislike of the dark or very noisy environments, dislike of bitter tastes and the liking of sweet tastes, and so on. Note that the visceral level responds to the immediate present and produces an affective state, relatively unaffected by context or history. It simply assesses the situation: no cause is assigned, no blame, and no credit.

FIGURE 2.3. Three Levels of Processing: Visceral, Behavioral, and Reflective. Visceral and behavioral levels are subconscious and the home of basic emotions. The reflective level is where conscious thought and decision-making reside, as well as the highest level of emotions.

The visceral level is tightly coupled to the body’s musculature— the motor system. This is what causes animals to fight or flee, or to relax. An animal’s (or person’s) visceral state can often be read by analyzing the tension of the body: tense means a negative state; relaxed, a positive state. Note, too, that we often determine our own body state by noting our own musculature. A common self-report might be something like, “I was tense, my fists clenched, and I was sweating.”

Visceral responses are fast and completely subconscious. They are sensitive only to the current state of things. Most scientists do not call these emotions: they are precursors to emotion. Stand at the edge of a cliff and you will experience a visceral response. Or bask in the warm, comforting glow after a pleasant experience, perhaps a nice meal.

For designers, the visceral response is about immediate perception: the pleasantness of a mellow, harmonious sound or the jarring, irritating scratch of fingernails on a rough surface. Here is where the style matters: appearances, whether sound or sight, touch or smell, drive the visceral response. This has nothing to do with how usable, effective, or understandable the product is. It is all about attraction or repulsion. Great designers use their aesthetic sensibilities to drive these visceral responses.

Engineers and other logical people tend to dismiss the visceral response as irrelevant. Engineers are proud of the inherent quality of their work and dismayed when inferior products sell better “just because they look better.” But all of us make these kinds of judgments, even those very logical engineers. That’s why they love some of their tools and dislike others. Visceral responses matter.

THE BEHAVIORAL LEVEL

The *behavioral* level is the home of learned skills, triggered by situations that match the appropriate patterns. Actions and analyses at this level are largely subconscious. Even though we are usually aware of our actions, we are often unaware of the details. When we speak, we often do not know what we are about to say until our conscious mind (the reflective part of the mind) hears ourselves uttering the words. When we play a sport, we are prepared for action, but our responses occur far too quickly for conscious control: it is the behavioral level that takes control.

When we perform a well-learned action, all we have to do is think of the goal and the behavioral level handles all the details: the conscious mind has little or no awareness beyond creating the desire to act. It’s actually interesting to keep trying it. Move the left hand, then the right. Stick out your tongue, or open your mouth. What did you do? You don’t know. All you know is that you “willed” the action and the correct thing happened. You can even make the actions more complex. Pick up a cup, and then with the same hand, pick up several more items. You automatically adjust the fingers and the hand’s orientation to make the task possible. You only need to pay conscious attention if the cup holds some liquid that you wish to avoid spilling. But even in that case, the actual control of the muscles is beneath conscious perception: concentrate on not spilling and the hands automatically adjust.

For designers, the most critical aspect of the behavioral level is that every action is associated with an expectation. Expect a positive outcome and the result is a positive affective response (a “positive valence,” in the scientific literature). Expect a negative outcome and the result is a negative affective response (a negative valence): dread and hope, anxiety and anticipation. The information in the feedback loop of evaluation confirms or disconfirms the expectations, resulting in satisfaction or relief, disappointment or frustration.

Behavioral states are learned. They give rise to a feeling of control when there is good understanding and knowledge of results, and frustration and anger when things do not go as planned, and especially when neither the reason nor the possible remedies are known. Feedback provides reassurance, even when it indicates a negative result. A lack of feedback creates a feeling of lack of control, which can be unsettling. Feedback is critical to managing expectations, and good design provides this. Feedback—knowledge of results—is how expectations are resolved and is critical to learning and the development of skilled behavior.

Expectations play an important role in our emotional lives. This is why drivers tense when trying to get through an intersection before the light turns red, or students become highly anxious before an exam. The release of the tension of expectation creates a sense of relief. The emotional system is especially responsive to changes in states—so an upward change is interpreted positively even if it is only from a very bad state to a not-so-bad state, just as a change is interpreted negatively even if it is from an extremely positive state to one only somewhat less positive.

THE REFLECTIVE LEVEL

The *reflective* level is the home of conscious cognition. As a consequence, this is where deep understanding develops, where reasoning and conscious decision-making take place. The visceral and behavioral levels are subconscious and, as a result, they respond rapidly, but without much analysis. Reflection is cognitive, deep, and slow. It often occurs after the events have happened. It is a reflection or looking back over them, evaluating the circumstances, actions, and outcomes, often assessing blame or responsibility. The highest levels of emotions come from the reflective level, for it is here that causes are assigned and where predictions of the future take place. Adding causal elements to experienced events leads to such emotional states as guilt and pride (when we assume ourselves to be the cause) and blame and praise (when others are thought to be the cause). Most of us have probably experienced the extreme highs and lows of anticipated future events, all imagined by a runaway reflective cognitive system but intense enough to create the physiological responses associated with extreme anger or pleasure. Emotion and cognition are tightly intertwined.

DESIGN MUST TAKE PLACE AT ALL LEVELS: VISCERAL, BEHAVIORAL, AND REFLECTIVE

To the designer, reflection is perhaps the most important of the levels of processing. Reflection is conscious, and the emotions produced at this level are the most protracted: those that assign agency and cause, such as guilt and blame or praise and pride. Reflective responses are part of our memory of events. Memories last far longer than the immediate experience or the period of usage, which are the domains of the visceral and behavioral levels. It is reflection that drives us to recommend a product, to recommend that others use it—or perhaps to avoid it.

Reflective memories are often more important than reality. If we have a strongly positive visceral response but disappointing usability problems at the behavioral level, when we reflect back upon the product, the reflective level might very well weigh the positive response strongly enough to overlook the severe behavioral difficulties (hence the phrase, “Attractive things work better”). Similarly, too much frustration, especially toward the ending stage of use, and our reflections about the experience might overlook the positive visceral qualities. Advertisers hope that the strong reflective value associated with a well-known, highly prestigious brand might overwhelm our judgment, despite a frustrating experience in using the product. Vacations are often remembered with fondness, despite the evidence from diaries of repeated discomfort and anguish.

All three levels of processing work together. All play essential roles in determining a person’s like or dislike of a product or service. One nasty experience with a service provider can spoil all future experiences. One superb experience can make up for past deficiencies. The behavioral level, which is the home of interaction, is also the home of all expectation-based emotions, of hope and joy, frustration and anger. Understanding arises at a combination of the behavioral and reflective levels. Enjoyment requires all three. Designing at all three levels is so important that I devote an entire book to the topic, *Emotional Design*.

In psychology, there has been a long debate about which happens first: emotion or cognition. Do we run and flee because some event happened that made us afraid? Or are we afraid because our conscious, reflective mind notices that we are running? The three-level analysis shows that both of these ideas can be correct. Sometimes the emotion comes first. An unexpected loud noise can cause automatic visceral and behavioral responses that make us flee. Then, the reflective system observes itself fleeing and deduces that it is afraid. The actions of running and fleeing occur first and set off the interpretation of fear.

But sometimes cognition occurs first. Suppose the street where we are walking leads to a dark and narrow section. Our reflective system might conjure numerous imagined threats that await us. At some point, the imagined depiction of potential harm is large enough to trigger the behavioral system, causing us to turn, run, and flee. Here is where the cognition sets off the fear and the action.

Most products do not cause fear, running, or fleeing, but badly designed devices can induce frustration and anger, a feeling of helplessness and despair, and possibly even hate. Well-designed devices can induce pride and enjoyment, a feeling of being in control and pleasure—possibly even love and attachment. Amusement parks are experts at balancing the conflicting responses of the emotional stages, providing rides and fun houses that trigger fear responses from the visceral and behavioral levels, while all the time providing reassurance at the reflective level that the park would never subject anyone to real danger.

All three levels of processing work together to determine a person's cognitive and emotional state. High-level reflective cognition can trigger lower-level emotions. Lower-level emotions can trigger higher-level reflective cognition.

The Seven Stages of Action and the Three Levels of Processing

The stages of action can readily be associated with the three different levels of processing, as shown in Figure 2.4. At the lowest level are the visceral levels of calmness or anxiety when approaching a task or evaluating the state of the world. Then, in the middle level, are the behavioral ones driven by expectations on the execution side—for example, hope and fear—and emotions driven by the confirmation of those expectations on the evaluation side—for example, relief or despair. At the highest level are the reflective emotions, ones that assess the results in terms of the presumed causal agents and the consequences, both immediate and long-term. Here is where satisfaction and pride occur, or perhaps blame and anger.

One important emotional state is the one that accompanies complete immersion into an activity, a state that the social scientist Mihaly Csikszentmihalyi has labeled "flow." Csikszentmihalyi has long studied how people interact with their work and play, and how their lives reflect this intermix of activities. When in the flow state, people lose track of time and the outside environment. They are at one with the task they are performing. The task, moreover, is at just the proper level of difficulty: difficult enough to provide a challenge and require continued attention, but not so difficult that it invokes frustration and anxiety.

Csikszentmihalyi's work shows how the behavioral level creates a powerful set of emotional responses. Here, the subconscious expectations established by the execution side of the action cycle set up emotional states dependent upon those expectations. When the results of our actions are evaluated against expectations, the resulting emotions affect our feelings as we continue through the many cycles of action. An easy task, far below our skill level, makes it so easy to meet expectations that there is no challenge. Very little or no processing effort is required, which leads to apathy or boredom. A difficult task, far above our skill, leads to so many failed expectations that it causes frustration, anxiety, and helplessness. The flow state occurs when the challenge of the activity just slightly exceeds our skill level, so full attention is continually required. Flow requires that the activity be neither too easy nor too difficult relative to our level of skill. The constant tension coupled with continual progress and success can be an engaging, immersive experience sometimes lasting for hours.

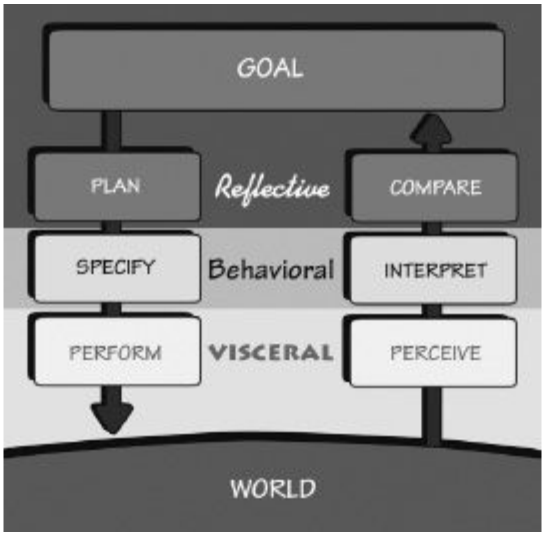


FIGURE 2.4. Levels of Processing and the Stages of the Action Cycle. Visceral response is at the lowest level: the control of simple muscles and sensing the state of the world and body. The behavioral level is about expectations, so it is sensitive to the expectations of the action sequence and then the interpretations of the feedback. The reflective level is a part of the goal- and plan-setting activity as well as affected by the comparison of expectations with what has actually happened.

People as Storytellers

Now that we have explored the way that actions get done and the three different levels of processing that integrate cognition and emotion, we are ready to look at some of the implications.

People are innately disposed to look for causes of events, to form explanations and stories. That is one reason storytelling is such a persuasive medium. Stories resonate with our experiences and provide examples of new instances. From our experiences and the stories of others we tend to form generalizations about the way people behave and things work. We attribute causes to events, and as long as these cause-and-effect pairings make sense, we accept them and use them for understanding future events. Yet these causal attributions are often erroneous. Sometimes they implicate the wrong causes, and for some things that happen, there is no single cause; rather, a complex chain of events that all contribute to the result: if any one of the events would not have occurred, the result would be different. But even when there is no single causal act, that doesn't stop people from assigning one.

Conceptual models are a form of story, resulting from our predisposition to find explanations. These models are essential in helping us understand our experiences, predict the outcome of our actions, and handle unexpected occurrences. We base our models on whatever knowledge we have, real or imaginary, naive or sophisticated.

Conceptual models are often constructed from fragmentary evidence, with only a poor understanding of what is happening, and with a kind of naive psychology that postulates causes, mechanisms, and relationships even where there are none. Some faulty models lead to the frustrations of everyday life, as in the case of my unsettable refrigerator, where my conceptual model of its operation (see again Figure 1.10A) did not correspond to reality (Figure 1.10B). Far more serious are faulty models of such complex systems as an industrial plant or passenger airplane. Misunderstanding there can lead to devastating accidents.

Consider the thermostat that controls room heating and cooling systems. How does it work? The average thermostat offers almost no evidence of its operation except in a highly roundabout manner. All we know is that if the room is too cold, we set a higher temperature into the thermostat. Eventually we feel warmer. Note that the same thing applies to the temperature control for almost any device whose temperature is to be regulated. Want to bake a cake? Set the oven thermostat and the oven goes to the desired temperature.

If you are in a cold room, in a hurry to get warm, will the room heat more quickly if you turn the thermostat to its maximum setting? Or if you want the oven to reach its working temperature faster, should you turn the temperature dial all the way to maximum, then turn it down once the desired temperature is reached? Or to cool a room most quickly, should you set the air conditioner thermostat to its lowest temperature setting?

If you think that the room or oven will cool or heat faster if the thermostat is turned all the way to the maximum setting, you are wrong—you hold an erroneous folk theory of the heating and cooling system. One commonly held folk theory of the working of a thermostat is that it is like a valve: the thermostat controls how much heat (or cold) comes out of the device. Hence, to heat or cool something most quickly, set the thermostat so that the device is on maximum. The theory is reasonable, and there exist devices that operate like this, but neither the heating or cooling equipment for a home nor the heating element of a traditional oven is one of them.

In most homes, the thermostat is just an on-off switch. Moreover, most heating and cooling devices are either fully on or fully off: all or nothing, with no in-between states. As a result, the thermostat turns the heater, oven, or air conditioner completely on, at full power, until the temperature setting on the thermostat is reached. Then it turns the unit completely off. Setting the thermostat at one extreme cannot affect how long it takes to reach the desired temperature. Worse, because this bypasses the automatic shutoff when the desired temperature is reached, setting it at the extremes invariably means that the temperature overshoots the target. If people were uncomfortably cold or hot before, they will become uncomfortable in the other direction, wasting considerable energy in the process.

But how are you to know? What information helps you understand how the thermostat works? The design problem with the refrigerator is that there are no aids to understanding, no way of forming the correct conceptual model. In fact, the information provided misleads people into forming the wrong, quite inappropriate model.

The real point of these examples is not that some people have erroneous beliefs; it is that everyone forms stories (conceptual models) to explain what they have observed. In the absence of external information, people can let their imagination run free as long as the conceptual models they develop account for the facts as they perceive them. As a result, people use their thermostats inappropriately, causing themselves unnecessary effort, and often resulting in large temperature swings, thus wasting energy, which is both a needless expense and bad for the environment. (Later in this chapter, page 69, I provide an example of a thermostat that does provide a useful conceptual model.)

Blaming the Wrong Things

People try to find causes for events. They tend to assign a causal relation whenever two things occur in succession. If some unexpected event happens in my home just after I have taken some action, I am apt to conclude that it was caused by that action, even if there really was no relationship between the two. Similarly, if I do something expecting a result and nothing happens, I am apt to interpret this lack of informative feedback as an indication that I didn't do the action correctly: the most likely thing to do, therefore, is to repeat the action, only with more force. Push a door and it fails to open? Push again, harder. With electronic devices, if the feedback is delayed sufficiently, people often are led to conclude that the press wasn't recorded, so they do the same action again, sometimes repeatedly, unaware that all of their presses were recorded. This can lead to unintended results. Repeated presses might intensify the response much more than was intended. Alternatively, a second request might cancel the previous one, so that an odd number of pushes produces the desired result, whereas an even number leads to no result.

The tendency to repeat an action when the first attempt fails can be disastrous. This has led to numerous deaths when people tried to escape a burning building by attempting to push open exit doors that opened inward, doors that should have been pulled. As a result, in many countries, the law requires doors in public places to open outward, and moreover to be operated by so-called panic bars, so that they automatically open when people, in a panic to escape a fire, push their bodies against them. This is a great application of appropriate affordances: see the door in Figure 2.5.

Modern systems try hard to provide feedback within 0.1 second of any operation, to reassure the user that the request was received. This is especially important if the operation will take considerable time. The presence of a filling hourglass or rotating clock hands is a reassuring sign that work is in progress. When the delay can be predicted, some systems provide time estimates as well as progress bars to indicate how far along the task has gone. More systems should adopt these sensible displays to provide timely and meaningful feedback of results.

FIGURE 2.5. Panic Bars on Doors. People fleeing a fire would die if they encountered exit doors that opened inward, because they would keep trying to push them outward, and when that failed, they would push harder. The proper design, now required by law in many places, is to change the design of doors so that they open when pushed. Here is one example: an excellent design strategy for dealing with real behavior by the use of the proper affordances coupled with a graceful signifier, the black bar, which indicates where to push. (Photograph by author at the Ford Design Center, Northwestern University.)

Some studies show it is wise to underpredict—that is, to say an operation will take longer than it actually will. When the system computes the amount of time, it can compute the range of possible times. In that case it ought to display the range, or if only a single value is desirable, show the slowest, longest value. That way, the expectations are liable to be exceeded, leading to a happy result.

When it is difficult to determine the cause of a difficulty, where do people put the blame? Often people will use their own conceptual models of the world to determine the perceived causal relationship between the thing being blamed and the result. The word *perceived* is critical: the causal relationship does not have to exist; the person simply has to think it is there. Sometimes the result is to attribute cause to things that had nothing to do with the action.

Suppose I try to use an everyday thing, but I can't. Who is at fault: me or the thing? We are apt to blame ourselves, especially if others are able to use it. Suppose the fault really lies in the device, so that lots of people have the same problems. Because everyone perceives the fault to be his or her own, nobody wants to admit to having trouble. This creates a conspiracy of silence, where the feelings of guilt and helplessness among people are kept hidden.

Interestingly enough, the common tendency to blame ourselves for failures with everyday objects goes against the normal attributions we make about ourselves and others. Everyone sometimes acts in a way that seems strange, bizarre, or simply wrong and inappropriate. When we do this, we tend to attribute our behavior to the environment. When we see others do it, we tend to attribute it to their personalities.

Here is a made-up example. Consider Tom, the office terror. Today, Tom got to work late, yelled at his colleagues because the office coffee machine was empty, then ran to his office and slammed the door shut. "Ah," his colleagues and staff say to one another, "there he goes again."

Now consider Tom's point of view. "I really had a hard day," Tom explains. "I woke up late because my alarm clock failed to go off: I didn't even have time for my morning coffee. Then I couldn't find a parking spot because I was late. And there wasn't any coffee in the office machine; it was all out. None of this was my fault—I had a run of really bad events. Yes, I was a bit curt, but who wouldn't be under the same circumstances?"

Tom's colleagues don't have access to his inner thoughts or to his morning's activities. All they see is that Tom yelled at them simply because the office coffee machine was empty. This reminds them of another similar event. "He does that all the time," they conclude, "always blowing up over the most minor things." Who is correct? Tom or his colleagues? The events can be seen from two different points of view with two different interpretations: common responses to the trials of life or the result of an explosive, irascible personality.

It seems natural for people to blame their own misfortunes on the environment. It seems equally natural to blame other people's misfortunes on their personalities. Just the opposite attribution, by the way, is made when things go well. When things go right, people credit their own abilities and intelligence. The onlookers do the reverse. When they see things go well for someone else, they sometimes credit the environment, or luck.

In all such cases, whether a person is inappropriately accepting blame for the inability to work simple objects or attributing behavior to environment or personality, a faulty conceptual model is at work.

LEARNED HELPLESSNESS

The phenomenon called *learned helplessness* might help explain the self-blame. It refers to the situation in which people experience repeated failure at a task. As a result, they decide that the task cannot be done, at least not by them: they are helpless. They stop trying. If this feeling covers a group of tasks, the result can be severe difficulties coping with life. In the extreme case, such learned helplessness leads to depression and to a belief that the individuals cannot cope with everyday life at all. Sometimes all it takes to get such a feeling of helplessness are a few experiences that accidentally turn out bad. The phenomenon has been most frequently studied as a precursor to the clinical problem of depression, but I have seen it happen after a few bad experiences with everyday objects.

Do common technology and mathematics phobias result from a kind of learned helplessness? Could a few instances of failure in what appear to be straightforward situations generalize to every technological object, every mathematics problem? Perhaps. In fact, the design of everyday things (and the design of mathematics courses) seems almost guaranteed to cause this. We could call this phenomenon taught helplessness.

When people have trouble using technology, especially when they perceive (usually incorrectly) that nobody else is having the same problems, they tend to blame themselves. Worse, the more they have trouble, the more helpless they may feel, believing that they must be technically or mechanically inept. This is just the opposite of the more normal situation where people blame their own difficulties on the environment. This false blame is especially ironic because the culprit here is usually the poor design of the technology, so blaming the environment (the technology) would be completely appropriate.

Consider the normal mathematics curriculum, which continues relentlessly on its way, each new lesson assuming full knowledge and understanding of all that has passed before. Even though each point may be simple, once you fall behind it is hard to catch up. The result: mathematics phobia—not because the material is difficult, but because it is taught so that difficulty in one stage hinders further progress. The problem is that once failure starts, it is soon generalized by self-blame to all of mathematics. Similar processes are at work with technology. The vicious cycle starts: if you fail at something, you think it is your fault. Therefore you think you can’t do that task. As a result, next time you have to do the task, you believe you can’t, so you don’t even try. The result is that you can’t, just as you thought.

You’re trapped in a self-fulfilling prophecy.

POSITIVE PSYCHOLOGY

Just as we learn to give up after repeated failure, we can learn optimistic, positive responses to life. For years, psychologists focused upon the gloomy story of how people failed, on the limits of human abilities, and on psychopathologies—depression, mania, paranoia, and so on. But the twenty-first century sees a new approach: to focus upon a positive psychology, a culture of positive thinking, of feeling good about oneself. In fact, the normal emotional state of most people is positive. When something doesn't work, it can be considered an interesting challenge, or perhaps just a positive learning experience.

We need to remove the word *failure* from our vocabulary, replacing it instead with *learning experience*. To fail is to learn: we learn more from our failures than from our successes. With success, sure, we are pleased, but we often have no idea why we succeeded. With failure, it is often possible to figure out why, to ensure that it will never happen again.

Scientists know this. Scientists do experiments to learn how the world works. Sometimes their experiments work as expected, but often they don't. Are these failures? No, they are learning experiences. Many of the most important scientific discoveries have come from these so-called failures.

Failure can be such a powerful learning tool that many designers take pride in their failures that happen while a product is still in development. One design firm, IDEO, has it as a creed: "Fail often, fail fast," they say, for they know that each failure teaches them a lot about what to do right. Designers need to fail, as do researchers. I have long held the belief—and encouraged it in my students and employees—that failures are an essential part of exploration and creativity. If designers and researchers do not sometimes fail, it is a sign that they are not trying hard enough—they are not thinking the great creative thoughts that will provide breakthroughs in how we do things. It is possible to avoid failure, to always be safe. But that is also the route to a dull, uninteresting life.

The designs of our products and services must also follow this philosophy. So, to the designers who are reading this, let me give some advice:

- Do not blame people when they fail to use your products properly.

- Take people's difficulties as signifiers of where the product can be improved.

- Eliminate all error messages from electronic or computer systems. Instead, provide help and guidance.

- Make it possible to correct problems directly from help and guidance messages. Allow people to continue with their task: Don't impede progress—help make it smooth and continuous. Never make people start over.

- Assume that what people have done is partially correct, so if it is inappropriate, provide the guidance that allows them to correct the problem and be on their way.

- Think positively, for yourself and for the people you interact with.

Falsely Blaming Yourself

I have studied people making errors—sometimes serious ones— with mechanical devices, light switches and fuses, computer operating systems and word processors, even airplanes and nuclear power plants. Invariably people feel guilty and either try to hide the error or blame themselves for "stupidity" or "clumsiness." I often have difficulty getting permission to watch: nobody likes to be observed performing badly. I point out that the design is faulty and that others make the same errors, yet if the task appears simple or trivial, people still blame themselves. It is almost as if they take perverse pride in thinking of themselves as mechanically incompetent.

I once was asked by a large computer company to evaluate a brand-new product. I spent a day learning to use it and trying it out on various problems. In using the keyboard to enter data, it was necessary to differentiate between the Return key and the Enter key. If the wrong key was pressed, the last few minutes' work was irrevocably lost.

I pointed out this problem to the designer, explaining that I, myself, had made the error frequently and that my analyses indicated that this was very likely to be a frequent error among users. The designer's first response was: "Why did you make that error? Didn't you read the manual?" He proceeded to explain the different functions of the two keys.

"Yes, yes," I explained, "I understand the two keys, I simply confuse them. They have similar functions, are located in similar locations on the keyboard, and as a skilled typist, I often hit Return automatically, without thought. Certainly others have had similar problems."

"Nope," said the designer. He claimed that I was the only person who had ever complained, and the company's employees had been using the system for many months. I was skeptical, so we went together to some of the employees and asked them whether they had ever hit the Return key when they should have hit Enter. And did they ever lose their work as a result?

"Oh, yes," they said, "we do that a lot."

Well, how come nobody ever said anything about it? After all, they were encouraged to report all problems with the system. The reason was simple: when the system stopped working or did something strange, they dutifully reported it as a problem. But when they made the Return versus Enter error, they blamed themselves. After all, they had been told what to do. They had simply erred.

The idea that a person is at fault when something goes wrong is deeply entrenched in society. That's why we blame others and even ourselves. Unfortunately, the idea that a person is at fault is imbedded in the legal system. When major accidents occur, official courts of inquiry are set up to assess the blame. More and more often the blame is attributed to "human error." The person involved can be fined, punished, or fired. Maybe training procedures are revised. The law rests comfortably. But in my experience, human error usually is a result of poor design: it should be called system error. Humans err continually; it is an intrinsic part of our nature. System design should take this into account. Pinning the blame on the person may be a comfortable way to proceed, but why was the system ever designed so that a single act by a single person could cause calamity? Worse, blaming the person without fixing the root, underlying cause does not fix the problem: the same error is likely to be repeated by someone else. I return to the topic of human error in Chapter 5.

Of course, people do make errors. Complex devices will always require some instruction, and someone using them without instruction should expect to make errors and to be confused. But designers should take special pains to make errors as cost-free as possible. Here is my credo about errors:

Eliminate the term *human error*. Instead, talk about communication and interaction: what we call an error is usually bad communication or interaction. When people collaborate with one another, the word error is never used to characterize another person's utterance. That's because each person is trying to understand and respond to the other, and when something is not understood or seems inappropriate, it is questioned, clarified, and the collaboration continues. Why can't the interaction between a person and a machine be thought of as collaboration?

Machines are not people. They can't communicate and understand the same way we do. This means that their designers have a special obligation to ensure that the behavior of machines is understandable to the people who interact with them. True collaboration requires each party to make some effort to accommodate and understand the other. When we collaborate with machines, it is people who must do all the accommodation. Why shouldn't the machine be more friendly? The machine should accept normal human behavior, but just as people often subconsciously assess the accuracy of things being said, machines should judge the quality of information given it, in this case to help its operators avoid grievous errors because of simple slips (discussed in Chapter 5). Today, we insist that people perform abnormally, to adapt themselves to the peculiar demands of machines, which includes always giving precise, accurate information. Humans are particularly bad at this, yet when they fail to meet the arbitrary, inhuman requirements of machines, we call it human error. No, it is design error.

Designers should strive to minimize the chance of inappropriate actions in the first place by using affordances, signifiers, good mapping, and constraints to guide the actions. If a person performs an inappropriate action, the design should maximize the chance that this can be discovered and then rectified. This requires good, intelligible feedback coupled with a simple, clear conceptual model. When people understand what has happened, what state the system is in, and what the most appropriate set of actions is, they can perform their activities more effectively.

People are not machines. Machines don't have to deal with continual interruptions. People are subjected to continual interruptions. As a result, we are often bouncing back and forth between tasks, having to recover our place, what we were doing, and what we were thinking when we return to a previous task. No wonder we sometimes forget our place when we return to the original task, either skipping or repeating a step, or imprecisely retaining the information we were about to enter.

Our strengths are in our flexibility and creativity, in coming up with solutions to novel problems. We are creative and imaginative, not mechanical and precise. Machines require precision and accuracy; people don't. And we are particularly bad at providing precise and accurate inputs. So why are we always required to do so? Why do we put the requirements of machines above those of people?

When people interact with machines, things will not always go smoothly. This is to be expected. So designers should anticipate this. It is easy to design devices that work well when everything goes as planned. The hard and necessary part of design is to make things work well even when things do not go as planned.

HOW TECHNOLOGY CAN ACCOMMODATE HUMAN BEHAVIOR

In the past, cost prevented many manufacturers from providing useful feedback that would assist people in forming accurate conceptual models. The cost of color displays large and flexible enough to provide the required information was prohibitive for small, inexpensive devices.

But as the cost of sensors and displays has dropped, it is now possible to do a lot more.

Thanks to display screens, telephones are much easier to use than ever before, so my extensive criticisms of phones found in the earlier edition of this book have been removed. I look forward to great improvements in all our devices now that the importance of these design principles are becoming recognized and the enhanced quality and lower costs of displays make it possible to implement the ideas.

PROVIDING A CONCEPTUAL MODEL FOR A HOME THERMOSTAT

My thermostat, for example (designed by Nest Labs), has a colorful display that is normally off, turning on only when it senses that I am nearby. Then it provides me with the current temperature of the room, the temperature to which it is set, and whether it is heating or cooling the room (the background color changes from black when it is neither heating nor cooling, to orange while heating, or to blue while cooling). It learns my daily patterns, so it changes temperature automatically, lowering it at bedtime, raising it again in the morning, and going into “away” mode when it detects that nobody is in the house. All the time, it explains what it is doing. Thus, when it has to change the room temperature substantially (either because someone has entered a manual change or because it has decided that it is now time to switch), it gives a prediction: “Now 75°, will be 72° in 20 minutes.” In addition, Nest can be connected wirelessly to smart devices that allow for remote operation of the thermostat and also for larger screens to provide a detailed analysis of its performance, aiding the home occupant’s development of a conceptual model both of Nest and also of the home’s energy consumption. Is Nest perfect? No, but it marks improvement in the collaborative interaction of people and everyday things.

A.

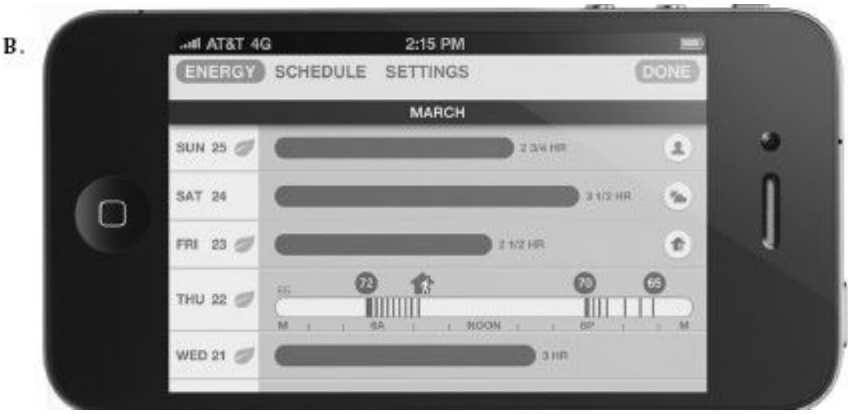


FIGURE 2.6. A Thermostat with an Explicit Conceptual Model. This thermostat, manufactured by Nest Labs, helps people form a good conceptual model of its operation. Photo A shows the thermostat. The background, blue, indicates that it is now cooling the home. The current temperature is 75°F (24°C) and the target temperature is 72°F (22°C), which it expects to reach in 20 minutes. Photo B shows its use of a smart phone to deliver a summary of its settings and the home's energy use. Both A and B combine to help the home dweller develop conceptual models of the thermostat and the home's energy consumption. (Photographs courtesy of Nest Labs, Inc.)

ENTERING DATES, TIMES, AND TELEPHONE NUMBERS

Many machines are programmed to be very fussy about the form of input they require, where the fussiness is not a requirement of the machine but due to the lack of consideration for people in the design of the software. In other words: inappropriate programming. Consider these examples.

Many of us spend hours filling out forms on computers—forms that require names, dates, addresses, telephone numbers, monetary sums, and other information in a fixed, rigid format. Worse, often we are not even told the correct format until we get it wrong. Why not figure out the variety of ways a person might fill out a form and accommodate all of them? Some companies have done excellent jobs at this, so let us celebrate their actions.

Consider Microsoft's calendar program. Here, it is possible to specify dates any way you like: "November 23, 2015," "23 Nov. 15," or

"11.23.15." It even accepts phrases such as "a week from Thursday," "tomorrow," "a week from tomorrow," or "yesterday." Same with time. You can enter the time any way you want: "3:45 PM," "15.35," "an hour," "two and one-half hours." Same with telephone numbers: Want to start with a + sign (to indicate the code for international dialing)? No problem. Like to separate the number fields with spaces, dashes, parentheses, slashes, periods? No problem. As long as the program can decipher the date, time, or telephone number into a legal format, it is accepted. I hope the team that worked on this got bonuses and promotions.

Although I single out Microsoft for being the pioneer in accepting a wide variety of formats, it is now becoming standard practice. By the time you read this, I would hope that every program would permit any intelligible format for names, dates, phone numbers, street addresses, and so on, transforming whatever is entered into whatever form the internal programming needs. But I predict that even in the twenty-second century, there will still be forms that require precise accurate (but arbitrary) formats for no reason except the laziness of the programming team. Perhaps in the years that pass between this book's publication and when you are reading this, great improvements will have been made. If we are all lucky, this section will be badly out of date. I hope so.

The Seven Stages of Action: Seven Fundamental Design Principles

The seven-stage model of the action cycle can be a valuable design tool, for it provides a basic checklist of questions to ask. In general, each stage of action requires its own special design strategies and, in turn, provides its own opportunity for disaster. Figure 2.7 summarizes the questions:

1. What do I want to accomplish?
2. What are the alternative action sequences?
3. What action can I do now?
4. How do I do it?

5. What happened?
6. What does it mean?
7. Is this okay? Have I accomplished my goal?

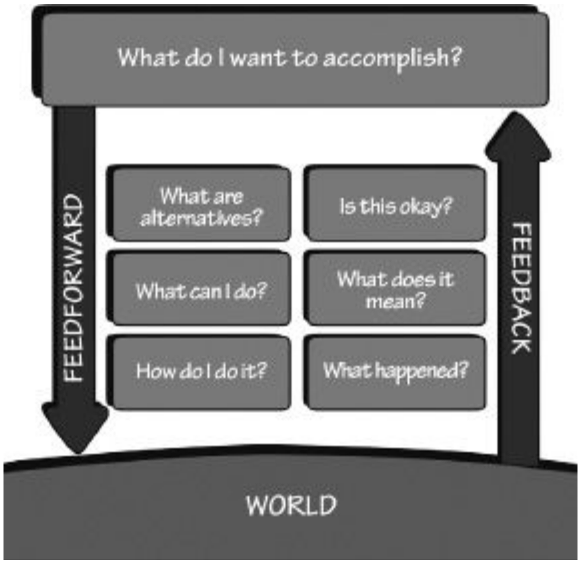


FIGURE 2.7. The Seven Stages of Action as Design Aids. Each of the seven stages indicates a place where the person using the system has a question. The seven questions pose seven design themes. How should the design convey the information required to answer the user's question? Through appropriate constraint and mappings, signifiers and conceptual models, feedback and visibility. The information that helps answer questions of execution (doing) is *feedforward*. The information that aids in understanding what has happened is *feedback*.

Anyone using a product should always be able to determine the answers to all seven questions. This puts the burden on the designer to ensure that at each stage, the product provides the information required to answer the question.

The information that helps answer questions of execution (doing) is *feedforward*. The information that aids in understanding what has happened is *feedback*. Everyone knows what feedback is. It helps you know what happened. But how do you know what you can do? That’s the role of feedforward, a term borrowed from control theory.

Feedforward is accomplished through appropriate use of signifiers, constraints, and mappings. The conceptual model plays an important role. Feedback is accomplished through explicit information about the impact of the action. Once again, the conceptual model plays an important role.

Both feedback and feedforward need to be presented in a form that is readily interpreted by the people using the system. The presentation has to match how people view the goal they are trying to achieve and their expectations. Information must match human needs.

The insights from the seven stages of action lead us to seven fundamental principles of design:

1. **Discoverability.** It is possible to determine what actions are possible and the current state of the device.

2. **Feedback.** There is full and continuous information about the results of actions and the current state of the product or service. After an action has been executed, it is easy to determine the new state.

3. **Conceptual model.** The design projects all the information needed to create a good conceptual model of the system, leading to understanding and a feeling of control. The conceptual model enhances both discoverability and evaluation of results.

4. **Affordances.** The proper affordances exist to make the desired actions possible.

5. **Signifiers.** Effective use of signifiers ensures discoverability and that the feedback is well communicated and intelligible.

6. **Mappings.** The relationship between controls and their actions follows the principles of good mapping, enhanced as much as possible through spatial layout and temporal contiguity.

7. **Constraints.** Providing physical, logical, semantic, and cultural constraints guides actions and eases interpretation.

The next time you can’t immediately figure out the shower control in a hotel room or have trouble using an unfamiliar television set or kitchen appliance, remember that the problem is in the design. Ask yourself where the problem lies. At which of the seven stages of action does it fail? Which design principles are deficient?

But it is easy to find fault: the key is to be able to do things better. Ask yourself how the difficulty came about. Realize that many different groups of people might have been involved, each of which might have had intelligent, sensible reasons for their actions. For example, a troublesome bathroom shower was designed by people who were unable to know how it would be installed, then the shower controls might have been selected by a building contractor to fit the home plans provided by yet another person. Finally, a plumber, who may not have had contact with any of the other people, did the installation. Where did the problems arise? It could have been at any one (or several) of these stages. The result may appear to be poor design, but it may actually arise from poor communication.

One of my self-imposed rules is, “Don’t criticize unless you can do better.” Try to understand how the faulty design might have occurred: try to determine how it could have been done otherwise. Thinking about the causes and possible fixes to bad design should make you better appreciate good design. So, the next time you come across a well-designed object, one that you can use smoothly and effortlessly on the first try, stop and examine it. Consider how well it masters the seven stages of action and the principles of design. Recognize that most of our interactions with products are actually interactions with a complex system: good design requires consideration of the entire system to ensure that the requirements, intentions, and desires at each stage are faithfully understood and respected at all the other stages.

CHAPTER THREE

KNOWLEDGE IN THE HEAD AND IN THE WORLD

A friend kindly let me borrow his car, an older, classic Saab. Just before I was about to leave, I found a note waiting for me: "I should have mentioned that to get the key out of the ignition, the car needs to be in reverse." The car needs to be in reverse! If I hadn't seen the note, I never could have figured that out. There was no visible cue in the car: the knowledge needed for this trick had to reside in the head. If the driver lacks that knowledge, the key stays in the ignition forever.

Every day we are confronted by numerous objects, devices, and services, each of which requires us to behave or act in some particular manner. Overall, we manage quite well. Our knowledge is often quite incomplete, ambiguous, or even wrong, but that doesn't matter: we still get through the day just fine. How do we manage? We combine knowledge in the head with knowledge in the world. Why combine? Because neither alone will suffice.

It is easy to demonstrate the faulty nature of human knowledge and memory. The psychologists Ray Nickerson and Marilyn Adams showed that people do not remember what common coins look like (Figure 3.1). Even though the example is for the American one-cent piece, the penny, the finding holds true for currencies across the world. But despite our ignorance of the coins' appearance, we use our money properly.

Why the apparent discrepancy between the precision of behavior and the imprecision of knowledge? Because not all of the knowledge required for precise behavior has to be in the head. It can be distributed

—partly in the head, partly in the world, and partly in the constraints of the world.

FIGURE 3.1. Which Is the US One-Cent Coin, the Penny? Fewer than half of the American college students who were given this set of drawings and asked to select the correct image could do so. Pretty bad performance, except that the students, of course, have no difficulty using the money. In normal life, we have to distinguish between the penny and other coins, not among several versions of one denomination. Although this is an old study using American coins, the results still hold true today using coins of any currency. (From Nickerson & Adams, 1979, *Cognitive Psychology*, 11 (3). Reproduced with permission of Academic Press via Copyright Clearance Center.)

Precise Behavior from Imprecise Knowledge

Precise behavior can emerge from imprecise knowledge for four reasons:

1. **Knowledge is both in the head and in the world.** Technically, knowledge can only be in the head, because knowledge requires interpretation and understanding, but once the world's structure has been interpreted and understood, it counts as knowledge. Much of the knowledge a person needs to do a task can be derived from the information in the world. Behavior is determined by combining the knowledge in the head with that in the world. For this chapter, I will use the term "knowledge" for both what is in the head and what is in the

world. Although technically imprecise, it simplifies the discussion and understanding.

2. **Great precision is not required.** Precision, accuracy, and completeness of knowledge are seldom required. Perfect behavior results if the combined knowledge in the head and in the world is sufficient to distinguish an appropriate choice from all others.

3. **Natural constraints exist in the world.** The world has many natural, physical constraints that restrict the possible behavior: such things as the order in which parts can go together and the ways by which an object can be moved, picked up, or otherwise manipulated. This is knowledge in the world. Each object has physical features—projections, depressions, screw threads, appendages— that limit its relationships with other objects, the operations that can be performed on it, what can be attached to it, and so on.

4. **Knowledge of cultural constraints and conventions exists in the head.** Cultural constraints and conventions are learned artificial restrictions on behavior that reduce the set of likely actions, in many cases leaving only one or two possibilities. This is knowledge in the head. Once learned, these constraints apply to a wide variety of circumstances.

Because behavior can be guided by the combination of internal and external knowledge and constraints, people can minimize the amount of material they must learn, as well as the completeness, precision, accuracy, or depth of the learning. They also can deliberately organize the environment to support behavior. This is how nonreaders can hide their inability, even in situations where their job requires reading skills. People with hearing deficits (or with normal hearing but in noisy environments) learn to use other cues. Many of us manage quite well when in novel, confusing situations where we do not know what is expected of us. How do we do this? We arrange things so that we do not need to have complete knowledge or we rely upon the knowledge of the people around us, copying their behavior or getting them to do the required actions. It is actually quite amazing how often it is possible to hide one’s ignorance, to get by without understanding or even much interest.

Although it is best when people have considerable knowledge and experience using a particular product—knowledge in the head— the designer can put sufficient cues into the design—knowledge in the world—that good performance results even in the absence of previous knowledge. Combine the two, knowledge in the head and in the world, and performance is even better. How can the designer put knowledge into the device itself?

Chapters 1 and 2 introduced a wide range of fundamental design principles derived from research on human cognition and emotion. This chapter shows how knowledge in the world combines with knowledge in the head. Knowledge in the head is knowledge in the human memory system, so this chapter contains a brief review of the critical aspects of memory necessary for the design of usable products. I emphasize that for practical purposes, we do not need to know the details of scientific theories but simpler, more general, useful approximations. Simplified models are the key to successful application. The chapter concludes with a discussion of how natural mappings present information in the world in a manner readily interpreted and usable.

KNOWLEDGE IS IN THE WORLD

Whenever knowledge needed to do a task is readily available in the world, the need for us to learn it diminishes. For example, we lack knowledge about common coins, even though we recognize them just fine (Figure 3.1). In knowing what our currency looks like, we don’t need to know all the details, simply sufficient knowledge to distinguish one value of currency from another. Only a small minority of people must know enough to distinguish counterfeit from legitimate money.

Or consider typing. Many typists have not memorized the keyboard. Usually each key is labeled, so nontypists can hunt and peck letter by letter, relying on knowledge in the world and minimizing the time required for learning. The problem is that such typing is slow and difficult. With experience, of course, hunt-and-peckers learn the positions of many of the letters on the keyboard, even without instruction, and typing speed increases notably, quickly surpassing handwriting speeds and, for some, reaching quite respectable rates. Peripheral vision and the feel of the keyboard provide some knowledge about key locations. Frequently used keys become completely learned, infrequently used keys are not learned well, and the other keys are partially learned. But as long as a typist needs to watch the keyboard, the speed is limited. The knowledge is still mostly in the world, not in the head.

If a person needs to type large amounts of material regularly, further investment is worthwhile: a course, a book, or an interactive program. The important thing is to learn the proper placement of fingers on the keyboard, to learn to type without looking, to get knowledge about the keyboard from the world into the head. It takes a few weeks to learn the system and several months of practice to become expert. But the payoff for all this effort is increased typing speed, increased accuracy, and decreased mental load and effort at the time of typing.

We only need to remember sufficient knowledge to let us get our tasks done. Because so much knowledge is available in the environment, it is surprising how little we need to learn. This is one reason people can function well in their environment and still be unable to describe what they do.

People function through their use of two kinds of knowledge: knowledge *of* and knowledge *how*. Knowledge *of*—what psychologists call *declarative knowledge*—includes the knowledge of facts and rules. “Stop at red traffic lights.” “New York City is north of Rome.” “China has twice as many people as India.” “To get the key out of the ignition of a Saab car, the gearshift must be in reverse.” Declarative knowledge is easy to write and to teach. Note that knowledge of the rules does not mean they are followed. The drivers in many cities are often quite knowledgeable about the official driving regulations, but they do not necessarily obey them. Moreover, the knowledge does not have to be true. New York City is actually south of Rome. China has only slightly more people than India (roughly 10 percent). People may know many things: that doesn’t mean they are true.

Knowledge *how*—what psychologists call *procedural knowledge*—is the knowledge that enables a person to be a skilled musician, to return a serve in tennis, or to move the tongue properly when saying the phrase “frightening witches.” Procedural knowledge is difficult or impossible to write down and difficult to teach. It is best taught by demonstration and best learned through practice. Even the best teachers cannot usually describe what they are doing. Procedural knowledge is largely subconscious, residing at the behavioral level of processing.

Knowledge in the world is usually easy to come by. Signifiers, physical constraints, and natural mappings are all perceivable cues that act as knowledge in the world. This type of knowledge occurs so commonly that we take it for granted. It is everywhere: the locations of letters on a keyboard; the lights and labels on controls that remind us of their purpose and give information about the current state of the device. Industrial equipment is replete with signal lights, indicators, and other reminders. We make extensive use of written notes. We place items in specific locations as reminders. In general, people structure their environment to provide a considerable amount of the knowledge required for something to be remembered.

Many organize their lives spatially in the world, creating a pile here, a pile there, each indicating some activity to be done, some event in progress. Probably everybody uses such a strategy to some extent. Look around you at the variety of ways people arrange their rooms and desks. Many styles of organization are possible, but invariably the physical layout and visibility of the items convey information about relative importance.

WHEN PRECISION IS UNEXPECTEDLY REQUIRED

Normally, people do not need precision in their judgments. All that is needed is the combination of knowledge in the world and in the head that makes decisions unambiguous. Everything works just fine unless the environment changes so that the combined knowledge is no longer sufficient: this can lead to havoc. At least three countries discovered this fact the hard way: the United States, when it introduced the Susan B. Anthony one-dollar coin; Great Britain, a one-pound coin (before the switch to decimal currency); and France, a ten-franc coin (before the conversion to the common European currency, the euro). The US dollar coin was confused with the existing twenty-five-cent piece (the quarter), and the British pound coin with the then five-pence piece that had the same diameter. Here is what happened in France:

> *PARIS With a good deal of fanfare, the French government released the new 10-franc coin (worth a little more than $1.50) on Oct. 22 [1986]. The public looked at it, weighed it, and began confusing it so quickly with the half-franc coin (worth only 8 cents) that a crescendo of fury and ridicule fell on both the government and the coin*.
>
> *Five weeks later, Minister of Finance Edouard Balladur suspended circulation of the coin. Within another four weeks, he canceled it altogether*.
>
> *In retrospect, the French decision seems so foolish that it is hard to fathom how it could have been made. After much study, designers came up with a silver-colored coin made of nickel and featuring a modernistic drawing by artist Joaquim Jimenez of a Gallic rooster on one side and of Marianne, the female symbol of the French republic, on the other. The coin was light, sported special ridges on its rim for easy reading by electronic vending machines and seemed tough to counterfeit*.
>
> *But the designers and bureaucrats were obviously so excited by their creation that they ignored or refused to accept the new coin's similarity to the hundreds of millions of silver-colored, nickel-based half-franc coins in circulation [whose] size and weight were perilously similar*. (Stanley Meisler. Copyright © 1986, *Los Angeles Times*. Reprinted with permission.)

The confusions probably occurred because the users of coins had already formed representations in their memories that were only sufficiently precise to distinguish among the coins that they were accustomed to using. Psychological research suggests that people maintain only partial descriptions of the things to be remembered. In the three examples of new coins introduced in the United States, Great Britain, and France, the descriptions formed to distinguish among national currency were not precise enough to distinguish between a new coin and at least one of the old coins.

Suppose I keep all my notes in a small red notebook. If this is my only notebook, I can describe it simply as “my notebook.” If I buy several more notebooks, the earlier description will no longer work. Now I must identify the first one as small or red, or maybe both small and red, whichever allows me to distinguish it from the others. But what if I acquire several small red notebooks? Now I must find some other means of describing the first book, adding to the richness of the description and to its ability to discriminate among the several similar items. Descriptions need discriminate only among the choices in front of me, but what works for one purpose may not for another.

Not all similar-looking items cause confusion. In updating this edition of the book, I searched to see whether there might be more recent examples of coin confusions. I found this interesting item on the website Wikicoins.com:

> *Someday, a leading psychologist may weigh in on one of the perplexing questions of our time: if the American public was constantly confusing the Susan B. Anthony dollar with the roughly similar-sized quarter, how come they weren’t also constantly confusing the $20 bill with the identical-sized $1 bill?* (James A. Capp, “Susan B. Anthony Dollar,” at www.wikicoins.com. Retrieved May 29, 2012)

Here is the answer. Why not any confusion? We learn to discriminate among things by looking for distinguishing features. In the United States, size is one major way of distinguishing among coins, but not among paper money. With paper money, all the bills are the same size, so Americans ignore size and look at the printed numbers and images. Hence, we often confuse similar-size American coins but only seldom confuse similar-size American bills. But people who come from a country that uses size and color of their paper money to distinguish among the amounts (for example, Great Britain or any country that uses the euro) have learned to use size and color to distinguish among paper money and therefore are invariably confused when dealing with bills from the United States.

More confirmatory evidence comes from the fact that although long-term residents of Britain complained that they confused the one-pound coin with the five-pence coin, newcomers (and children) did not have the same confusion. This is because the long-term residents were working with their original set of descriptions, which did not easily accommodate the distinctions between these two coins. Newcomers, however, started off with no preconceptions and therefore formed a set of descriptions to distinguish among all the coins; in this situation, the one-pound coin offered no particular problem. In the United States, the Susan B. Anthony dollar coin never became popular and is no longer being made, so the equivalent observations cannot be made.

What gets confused depends heavily upon history: the aspects that have allowed us to distinguish among the objects in the past. When the rules for discrimination change, people can become confused and make errors. With time, they will adjust and learn to discriminate just fine and may even forget the initial period of confusion. The problem is that in many circumstances, especially one as politically charged as the size, shape, and color of currency, the public's outrage prevents calm discussion and does not allow for any adjustment time.

Consider this as an example of design principles interacting with the messy practicality of the real world. What appears good in principle can sometimes fail when introduced to the world. Sometimes, bad products succeed and good products fail. The world is complex.

CONSTRAINTS SIMPLIFY MEMORY

Before widespread literacy, and especially before the advent of sound recording devices, performers traveled from village to village, reciting epic poems thousands of lines long. This tradition still exists in some societies. How do people memorize such voluminous amounts of material? Do some people have huge amounts of knowledge in their heads? Not really. It turns out that external constraints exert control over the permissible choice of words, thus dramatically reducing the memory load. One of the secrets comes from the powerful constraints of poetry.

Consider the constraints of rhyming. If you wish to rhyme one word with another, there are usually a lot of alternatives. But if you must have a word with a particular meaning to rhyme with another, the joint constraints of meaning and rhyme can cause a dramatic reduction in the number of possible candidates, sometimes reducing a large set to a single choice. Sometimes there are no candidates at all. This is why it is much easier to memorize poetry than to create poems. Poems come in many different forms, but all have formal restrictions on their construction. The ballads and tales told by the traveling storytellers used multiple poetic constraints, including rhyme, rhythm, meter, assonance, alliteration, and onomatopoeia, while also remaining consistent with the story being told.

Consider these two examples:

One. I am thinking of three words: one means "a mythical being," the second is "the name of a building material," and the third is "a unit of time." What words do I have in mind?

Two. This time look for rhyming words. I am thinking of three words: one rhymes with "post," the second with "eel," and the third with "ear." What words am I thinking of? (From Rubin & Wallace, 1989.)

In both examples, even though you might have found answers, they were not likely to be the same three that I had in mind. There simply are not enough constraints. But suppose I now tell you that the words I seek are the same in both tasks: What is a word that means a mythical being and rhymes with "post"? What word is the name of a building material and rhymes with "eel"? And what word is a unit of time and rhymes with "ear"? Now the task is easy: the joint specification of the words completely constrains the selection. When the psychologists David Rubin and Wanda Wallace studied these examples in their laboratory, people almost never got the correct meanings or rhymes for the first two tasks, but most people correctly answered, "ghost," "steel," and "year" in the combined task.

The classic study of memory for epic poetry was done by Albert Bates Lord. In the mid-1900s he traveled throughout the former Yugoslavia (now a number of separate, independent countries) and found people who still followed the oral tradition. He demonstrated that the "singer of tales," the person who learns epic poems and goes from village to village reciting them, is really re-creating them, composing poetry on the fly in such a way that it obeys the rhythm, theme, story line, structure, and other characteristics of the poem. This is a prodigious feat, but it is not an example of rote memory.

The power of multiple constraints allows one singer to listen to another singer tell a lengthy tale once, and then after a delay of a few hours or a day, to recite “the same song, word for word, and line for line.” In fact, as Lord points out, the original and new recitations are not the same word for word, but both teller and listener perceive them as the same, even when the second version was twice as long as the first. They are the same in the ways that matter to the listener: they tell the same story, express the same ideas, and follow the same rhyme and meter. They are the same in all senses that matter to the culture. Lord shows just how the combination of memory for poetics, theme, and style combines with cultural structures into what he calls a “formula” for producing a poem perceived as identical to earlier recitations.

The notion that someone should be able to recite word for word is relatively modern. Such a notion can be held only after printed texts become available; otherwise who could judge the accuracy of a recitation? Perhaps more important, who would care?

All this is not to detract from the feat. Learning and reciting an epic poem, such as Homer’s *Odyssey* and *Iliad*, is clearly difficult even if the singer is re-creating it: there are twenty-seven thousand lines of verse in the combined written version. Lord points out that this length is excessive, probably produced only during the special circumstances in which Homer (or some other singer) dictated the story slowly and repetitively to the person who first wrote it down. Normally the length would be varied to accommodate the whims of the audience, and no normal audience could sit through twenty-seven thousand lines. But even at one-third the size, nine thousand lines, being able to recite the poem is impressive: at one second per line, the verses would take two and one-half hours to recite. It is impressive even allowing for the fact that the poem is re-created as opposed to memorized, because neither the singer nor the audience expect word-for-word accuracy (nor would either have any way of verifying that).

Most of us do not learn epic poems. But we do make use of strong constraints that serve to simplify what must be retained in memory. Consider an example from a completely different domain: taking apart and reassembling a mechanical device. Typical items in the home that an adventuresome person might attempt to repair include a door lock, toaster, and washing machine. The device is apt to have tens of parts. What has to be remembered to be able to put the parts together again in a proper order? Not as much as might appear from an initial analysis. In the extreme case, if there are ten parts, there are 10! (ten factorial) different ways in which to reassemble them—a little over 3.5 million alternatives.

But few of these possibilities are possible: there are numerous physical constraints on the ordering. Some pieces must be assembled before it is even possible to assemble the others. Some pieces are physically constrained from fitting into the spots reserved for others: bolts must fit into holes of an appropriate diameter and depth; nuts and washers must be paired with bolts and screws of appropriate sizes; and washers must always be put on before nuts. There are even cultural constraints: we turn screws clockwise to tighten, counterclockwise to loosen; the heads of screws tend to go on the visible part (front or top) of a piece, bolts on the less visible part (bottom, side, or interior); wood screws and machine screws look different and are inserted into different kinds of materials. In the end, the apparently large number of decisions is reduced to only a few choices that should have been learned or otherwise noted during the disassembly. The constraints by themselves are often not sufficient to determine the proper reassembly of the device—mistakes do get made—but the constraints reduce the amount that must be learned to a reasonable quantity. Constraints are powerful tools for the designer: they are examined in detail in Chapter 4.

Memory Is Knowledge in the Head

An old Arabic folk tale, "'Ali Baba and the Forty Thieves," tells how the poor woodcutter 'Ali Baba discovered the secret cave of a band of thieves. 'Ali Baba overheard the thieves entering the cave and learned the secret phrase that opened the cave: "Open Simsim." (*Simsim* means "sesame" in Persian, so many versions of the story translate the phrase as “Open Sesame.”) ‘Ali Baba’s brother-in-law, Kasim, forced him to reveal the secret. Kasim then went to the cave.

> *When he reached the entrance of the cavern, he pronounced the words, Open Simsim!*
>
> *The door immediately opened, and when he was in, closed on him. In examining the cave he was greatly astonished to find much more riches than he had expected from ‘Ali Baba’s relation*.
>
> *He quickly laid at the door of the cavern as many bags of gold as his ten mules could carry, but his thoughts were now so full of the great riches he should possess, that he could not think of the necessary words to make the door open. Instead of Open Simsim! he said Open Barley! and was much amazed to find that the door remained shut. He named several sorts of grain, but still the door would not open*.
>
> *Kasim never expected such an incident, and was so alarmed at the danger he was in that the more he endeavoured to remember the word Simsim the more his memory was confounded, and he had as much forgotten it as if he had never heard it mentioned*.
>
> *Kasim never got out. The thieves returned, cut off Kasim’s head, and quartered his body*. (From Colum’s 1953 edition of *The Arabian Nights.)*

Most of us will not get our head cut off if we fail to remember a secret code, but it can still be very hard to recall the code. It is one thing to have to memorize one or two secrets: a combination, or a password, or the secret to opening a door. But when the number of secret codes gets too large, memory fails. There seems to be a conspiracy, one calculated to destroy our sanity by overloading our memory. Many codes, such as postal codes and telephone numbers, exist primarily to make life easier for machines and their designers without any consideration of the burden placed upon people. Fortunately, technology has now permitted most of us to avoid having to remember this arbitrary knowledge but to let our technology do it for us: phone numbers, addresses and postal codes, Internet and e-mail addresses are all retrievable automatically, so we no longer have to learn them. Security codes, however, are a different matter, and in the never-ending, escalating battle between the white hats and the black, the good guys and the bad, the number of different arbitrary codes we must remember or special security devices we must carry with us continues to escalate in both number and complexity.

Many of these codes must be kept secret. There is no way that we can learn all those numbers or phrases. Quick: what magical command was Kasim trying to remember to open the cavern door?

How do most people cope? They use simple passwords. Studies show that five of the most common passwords are: “password,” “123456,” “12345678,” “qwerty,” and “abc123.” All of these are clearly selected for easy remembering and typing. All are therefore easy for a thief or mischief-maker to try. Most people (including me) have a small number of passwords that they use on as many different sites as possible. Even security professionals admit to this, thereby hypocritically violating their own rules.

Many of the security requirements are unnecessary, and needlessly complex. So why are they required? There are many reasons. One is that there are real problems: criminals impersonate identities to steal people’s money and possessions. People invade others’ privacy, for nefarious or even harmless purposes. Professors and teachers need to safeguard examination questions and grades. For companies and nations, it is important to maintain secrets. There are lots of reasons to keep things behind locked doors or password-protected walls. The problem, however, is the lack of proper understanding of human abilities.

We do need protection, but most of the people who enforce the security requirements at schools, businesses, and government are technologists or possibly law-enforcement officials. They understand crime, but not human behavior. They believe that “strong” passwords, ones difficult to guess, are required, and that they must be changed frequently. They do not seem to recognize that we now need so many passwords—even easy ones—that it is difficult to remember which goes with which requirement. This creates a new layer of vulnerability.

The more complex the password requirements, the less secure the system. Why? Because people, unable to remember all these combinations, write them down. And then where do they store this private, valuable knowledge? In their wallet, or taped under the computer keyboard, or wherever it is easy to find, because it is so frequently needed. So a thief only has to steal the wallet or find the list and then all secrets are known. Most people are honest, concerned workers. And it is these individuals that complex security systems impede the most, preventing them from getting their work done. As a result, it is often the most dedicated employee who violates the security rules and weakens the overall system.

When I was doing the research for this chapter, I found numerous examples of secure passwords that force people to use insecure memory devices for them. One post on the “Mail Online” forum of the British *Daily Mail* newspaper described the technique:

> *When I used to work for the local government organisation we HAD TO change our Passwords every three months. To ensure I could remember it, I used to write it on a Post-It note and stick it above my desk*.

How can we remember all these secret things? Most of us can’t, even with the use of mnemonics to make some sense of nonsensical material. Books and courses on improving memory can work, but the methods are laborious to learn and need continual practice to maintain. So we put the memory in the world, writing things down in books, on scraps of paper, even on the backs of our hands. But we disguise them to thwart would-be thieves. That creates another problem: How do we disguise the items, how do we hide them, and how do we remember what the disguise was or where we put it? Ah, the foibles of memory.

Where should you hide something so that nobody else will find it? In unlikely places, right? Money is hidden in the freezer; jewelry in the medicine cabinet or in shoes in the closet. The key to the front door is hidden under the mat or just below the window ledge. The car key is under the bumper. The love letters are in a flower vase. The problem is, there aren’t that many unlikely places in the home. You may not remember where the love letters or keys are hidden, but your burglar will. Two psychologists who examined the issue described the problem this way:

> *There is often a logic involved in the choice of unlikely places. For example, a friend of ours was required by her insurance company to acquire a safe if she wished to insure her valuable gems. Recognizing that she might forget the combination to the safe, she thought carefully about where to keep the combination. Her solution was to write it in her personal phone directory under the letter S next to "Mr. and Mrs. Safe," as if it were a telephone number. There is a clear logic here: Store numerical information with other numerical information. She was appalled, however, when she heard a reformed burglar on a daytime television talk show say that upon encountering a safe, he always headed for the phone directory because many people keep the combination there*. (From Winograd & Soloway, 1986, "On Forgetting the Locations of Things Stored in Special Places." Reprinted with permission.)

All the arbitrary things we need to remember add up to unwitting tyranny. It is time for a revolt. But before we revolt, it is important to know the solution. As noted earlier, one of my self-imposed rules is, "Never criticize unless you have a better alternative." In this case, it is not clear what the better system might be.

Some things can only be solved by massive cultural changes, which probably means they will never be solved. For example, take the problem of identifying people by their names. People's names evolved over many thousands of years, originally simply to distinguish people within families and groups who lived together. The use of multiple names (given names and surnames) is relatively recent, and even those do not distinguish one person from all the seven billion in the world. Do we write the given name first, or the surname? It depends upon what country you are in. How many names does a person have? How many characters in a name? What characters are legitimate? For example, can a name include a digit? (I know people who have tried to use such names as "h3nry." I know of a company named "Autonom3.")

How does a name translate from one alphabet to another? Some of my Korean friends have given names that are identical when written in the

Korean alphabet, Hangul, but that are different when transliterated into English.

Many people change their names when they get married or divorced, and in some cultures, when they pass significant life events. A quick search on the Internet reveals multiple questions from people in Asia who are confused about how to fill out American or European passport forms because their names don’t correspond to the requirements.

And what happens when a thief steals a person’s identity, masquerading as the other individual, using his or her money and credit? In the United States, these identity thieves can also apply for income tax rebates and get them, and when the legitimate taxpayers try to get their legitimate refund, they are told they already received it.

I once attended a meeting of security experts that was held at the corporate campus of Google. Google, like most corporations, is very protective of its processes and advanced research projects, so most of the buildings were locked and guarded. Attendees of the security meeting were not allowed access (except those who worked at Google, of course). Our meetings were held in a conference room in the public space of an otherwise secure building. But the toilets were all located inside a secure area. How did we manage? These world-famous, leading authorities on security figured out a solution: They found a brick and used it to prop open the door leading into the secure area. So much for security: Make something too secure, and it becomes less secure.

How do we solve these problems? How do we guarantee people’s access to their own records, bank accounts, and computer systems? Almost any scheme you can imagine has already been proposed, studied, and found to have defects. Biometric markers (iris or retina patterns, fingerprints, voice recognition, body type, DNA)? All can be forged or the systems’ databases manipulated. Once someone manages to fool the system, what recourse is there? It isn’t possible to change biometric markers, so once they point to the wrong person, changes are extremely difficult to make.

The strength of a password is actually pretty irrelevant because most passwords are obtained through “key loggers” or are stolen. A key logger is software hidden within your computer system that records what you type and sends it to the bad guys. When computer systems are broken into, millions of passwords might get stolen, and even if they are encrypted, the bad guys can often decrypt them. In both these cases, however secure the password, the bad guys know what it is.

The safest methods require multiple identifiers, the most common schemes requiring at least two different kinds: “something you have” plus “something you know.” The “something you have” is often a physical identifier, such as a card or key, perhaps even something implanted under the skin or a biometric identifier, such as fingerprints or patterns of the eye’s iris. The “something you know” would be knowledge in the head, most likely something memorized. The memorized item doesn’t have to be as secure as today’s passwords because it wouldn’t work without the “something you have.” Some systems allow for a second, alerting password, so that if the bad guys try to force someone to enter a password into a system, the individual would use the alerting one, which would warn the authorities of an illegal entry.

Security poses major design issues, ones that involve complex technology as well as human behavior. There are deep, fundamental difficulties. Is there a solution? No, not yet. We will probably be stuck with these complexities for a long time.

The Structure of Memory

Say aloud the numbers 1, 7, 4, 2, 8. Next, without looking back, repeat them. Try again if you must, perhaps closing your eyes, the better to “hear” the sound still echoing in mental activity. Have someone read a random sentence to you. What were the words? The memory of the just present is available immediately, clear and complete, without mental effort.

What did you eat for dinner three days ago? Now the feeling is different. It takes time to recover the answer, which is neither as clear nor as complete a remembrance as that of the just present, and *the recovery is likely to require considerable mental effort. Retrieval of the past differs from retrieval of the just present. More effort is required, less clarity results. Indeed, the "past" need not be so long ago. Without looking back, what were those digits? For some people, this retrieval now takes time and effort*. (From *Learning and Memory*, Norman, 1982.)

Psychologists distinguish between two major classes of memory: short-term or working memory, and long-term memory. The two are quite different, with different implications for design.

SHORT-TERM OR WORKING MEMORY

Short-term or working memory (STM) retains the most recent experiences or material that is currently being thought about. It is the memory of the just present. Information is retained automatically and retrieved without effort; but the amount of information that can be retained this way is severely limited. Something like five to seven items is the limit of STM, with the number going to ten or twelve if the material is continually repeated, what psychologists call "rehearsing."

Multiply 27 times 293 in your head. If you try to do it the same way you would with paper and pencil, you will almost definitely be unable to hold all the digits and intervening answers within STM. You will fail. The traditional method of multiplying is optimized for paper and pencil. There is no need to minimize the burden on working memory because the numbers written on the paper serve this function (knowledge in the world), so the burden on STM, on knowledge in the head, is quite limited. There are ways of doing mental multiplication, but the methods are quite different from those using paper and pencil and require considerable training and practice.

Short-term memory is invaluable in the performance of everyday tasks, in letting us remember words, names, phrases, and parts of tasks: hence its alternative name, working memory. But the material being maintained in STM is quite fragile. Get distracted by some other activity and, poof, the stuff in STM disappears. It is capable of holding a postal code or telephone number from the time you look it up until the time it is used—as long as no distractions occur. Nine- or ten-digit numbers give trouble, and when the number starts to exceed that—don’t bother. Write it down. Or divide the number into several shorter segments, transforming the long number into meaningful chunks.

Memory experts use special techniques, called *mnemonics*, to remember amazingly large amounts of material, often after only a single exposure. One method is to transform the digits into meaningful segments (one famous study showed how an athlete thought of digit sequences as running times, and after refining the method over a long period, could learn incredibly long sequences at one glance). One traditional method used to encode long sequences of digits is to first transform each digit into a consonant, then transform the consonant sequence into a memorable phrase. A standard table of conversions of digits to consonants has been around for hundreds of years, cleverly designed to be easy to learn because the consonants can be derived from the shape of the digits. Thus, “1” is translated into “t” (or the similar-sounding “d”), “2” becomes “n,” “3” becomes “m,” “4” is “r,” and “5” becomes “L” (as in the Roman numeral for 50). The full table and the mnemonics for learning the pairings are readily found on the Internet by searching for “number-consonant mnemonic.”

Using the number-consonant transformation, the string 4194780135092770 translates into the letters *rtbrkfstmlspncks*, which in turn may become, “A hearty breakfast meal has pancakes.” Most people are not experts at retaining long arbitrary strings of anything, so although it is interesting to observe memory wizards, it would be wrong to design systems that assumed this level of proficiency.

The capacity of STM is surprisingly difficult to measure, because how much can be retained depends upon the familiarity of the material. Retention, moreover, seems to be of meaningful items, rather than of some simpler measure such as seconds or individual sounds or letters. Retention is affected by both time and the number of items. The number of items is more important than time, with each new item decreasing the likelihood of remembering all of the preceding items. The capacity is items because people can remember roughly the same number of digits and words, and almost the same number of simple three- to five-word phrases. How can this be? I suspect that STM holds something akin to a pointer to an already encoded item in long-term memory, which means the memory capacity is the number of pointers it can keep. This would account for the fact that the length or complexity of the item has little impact—simply the number of items. It doesn't neatly account for the fact that we make acoustical errors in STM, unless the pointers are held in a kind of acoustical memory. This remains an open topic for scientific exploration.

The traditional measures of STM capacity range from five to seven, but from a practical point of view, it is best to think of it as holding only three to five items. Does that seem too small a number? Well, when you meet a new person, do you always remember his or her name? When you have to dial a phone number, do you have to look at it several times while entering the digits? Even minor distractions can wipe out the stuff we are trying to hold on to in STM.

What are the design implications? Don't count on much being retained in STM. Computer systems often enhance people's frustration when things go wrong by presenting critical information in a message that then disappears from the display just when the person wishes to make use of the information. So how can people remember the critical information? I am not surprised when people hit, kick, or otherwise attack their computers.

I have seen nurses write down critical medical information about their patients on their hands because the critical information would disappear if the nurse was distracted for a moment by someone asking a question. The electronic medical records systems automatically log out users when the system does not appear to be in use. Why the automatic logouts? To protect patient privacy. The cause may be well motivated, but the action poses severe challenges to nurses who are continually being interrupted in their work by physicians, co-workers, or patient requests. While they are attending to the interruption, the system logs them out, so they have to start over again. No wonder these nurses wrote down the knowledge, although this then negated much of the value of the computer system in minimizing handwriting errors. But what else were they to do? How else to get at the critical information? They couldn't remember it all: that's why they had computers.

The limits on our short-term memory systems caused by interfering tasks can be mitigated by several techniques. One is through the use of multiple sensory modalities. Visual information does not much interfere with auditory, actions do not interfere much with either auditory or written material. Haptics (touch) is also minimally interfering. To maximize efficiency of working memory it is best to present different information over different modalities: sight, sound, touch (haptics), hearing, spatial location, and gestures. Automobiles should use auditory presentation of driving instructions and haptic vibration of the appropriate side of the driver's seat or steering wheel to warn when drivers leave their lanes, or when there are other vehicles to the left or right, so as not to interfere with the visual processing of driving information. Driving is primarily visual, so the use of auditory and haptic modalities minimizes interference with the visual task.

LONG-TERM MEMORY

Long-term memory (LTM) is memory for the past. As a rule, it takes time for information to get into LTM and time and effort to get it out again. Sleep seems to play an important role in strengthening the memories of each day's experiences. Note that we do not remember our experiences as an exact recording; rather, as bits and pieces that are reconstructed and interpreted each time we recover the memories, which means they are subject to all the distortions and changes that the human explanatory mechanism imposes upon life. How well we can ever recover experiences and knowledge from LTM is highly dependent upon how the material was interpreted in the first place. What is stored in LTM under one interpretation probably cannot be found later on when sought under some other interpretation. As for how large the memory is, nobody really knows: giga- or tera-items. We don't even know what kinds of units should be used. Whatever the size, it is so large as not to impose any practical limit.

The role of sleep in the strengthening of LTM is still not well understood, but there are numerous papers investigating the topic. One possible mechanism is that of rehearsal. It has long been known that rehearsal of material—mentally reviewing it while still active in working memory (STM)—is an important component of the formation of long-term memory traces. "Whatever makes you rehearse during sleep is going to determine what you remember later, and conversely, what you're going to forget," said Professor Ken Paller of Northwestern University, one of the authors of a recent study on the topic (Oudiette, Antony, Creery, and Paller, 2013). But although rehearsal in sleep strengthens memories, it might also falsify them: "Memories in our brain are changing all of the time. Sometimes you improve memory storage by rehearsing all the details, so maybe later you remember better—or maybe worse if you've embellished too much."

Remember how you answered this question from Chapter 2?

In the house you lived in three houses ago, as you entered the front door, was the doorknob on the left or right?

For most people, the question requires considerable effort just to recall which house is involved, plus one of the special techniques described in Chapter 2 for putting yourself back at the scene and reconstructing the answer. This is an example of procedural memory, a memory for how we do things, as opposed to declarative memory, the memory for factual information. In both cases, it can take considerable time and effort to get to the answer. Moreover, the answer is not directly retrieved in a manner analogous to the way we read answers from books or websites. The answer is a reconstruction of the knowledge, so it is subject to biases and distortions. Knowledge in memory is meaningful, and at the time of retrieval, a person might subject it to a different meaningful interpretation than is wholly accurate.

A major difficulty with LTM is in organization. How do we find the things we are trying to remember? Most people have had the "tip of the tongue" experience when trying to remember a name or word: there is a feeling of knowing, but the knowledge is not consciously available. Sometime later, when engaged in some other, different activity, the name may suddenly pop into the conscious mind. The way by which people retrieve the needed knowledge is still unknown, but probably involves some form of pattern-matching mechanism coupled with a confirmatory process that checks for consistency with the required knowledge. This is why when you search for a name but continually retrieve the wrong name, you know it is wrong. Because this false retrieval impedes the correct retrieval, you have to turn to some other activity to allow the subconscious memory retrieval process to reset itself.

Because retrieval is a reconstructive process, it can be erroneous. We may reconstruct events the way we would prefer to remember them, rather than the way we experienced them. It is relatively easy to bias people so that they form false memories, “remembering” events in their lives with great clarity, even though they never occurred. This is one reason that eyewitness testimony in courts of law is so problematic: eyewitnesses are notoriously unreliable. A huge number of psychological experiments show how easy it is to implant false memories into people’s minds so convincingly that people refuse to admit that the memory is of an event that never happened.

Knowledge in the head is actually knowledge in memory: internal knowledge. If we examine how people use their memories and how they retrieve knowledge, we discover a number of categories. Two are important for us now:

1. **Memory for arbitrary things.** The items to be retained seem arbitrary, with no meaning and no particular relationship to one another or to things already known.

2. **Memory for meaningful things.** The items to be retained form meaningful relationships with themselves or with other things already known.

MEMORY FOR ARBITRARY AND MEANINGFUL THINGS

Arbitrary knowledge can be classified as the simple remembering of things that have no underlying meaning or structure. A good example is the memory of the letters of the alphabet and their ordering, the names of people, and foreign vocabulary, where there appears to be no obvious structure to the material. This also applies to the learning of the arbitrary key sequences, commands, gestures, and procedures of much of our modern technology: This is rote learning, the bane of modern existence.

Some things do require rote learning: the letters of the alphabet, for example, but even here we add structure to the otherwise meaningless list of words, turning the alphabet into a song, using the natural constraints of rhyme and rhythm to create some structure.

Rote learning creates problems. First, because what is being learned is arbitrary, the learning is difficult: it can take considerable time and effort. Second, when a problem arises, the memorized sequence of actions gives no hint of what has gone wrong, no suggestion of what might be done to fix the problem. Although some things are appropriate to learn by rote, most are not. Alas, it is still the dominant method of instruction in many school systems, and even for much adult training. This is how some people are taught to use computers, or to cook. It is how we have to learn to use some of the new (poorly designed) gadgets of our technology.

We learn arbitrary associations or sequences by artificially providing structure. Most books and courses on methods for improving memory (mnemonics) use a variety of standard methods for providing structure, even for things that might appear completely arbitrary, such as grocery lists, or matching the names of people to their appearance. As we saw in the discussion of these methods for STM, even strings of digits can be remembered if they can be associated with meaningful structures. People who have not received this training or who have not invented some methods themselves often try to manufacture some artificial structure, but these are often rather unsatisfactory, which is why the learning is so bad.

Most things in the world have a sensible structure, which tremendously simplifies the memory task. When things make sense, they correspond to knowledge that we already have, so the new material can be understood, interpreted, and integrated with previously acquired material. Now we can use rules and constraints to help understand what things go together. Meaningful structure can organize apparent chaos and arbitrariness.

Remember the discussion of conceptual models in Chapter 1? Part of the power of a good conceptual model lies in its ability to provide meaning to things. Let’s look at an example to show how a meaningful interpretation transforms an apparently arbitrary task into a natural one. Note that the appropriate interpretation may not at first be obvious; it, too, is knowledge and has to be discovered.

A Japanese colleague, Professor Yutaka Sayeki of the University of Tokyo, had difficulty remembering how to use the turn signal switch on his motorcycle's left handlebar. Moving the switch forward signaled a right turn; backward, a left turn. The meaning of the switch was clear and unambiguous, but the direction in which it should be moved was not. Sayeki kept thinking that because the switch was on the left handlebar, pushing it forward should signal a left turn. That is, he was trying to map the action "push the left switch forward" to the intention "turn left," which was wrong. As a result, he had trouble remembering which switch direction should be used for which turning direction. Most motorcycles have the turn-signal switch mounted differently, rotated 90 degrees, so that moving it left signals a left turn; moving it right, a right turn. This mapping is easy to learn (it is an example of a natural mapping, discussed at the end of this chapter). But the turn switch on Sayeki's motorcycle moved forward and back, not left and right. How could he learn it?

Sayeki solved the problem by reinterpreting the action. Consider the way the handlebars of the motorcycle turn. For a left turn, the left handlebar moves backward. For a right turn, the left handlebar moves forward. The required switch movements exactly paralleled the handlebar movements. If the task is conceptualized as signaling the direction of motion of the handlebars rather than the direction of the motorcycle, the switch motion can be seen to mimic the desired motion; finally we have a natural mapping.

When the motion of the switch seemed arbitrary, it was difficult to remember. Once Professor Sayeki had invented a meaningful relationship, he found it easy to remember the proper switch operation. (Experienced riders will point out that this conceptual model is wrong: to turn a bike, one first steers in the opposite direction of the turn. This is discussed as Example 3 in the next section, "Approximate Models.")

The design implications are clear: provide meaningful structures. Perhaps a better way is to make memory unnecessary: put the required information in the world. This is the power of the traditional graphical user interface with its old-fashioned menu structure. When in doubt, one could always examine all the menu items until the desired one was found. Even systems that do not use menus need to provide some structure: appropriate constraints and forcing functions, natural good mapping, and all the tools of feedforward and feedback. The most effective way of helping people remember is to make it unnecessary.

Approximate Models: Memory in the Real World

Conscious thinking takes time and mental resources. Well-learned skills bypass the need for conscious oversight and control: conscious control is only required for initial learning and for dealing with unexpected situations. Continual practice automates the action cycle, minimizing the amount of conscious thinking and problem-solving required to act. Most expert, skilled behavior works this way, whether it is playing tennis or a musical instrument, or doing mathematics and science. Experts minimize the need for conscious reasoning. Philosopher and mathematician Alfred North Whitehead stated this principle over a century ago:

> *It is a profoundly erroneous truism, repeated by all copy-books and by eminent people when they are making speeches, that we should cultivate the habit of thinking of what we are doing. The precise opposite is the case. Civilization advances by extending the number of important operations which we can perform without thinking about them*. (Alfred North Whitehead, 1911.)

One way to simplify thought is to use simplified models, approximations to the true underlying state of affairs. Science deals in truth, practice deals with approximations. Practitioners don’t need truth: they need results relatively quickly that, although inaccurate, are “good enough” for the purpose to which they will be applied. Consider these examples:

EXAMPLE 1: CONVERTING TEMPERATURES BETWEEN FAHRENHEIT AND CELSIUS

It is now 55 °F outside my home in California. What temperature is it in Celsius? Quick, do it in your head without using any technology: What is the answer?

I am sure all of you remember the conversion equation:

$$^\circ C = (^\circ F - 32) \times 5 / 9$$

Plug in 55 for °F, and $^\circ C = (55-32) \times 5 / 9 = 12.8^\circ$. But most people can’t do this without pencil and paper because there are too many intermediate numbers to maintain in STM.

Want a simpler way? Try this approximation—you can do it in your head, there is no need for paper or pencil:

$$^\circ C = (^\circ F - 30) / 2$$

Plug in 55 for °F, and $^\circ C = (55-30) / 2 = 12.5^\circ$. Is the equation an exact conversion? No, but the approximate answer of 12.5 is close enough to the correct value of 12.8. After all, I simply wanted to know whether I should wear a sweater. Anything within 5ºF of the real value would work for this purpose.

Approximate answers are often good enough, even if technically wrong. This simple approximation method for temperature conversion is “good enough” for temperatures in the normal range of interior and outside temperatures: it is within 3ºF (or 1.7ºC) in the range of −5º to 25ºC (20º to 80ºF). It gets further off at lower or higher temperatures, but for everyday use, it is wonderful. Approximations are good enough for practical use.

EXAMPLE 2: A MODEL OF SHORT-TERM MEMORY

Here is an approximate model for STM:

There are five memory slots in short-term memory. Each time a new item is added, it occupies a slot, knocking out whatever was there beforehand.

Is this model true? No, not a single memory researcher in the entire world believes this to be an accurate model of STM. But it is good enough for applications. Make use of this model, and your designs will be more usable.

EXAMPLE 3: STEERING A MOTORCYCLE

In the preceding section, we learned how Professor Sayeki mapped the turning directions of his motorcycle to his turn signals, enabling him to remember their correct usage. But there, I also pointed out that the conceptual model was wrong.

Why is the conceptual model for steering a motorcycle useful even though it is wrong? Steering a motorcycle is counterintuitive: to turn to the left, the handlebars must first be turned to the right. This is called countersteering, and it violates most people's conceptual models. Why is this true? Shouldn't we rotate the handlebars left to turn the bike left? The most important component of turning a two-wheeled vehicle is lean: when the bike is turning left, the rider is leaning to the left. Countersteering causes the rider to lean properly: when the handlebars are turned to the right, the resulting forces upon the rider cause the body to lean left. This weight shift then causes the bike to turn left.

Experienced riders often do the correct operations subconsciously, unaware that they start a turn by rotating the handlebars opposite from the intended direction, thus violating their own conceptual models. Motorcycle training courses have to conduct special exercises to convince riders that this is what they are doing.

You can test this counterintuitive concept on a bicycle or motorcycle by getting up to a comfortable speed, placing the palm of the hand on the end of the left handlebar, and gently pushing it forward. The handlebars and front wheel will turn to the right and the body will lean to the left, resulting in the bike—and the handlebars— turning to the left.

Professor Sayeki was fully aware of this contradiction between his mental scheme and reality, but he wanted his memory aid to match his conceptual model. Conceptual models are powerful explanatory devices, useful in a variety of circumstances. They do not have to be accurate as long as they lead to the correct behavior in the desired situation.

EXAMPLE 4: “GOOD ENOUGH” ARITHMETIC

Most of us can’t multiply two large numbers in our head: we forget where we are along the way. Memory experts can multiply two large numbers quickly and effortlessly in their heads, amazing audiences with their skills. Moreover, the numbers come out left to right, the way we use them, not right to left, as we write them while laboriously using pencil and paper to compute the answers. These experts use special techniques that minimize the load on working memory, but they do so at the cost of having to learn numerous special methods for different ranges and forms of problems.

Isn’t this something we should all learn? Why aren’t school systems teaching this? My answer is simple: Why bother? I can estimate the answer in my head with reasonable accuracy, often good enough for the purpose. When I need precision and accuracy, well, that’s what calculators are for.

Remember my earlier example, to multiply 27 times 293 in your head? Why would anyone need to know the precise answer? an approximate answer is good enough, and pretty easy to get. Change 27 to 30, and 293 to 300: 30 × 300 = 9,000 (3 × 3 = 9, and add back the three zeros). The accurate answer is 7,911, so the estimate of 9,000 is only 14 percent too large. In many instances, this is good enough. Want a bit more accuracy? We changed 27 to 30 to make the multiplication easier. That’s 3 too large. So subtract 3 × 300 from the answer (9,000 – 900). Now we get 8,100, which is accurate within 2 percent.

It is rare that we need to know the answers to complex arithmetic problems with great precision: almost always, a rough estimate is good enough. When precision is required, use a calculator. That’s what machines are good for: providing great precision. For most purposes, estimates are good enough. Machines should focus on solving arithmetic problems. People should focus on higher-level issues, such as the reason the answer was needed.

Unless it is your ambition to become a nightclub performer and amaze people with great skills of memory, here is a simpler way to dramatically enhance both memory and accuracy: write things down. Writing is a powerful technology: why not use it? Use a pad of paper, or the back of your hand. Write it or type it. Use a phone or a computer. Dictate it. This is what technology is for.

The unaided mind is surprisingly limited. It is things that make us smart. Take advantage of them.

SCIENTIFIC THEORY VERSUS EVERYDAY PRACTICE

Science strives for truth. As a result, scientists are always debating, arguing, and disagreeing with one another. The scientific method is one of debate and conflict. Only ideas that have passed through the critical examination of multiple other scientists survive. This continual disagreement often seems strange to the nonscientist, for it appears that scientists don't know anything. Select almost any topic, and you will discover that scientists who work in that area are continually disagreeing.

But the disagreements are illusory. That is, most scientists usually agree about the broad details: their disagreements are often about tiny details that are important for distinguishing between two competing theories, but that might have very little impact in the real world of practice and applications.

In the real, practical world, we don't need absolute truth: approximate models work just fine. Professor Sayeki's simplified conceptual model of steering his motorcycle enabled him to remember which way to move the switches for his turn signals; the simplified equation for temperature conversion and the simplified model of approximate arithmetic enabled "good enough" answers in the head. The simplified model of STM provides useful design guidance, even if it is scientifically wrong. Each of these approximations is wrong, yet all are valuable in minimizing thought, resulting in quick, easy results whose accuracy is "good enough."

Knowledge in the Head

Knowledge in the world, external knowledge, is a valuable tool for remembering, but only if it is available at the right place, at the right time, in the appropriate situation. Otherwise, we must use knowledge in the head, in the mind. A folk saying captures this situation well: “Out of sight, out of mind.” Effective memory uses all the clues available: knowledge in the world and in the head, combining world and mind. We have already seen how the combination allows us to function quite well in the world even though either source of knowledge, by itself, is insufficient.

HOW PILOTS REMEMBER WHAT AIR-TRAFFIC CONTROL TELLS THEM

Airplane pilots have to listen to commands from air-traffic control delivered at a rapid pace, and then respond accurately. Their lives depend upon being able to follow the instructions accurately. One website, discussing the problem, gave this example of instructions to a pilot about to take off for a flight:

> *Frasca 141, cleared to Mesquite airport, via turn left heading 090, radar vectors to Mesquite airport. Climb and maintain 2,000. Expect 3,000 10 minutes after departure. Departure frequency 124.3, squawk 5270*.
>
> (Typical Air traffic control sequence, usually spoken extremely rapidly. Text from “ATC Phraseology,” on numerous websites, with no credit for originator.)

“How can we remember all that,” asked one novice pilot, “when we are trying to focus on taking off?” Good question. Taking off is a busy, dangerous procedure with a lot going on, both inside and outside the airplane. How do pilots remember? Do they have superior memories?

Pilots use three major techniques:

1. They write down the critical information.

2. They enter it into their equipment as it is told to them, so minimal memory is required.

3. They remember some of it as meaningful phrases.

Although to the outside observer, all the instructions and numbers seem random and confusing, to the pilots they are familiar names, familiar numbers. As one respondent pointed out, those are common numbers and a familiar pattern for a takeoff. “Frasca 141” is the name of the airplane, announcing the intended recipient of these instructions. The first critical item to remember is to turn left to a compass direction of 090, then climb to an altitude of 2,000 feet. Write those two numbers down. Enter the radio frequency 124.3 into the radio as you hear it—but most of the time this frequency is known in advance, so the radio is probably already set to it. All you have to do is look at it and see that it is set properly. Similarly, setting the “squawk box to 5270” is the special code the airplane sends whenever it is hit by a radar signal, identifying the airplane to the air-traffic controllers. Write it down, or set it into the equipment as it is being said. As for the one remaining item, “Expect 3,000 10 minutes after departure,” nothing need be done. This is just reassurance that in ten minutes, Frasca 141 will probably be advised to climb to 3,000 feet, but if so, there will be a new command to do so.

How do pilots remember? They transform the new knowledge they have just received into memory in the world, sometimes by writing, sometimes by using the airplane’s equipment.

The design implication? The easier it is to enter the information into the relevant equipment as it is heard, the less chance of memory error. The air-traffic control system is evolving to help. The instructions from the air-traffic controllers will be sent digitally, so that they can remain displayed on a screen as long as the pilot wishes. The digital transmission also makes it easy for automated equipment to set itself to the correct parameters. Digital transmission of the controller’s commands has some disadvantages, however. Other aircraft will not hear the commands, which reduces pilot awareness of what all the airplanes in the vicinity are going to do. Researchers in air-traffic control and aviation safety are looking into these issues. Yes, it’s a design issue.

REMINDING: PROSPECTIVE MEMORY

The phrases *prospective memory* or *memory for the future* might sound counterintuitive, or perhaps like the title of a science-fiction novel, but to memory researchers, the first phrase simply denotes the task of remembering to do some activity at a future time. The second phrase denotes planning abilities, the ability to imagine future scenarios. Both are closely related.

Consider reminding. Suppose you have promised to meet some friends at a local café on Wednesday at three thirty in the afternoon. The knowledge is in your head, but how are you going to remember it at the proper time? You need to be reminded. This is a clear instance of prospective memory, but your ability to provide the required cues involves some aspect of memory for the future as well. Where will you be Wednesday just before the planned meeting? What can you think of now that will help you remember then?

There are many strategies for reminding. One is simply to keep the knowledge in your head, trusting yourself to recall it at the critical time. If the event is important enough, you will have no problem remembering it. It would be quite strange to have to set a calendar alert to remind yourself, “Getting married at 3 PM.”

Relying upon memory in the head is not a good technique for commonplace events. Ever forget a meeting with friends? It happens a lot. Not only that, but even if you might remember the appointment, will you remember all the details, such as that you intended to loan a book to one of them? Going shopping, you may remember to stop at the store on the way home, but will you remember all the items you were supposed to buy?

If the event is not personally important and several days away, it is wise to transfer some of the burden to the world: notes, calendar reminders, special cell phone or computer reminding services. You can ask friends to remind you. Those of us with assistants put the burden on them. They, in turn, write notes, enter events on calendars, or set alarms on their computer systems.

Why burden other people when we can put the burden on the thing itself? Do I want to remember to take a book to a colleague? I put the book someplace where I cannot fail to see it when I leave the house. A good spot is against the front door so that I can't leave without tripping over it. Or I can put my car keys on it, so when I leave, I am reminded. Even if I forget, I can't drive away without the keys. (Better yet, put the keys under the book, else I might still forget the book.)

There are two different aspects to a reminder: the signal and the message. Just as in doing an action we can distinguish between knowing *what* can be done and knowing *how* to do it, in reminding we must distinguish between the *signal*—knowing that something is to be remembered, and the *message*—remembering the information itself. Most popular reminding methods typically provide only one or the other of these two critical aspects. The famous "tie a string around your finger" reminder provides only the signal. It gives no hint of what is to be remembered. Writing a note to yourself provides only the message; it doesn't remind you ever to look at it. The ideal reminder has to have both components: the signal that something is to be remembered, and then the message of what it is.

The signal that something is to be remembered can be a sufficient memory cue if it occurs at the correct time and place. Being reminded too early or too late is just as useless as having no reminder. But if the reminder comes at the correct time or location, the environmental cue can suffice to provide enough knowledge to aid retrieval of the to-be-remembered item. Time-based reminders can be effective: the *bing* of my cell phone reminds me of the next appointment. Location-based reminders can be effective in giving the cue at the precise place where it will be needed. All the knowledge needed can reside in the world, in our technology.

The need for timely reminders has created loads of products that make it easier to put the knowledge in the world—timers, diaries, calendars. The need for electronic reminders is well known, as the proliferation of apps for smart phones, tablets, and other portable devices attests. Yet surprisingly in this era of screen-based devices, paper tools are still enormously popular and effective, as the number of paper-based diaries and reminders indicates.

The sheer number of different reminder methods also indicates that there is indeed a great need for assistance in remembering, but that none of the many schemes and devices is completely satisfactory. After all, if any one of them was, then we wouldn't need so many. The less effective ones would disappear and new schemes would not continually be invented.

The Tradeoff Between Knowledge in the World and in the Head

Knowledge in the world and knowledge in the head are both essential in our daily functioning. But to some extent we can choose to lean more heavily on one or the other. That choice requires a tradeoff—gaining the advantages of knowledge in the world means losing the advantages of knowledge in the head (Table 3.1).

Knowledge in the world acts as its own reminder. It can help us recover structures that we otherwise would forget. Knowledge in the head is efficient: no search and interpretation of the environment is required. The tradeoff is that to use our knowledge in the head, we have to be able to store and retrieve it, which might require considerable amounts of learning. Knowledge in the world requires no learning, but can be more difficult to use. And it relies heavily upon the continued physical presence of the knowledge; change the environment and the knowledge might be lost. Performance relies upon the physical stability of the task environment.

TABLE 3.1. Tradeoffs Between Knowledge in the World and in the Head

Knowledge in the World	Knowledge in the Head
Information is readily and easily available whenever perceivable.	Material in working memory is readily available. Otherwise considerable search and effort may be required.

Interpretation substitutes for learning. How easy it is to interpret knowledge in the world depends upon the skill of the designer.	Requires learning, which can be considerable. Learning is made easier if there is meaning or structure to the material or if there is a good conceptual model.
Slowed by the need to find and interpret the knowledge.	Can be efficient, especially if so well-learned that it is automated.
Ease of use at first encounter is high.	Ease of use at first encounter is low.
Can be ugly and inelegant, especially if there is a need to maintain a lot of knowledge. This can lead to clutter. Here is where the skills of the graphics and industrial designer play major roles.	Nothing needs to be visible, which gives more freedom to the designer. This leads to cleaner, more pleasing appearance—at the cost of ease of use at first encounter, learning, and remembering.

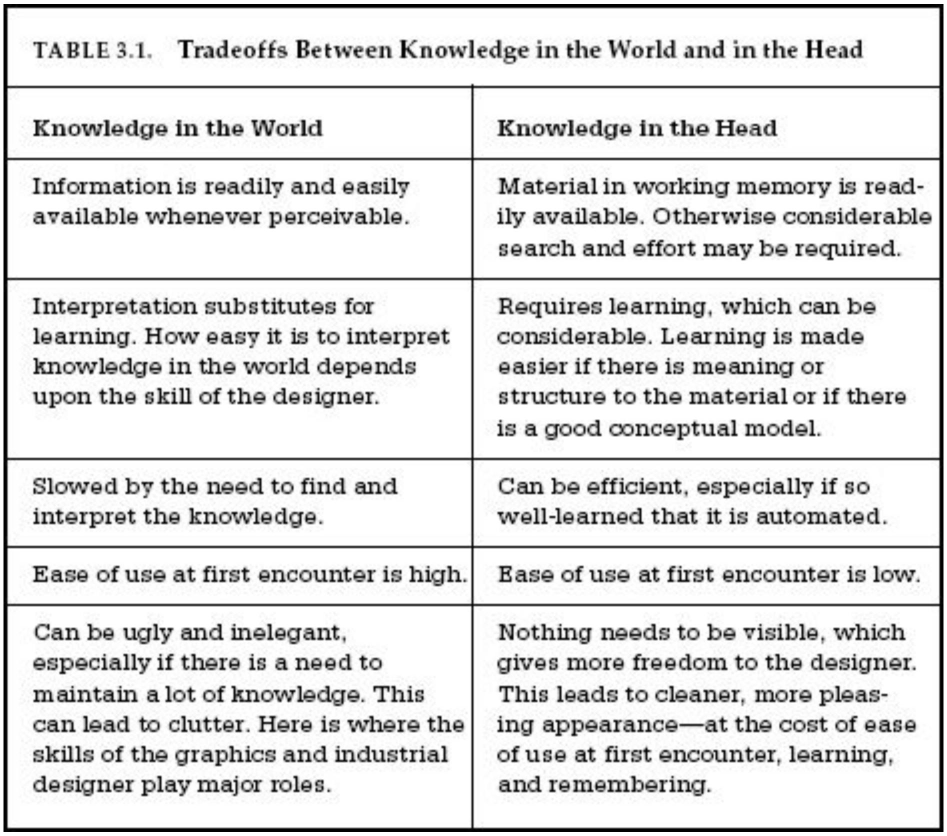


TABLE 3.1. **Tradeoffs Between Knowledge in the World and in the Head**

Knowledge in the World	**Knowledge in the Head**
Information is readily and easily available whenever perceivable.	Material in working memory is readily available. Otherwise considerable search and effort may be required.
Interpretation substitutes for learning. How easy it is to interpret knowledge in the world depends upon the skill of the designer.	Requires learning, which can be considerable. Learning is made easier if there is meaning or structure to the material or if there is a good conceptual model.
Slowed by the need to find and interpret the knowledge.	Can be efficient, especially if so well-learned that it is automated.
Ease of use at first encounter is high.	Ease of use at first encounter is low.
Can be ugly and inelegant, especially if there is a need to maintain a lot of knowledge. This can lead to clutter. Here is where the skills of the graphics and industrial designer play major roles.	Nothing needs to be visible, which gives more freedom to the designer. This leads to cleaner, more pleasing appearance—at the cost of ease of use at first encounter, learning, and remembering.

As we just discussed, reminders provide a good example of the relative tradeoffs between knowledge in the world versus in the head. Knowledge in the world is accessible. It is self-reminding. It is always there, waiting to be seen, waiting to be used. That is why we structure our offices and our places of work so carefully. We put piles of papers where they can be seen, or if we like a clean desk, we put them in standardized locations and teach ourselves (knowledge in the head) to look in these standard places routinely. We use clocks and calendars and notes. Knowledge in the mind is ephemeral: here now, gone later. We can't count on something being present in mind at any particular time, unless it is triggered by some external event or unless we deliberately keep it in mind through constant repetition (which then prevents us from having other conscious thoughts). Out of sight, out of mind.

As we move away from many physical aids, such as printed books and magazines, paper notes, and calendars, much of what we use today as knowledge in the world will become invisible. Yes, it will all be available on display screens, but unless the screens always show this material, we will have added to the burden of memory in the head. We may not have to remember all the details of the information stored away for us, but we will have to remember that it is there, that it needs to be redisplayed at the appropriate time for use or for reminding.

Memory in Multiple Heads, Multiple Devices

If knowledge and structure in the world can combine with knowledge in the head to enhance memory performance, why not use the knowledge in multiple heads, or in multiple devices?

Most of us have experienced the power of multiple minds in remembering things. You are with a group of friends trying to remember the name of a movie, or perhaps a restaurant, and failing. But others try to help. The conversation goes something like this:

“That new place where they grill meat”

“Oh, the Korean barbecue on Fifth Street?”

“No, not Korean, South American, um,“

“Oh, yeah, Brazilian, it’s what’s its name?”

“Yes, that’s the one!”

“Pampas something.”

“Yes, Pampas Chewy. Um, Churry, um,”

“Churrascaria. Pampas Churrascaria.”

How many people are involved? It could be any number, but the point is that each adds their bit of knowledge, slowly constraining the choices, recalling something that no single one of them could have done alone. Daniel Wegner, a Harvard professor of psychology, has called this “transactive memory.”

Of course, we often turn to technological aids to answer our questions, reaching for our smart devices to search our electronic resources and the Internet. When we expand from seeking aids from other people to seeking aids from our technologies, which Wegner labels as “cybermind,” the principle is basically the same. The cybermind doesn’t always produce the answer, but it can produce sufficient clues so that we can generate the answer. Even where the technology produces the answer, it is often buried in a list of potential answers, so we have to use our own knowledge— or the knowledge of our friends—to determine which of the potential items is the correct one.

What happens when we rely too much upon external knowledge, be it knowledge in the world, knowledge of friends, or knowledge provided by our technology? On the one hand, there no such thing as “too much.” The more we learn to use these resources, the better our performance. External knowledge is a powerful tool for enhanced intelligence. On the other hand, external knowledge is often erroneous: witness the difficulties of trusting online sources and the controversies that arise over Wikipedia entries. It doesn’t matter where our knowledge comes from. What matters is the quality of the end result.

In an earlier book, *Things That Make Us Smart*, I argued that it is this combination of technology and people that creates super-powerful beings. Technology does not make us smarter. People do not make technology smart. It is the combination of the two, the person plus the artifact, that is smart. Together, with our tools, we are a powerful combination. On the other hand, if we are suddenly without these external devices, then we don’t do very well. In many ways, we do become less smart.

Take away their calculator, and many people cannot do arithmetic. Take away a navigation system, and people can no longer get around, even in their own cities. Take away a phone’s or computer’s address book, and people can no longer reach their friends (in my case, I can no longer remember my own phone number). Without a keyboard, I can’t write. Without a spelling corrector, I can’t spell.

What does all of this mean? Is this bad or good? It is not a new phenomenon. Take away our gas supply and electrical service and we might starve. Take away our housing and clothes and we might freeze.

We rely on commercial stores, transportation, and government services to provide us with the essentials for living. Is this bad?

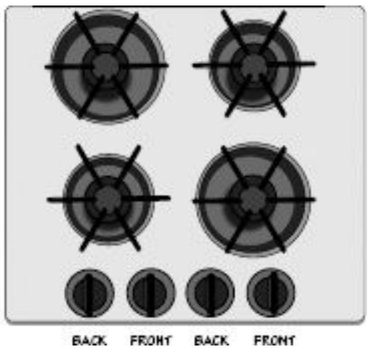
The partnership of technology and people makes us smarter, stronger, and better able to live in the modern world. We have become reliant on the technology and we can no longer function without it. The dependence is even stronger today than ever before, including mechanical, physical things such as housing, clothing, heating, food preparation and storage, and transportation. Now this range of dependencies is extended to information services as well: communication, news, entertainment, education, and social interaction. When things work, we are informed, comfortable, and effective. When things break, we may no longer be able to function. This dependence upon technology is very old, but every decade, the impact covers more and more activities.

Natural Mapping

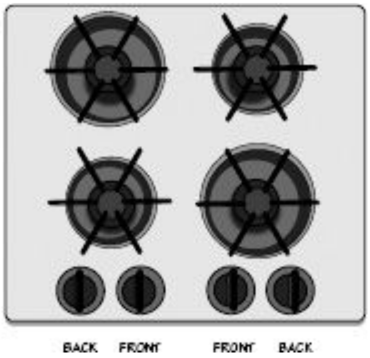
Mapping, a topic from Chapter 1, provides a good example of the power of combining knowledge in the world with that in the head. Did you ever turn the wrong burner of a stove on or off? You would think that doing it correctly would be an easy task. A simple control turns the burner on, controls the temperature, and allows the burner to be turned off. In fact, the task appears to be so simple that when people do it wrong, which happens more frequently than you might have thought, they blame themselves: “How could I be so stupid as to do this simple task wrong?” they think to themselves. Well, it isn’t so simple, and it is not their fault: even as simple a device as the everyday kitchen stove is frequently badly designed, in a way that guarantees the errors.

Most stoves have only four burners and four controls in one-to-one correspondence. Why is it so hard to remember four things? In principle, it should be easy to remember the relationship between the controls and the burners. In practice, however, it is almost impossible. Why? Because of the poor mappings between the controls and the burners. Look at Figure 3.2, which depicts four possible mappings between the four burners and controls. Figures 3.2A and B show how not to map one dimension onto two. Figures 3.2C and D show two ways of doing it properly: arrange the controls in two dimensions (C) or stagger the burners (D) so they can be ordered left to right.

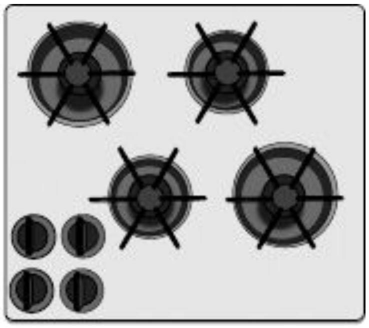
A.



B.



C.



D.

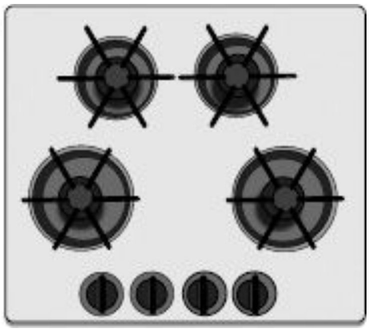


FIGURE 3.2. Mappings of Stove Controls with Burners. With the traditional arrangement of stove burners shown in Figures A and B, the burners are arranged in a rectangle and the controls in a linear line. Usually there is a partial natural mapping, with the left two controls operating the left burners and the right two controls operating the right burners. Even so, there are four possible mappings of controls to burners, all four of which are used on commercial stoves. The only way to know which control works which burner is to read the labels. But if the controls were also in a rectangle (Figure C) or the burners staggered (Figure D), no labels would be needed. Learning would be easy; errors would be reduced.

To make matters worse, stove manufacturers cannot agree upon what the mapping should be. If all stoves used the same arrangement of controls, even if it is unnatural, everyone could learn it once and forever after get things right. As the legend of Figure 3.2 points out, even if the stove manufacturer is nice enough to ensure that each pair of controls operates the pair of burners on its side, there are still four possible mappings. All four are in common use. Some stoves arrange the controls in a vertical line, giving even more possible mappings. Every stove seems to be different. Even different stoves from the same manufacturer differ. No wonder people have trouble, leading their food to go uncooked, and in the worst cases, leading to fire.

Natural mappings are those where the relationship between the controls and the object to be controlled (the burners, in this case) is obvious. Depending upon circumstances, natural mappings will employ spatial cues. Here are three levels of mapping, arranged in decreasing effectiveness as memory aids:

- **Best mapping:** Controls are mounted directly on the item to be controlled.

- **Second-best mapping:** Controls are as close as possible to the object to be controlled.

- **Third-best mapping:** Controls are arranged in the same spatial configuration as the objects to be controlled.

In the ideal and second-best cases, the mappings are indeed clear and unambiguous.

Want excellent examples of natural mapping? Consider gesture-controlled faucets, soap dispensers, and hand dryers. Put your hands under the faucet or soap dispenser and the water or soap appears. Wave your hand in front of the paper towel dispenser and out pops a new towel, or in the case of blower-controlled hand dryers, simply put your hands beneath or into the dryer and the drying air turns on. Mind you, although the mappings of these devices are appropriate, they do have problems. First, they often lack signifiers, hence they lack discoverability. The controls are often invisible, so we sometimes put our hands under faucets expecting to receive water, but wait in vain: these are mechanical faucets that require handle turning. Or the water turns on and then stops, so we wave our hands up and down, hoping to find the precise location where the water turns on. When I wave my hand in front of the towel dispenser but get no towel, I do not know whether this means the dispenser is broken or out of towels; or that I did the waving wrong, or in the wrong place; or that maybe this doesn’t work by gesture, but I must push, pull, or turn something. The lack of signifiers is a real drawback. These devices aren’t perfect, but at least they got the mapping right.

In the case of stove controls, it is obviously not possible to put the controls directly on the burners. In most cases, it is also dangerous to put the controls adjacent to the burners, not only for fear of burning the person using the stove, but also because it would interfere with the placement of cooking utensils. Stove controls are usually situated on the side, back, or front panel of the stove, in which case they ought to be arranged in spatial harmony with the burners, as in Figures 3.2 C and D.

With a good natural mapping, the relationship of the controls to the burner is completely contained in the world; the load on human memory is much reduced. With a bad mapping, however, a burden is placed upon memory, leading to more mental effort and a higher chance of error. Without a good mapping, people new to the stove cannot readily determine which burner goes with which control and even frequent users will still occasionally err.

Why do stove designers insist on arranging the burners in a two-dimensional rectangular pattern, and the controls in a one-dimensional row? We have known for roughly a century just how bad such an arrangement is. Sometimes the stove comes with clever little diagrams to indicate which control works which burner. Sometimes there are labels. But the proper natural mapping requires no diagrams, no labels, and no instructions.

The irony about stove design is that it isn't hard to do right. Textbooks of ergonomics, human factors, psychology, and industrial engineering have been demonstrating both the problems and the solutions for over fifty years. Some stove manufacturers do use good designs. Oddly, sometimes the best and the worst designs are manufactured by the same companies and are illustrated side by side in their catalogs. Why do users still purchase stoves that cause so much trouble? Why not revolt and refuse to buy them unless the controls have an intelligent relationship to the burners?

The problem of the stovetop may seem trivial, but similar mapping problems exist in many situations, including commercial and industrial settings, where selecting the wrong button, dial, or lever can lead to major economic impact or even fatalities.

In industrial settings good mapping is of special importance, whether it is a remotely piloted airplane, a large building crane where the operator is at a distance from the objects being manipulated, or even in an automobile where the driver might wish to control temperature or windows while driving at high speeds or in crowded streets. In these cases, the best controls usually are spatial mappings of the controls to the items being controlled. We see this done properly in most automobiles where the driver can operate the windows through switches that are arranged in spatial correspondence to the windows.

Usability is not often thought about during the purchasing process. Unless you actually test a number of units in a realistic environment, doing typical tasks, you are not likely to notice the ease or difficulty of use. If you just look at something, it appears straightforward enough, and the array of wonderful features seems to be a virtue. You may not realize that you won't be able to figure out how to use those features. I urge you to test products before you buy them. Before purchasing a new stovetop, pretend you are cooking a meal. Do it right there in the store. Do not be afraid to make mistakes or ask stupid questions. Remember, any problems you have are probably the design's fault, not yours.

A major obstacle is that often the purchaser is not the user. Appliances may be in a home when people move in. In the office, the purchasing department orders equipment based upon such factors as price, relationships with the supplier, and perhaps reliability: usability is seldom considered. Finally, even when the purchaser is the end user, it is sometimes necessary to trade off one desirable feature for an undesirable one. In the case of my family's stove, we did not like the arrangement of controls, but we bought the stove anyway: we traded off the layout of the burner controls for another design feature that was more important to us and available only from one manufacturer. But why should we have to make a tradeoff? It wouldn't be hard for all stove manufacturers to use natural mappings, or at the least, to standardize their mappings.

Culture and Design: Natural Mappings Can Vary with Culture

I was in Asia, giving a talk. My computer was connected to a projector and I was given a remote controller for advancing through the illustrations for my talk. This one had two buttons, one above the other. The title was already displayed on the screen, so when I started, all I had to do was to advance to the first photograph in my presentation, but when I pushed the upper button, to my amazement I went backward through my illustrations, not forward.

"How could this happen?" I wondered. To me, top means forward; bottom, backward. The mapping is clear and obvious. If the buttons had been side by side, then the control would have been ambiguous: which comes first, right or left? This controller appeared to use an appropriate mapping of top and bottom. Why was it working backward? Was this yet another example of poor design?

I decided to ask the audience. I showed them the controller and asked: "To get to my next picture, which button should I push, the top or the bottom?” To my great surprise, the audience was split in their responses. Many thought that it should be the top button, just as I had thought. But a large number thought it should be the bottom.

What’s the correct answer? I decided to ask this question to my audiences around the world. I discovered that they, too, were split in their opinions: some people firmly believe that it is the top button and some, just as firmly, believe it is the bottom button. Everyone is surprised to learn that someone else might think differently.

I was puzzled until I realized that this was a point-of-view problem, very similar to the way different cultures view time. In some cultures, time is represented mentally as if it were a road stretching out ahead of the person. As a person moves through time, the person moves forward along the time line. Other cultures use the same representation, except now it is the person who is fixed and it is time that moves: an event in the future moves toward the person.

This is precisely what was happening with the controller. Yes, the top button does cause something to move forward, but the question is, what is moving? Some people thought that the person would move through the images, other people thought the images would move. People who thought that they moved through the images wanted the top button to indicate the next one. People who thought it was the illustrations that moved would get to the next image by pushing the bottom button, causing the images to move toward them.

Some cultures represent the time line vertically: up for the future, down for the past. Other cultures have rather different views. For example, does the future lie ahead or behind? To most of us, the question makes no sense: of course, the future lies ahead—the past is behind us. We speak this way, discussing the “arrival” of the future; we are pleased that many unfortunate events of the past have been “left behind.”

But why couldn’t the past be in front of us and the future behind? Does that sound strange? Why? We can see what is in front of us, but not what is behind, just as we can remember what happened in the past, but we can’t remember the future. Not only that, but we can remember recent events much more clearly than long-past events, captured neatly by the visual metaphor in which the past lines up before us, the most recent events being the closest so that they are clearly perceived (remembered), with long-past events far in the distance, remembered and perceived with difficulty. Still sound weird? This is how the South American Indian group, the Aymara, represent time. When they speak of the future, they use the phrase *back days* and often gesture behind them. Think about it: it is a perfectly logical way to view the world.

If time is displayed along a horizontal line, does it go from left to right or right to left? Either answer is correct because the choice is arbitrary, just as the choice of whether text should be strung along the page from left to right or right to left is arbitrary. The choice of text direction also corresponds to people's preference for time direction. People whose native language is Arabic or Hebrew prefer time to flow from right to left (the future being toward the left), whereas those who use a left-to-right writing system have time flowing in the same direction, so the future is to the right.

But wait: I'm not finished. Is the time line relative to the person or relative to the environment? In some Australian Aborigine societies, time moves relative to the environment based on the direction in which the sun rises and sets. Give people from this community a set of photographs structured in time (for example, photographs of a person at different ages or a child eating some food) and ask them to order the photographs in time. People from technological cultures would order the pictures from left to right, most recent photo to the right or left, depending upon how their printed language was written. But people from these Australian communities would order them east to west, most recent to the west. If the person were facing south, the photo would be ordered left to right. If the person were facing north, the photos would be ordered right to left. If the person were facing west, the photos would be ordered along a vertical line extending from the body outward, outwards being the most recent. And, of course, were the person facing east, the photos would also be on a line extending out from the body, but with the most recent photo closest to the body.

The choice of metaphor dictates the proper design for interaction. Similar issues show up in other domains. Consider the standard problem of scrolling the text in a computer display. Should the scrolling control move the text or the window? This was a fierce debate in the early years of display terminals, long before the development of modern computer systems. Eventually, there was mutual agreement that the cursor arrow keys—and then, later on, the mouse—would follow the moving window metaphor. Move the window down to see more text at the bottom of the screen. What this meant in practice is that to see more text at the bottom of the screen, move the mouse down, which moves the window down, so that the text moves up: the mouse and the text move in opposite directions. With the moving text metaphor, the mouse and the text move in the same directions: move the mouse up and the text moves up. For over two decades, everyone moved the scrollbars and mouse down in order to make the text move up.

But then smart displays with touch-operated screens arrived. Now it was only natural to touch the text with the fingers and move it up, down, right, or left directly: the text moved in the same direction as the fingers. The moving text metaphor became prevalent. In fact, it was no longer thought of as a metaphor: it was real. But as people switched back and forth between traditional computer systems that used the moving window metaphor and touch-screen systems that used the moving text model, confusion reigned. As a result, one major manufacturer of both computers and smart screens, Apple, switched everything to the moving text model, but no other company followed Apple's lead. As I write this, the confusion still exists. How will it end? I predict the demise of the moving window metaphor: touch-screens and control pads will dominate, which will cause the moving text model to take over. All systems will move the hands or controls in the same direction as they wish the screen images to move. Predicting technology is relatively easy compared to predictions of human behavior, or in this case, the adoption of societal conventions. Will this prediction be true? You will be able to judge for yourself.

Similar issues occurred in aviation with the pilot's attitude indicator, the display that indicates the airplane's orientation (roll or bank and pitch). The instrument shows a horizontal line to indicate the horizon with a silhouette of an airplane seen from behind. If the wings are level and on a line with the horizon, the airplane is flying in level flight. Suppose the airplane turns to the left, so it banks (tilts) left. What should the display look like? Should it show a left-tilting airplane against a fixed horizon, or a fixed airplane against a right-tilting horizon? The first is correct from the viewpoint of someone watching the airplane from behind, where the horizon is always horizontal: this type of display is called *outside-in*. The second is correct from the viewpoint of the pilot, where the airplane is always stable and fixed in position, so that when the airplane banks, the horizon tilts: this type of display is called *inside-out*.

In all these cases, every point of view is correct. It all depends upon what you consider to be moving. What does all this mean for design? What is natural depends upon point of view, the choice of metaphor, and therefore, the culture. The design difficulties occur when there is a switch in metaphors. Airplane pilots have to undergo training and testing before they are allowed to switch from one set of instruments (those with an outside-in metaphor, for example) to the other (those with the inside-out metaphor). When countries decided to switch which side of the road cars would drive on, the temporary confusion that resulted was dangerous. (Most places that switched moved from left-side driving to right-side, but a few, notably Okinawa, Samoa, and East Timor, switched from right to left.) In all these cases of convention switches, people eventually adjusted. It is possible to break convention and switch metaphors, but expect a period of confusion until people adapt to the new system.