1 What psychology is

1.1 A science of mind

We spend an enormous number of our waking hours thinking and talking about our thoughts, emotions, and experiences. For example, we wonder: Why did the waiter give me that unusual smile? Did my co-worker see me stealing those office supplies? How can I deflect my unwanted admirer's attention – or attract the attention of someone else? In trying to answer such questions, and in interpreting one another's behavior more generally, we make use of a vast body of lore about how people perceive, reason, desire, feel, and so on. So we say such things as: the waiter is smiling obsequiously because he hopes I will give him a larger tip; my co-worker does know, but he won't tell anyone, because he's afraid I'll reveal his gambling problem; and so on. Formulating such explanations is part of what enables us to survive in a shared social environment.

This everyday understanding of our minds, and those of others, is referred to as "folk psychology." The term is usually taken as picking out our ability to attribute psychological states and to use those attributions for a variety of practical ends, including prediction, explanation, manipulation, and deception. It encompasses our ability to verbally produce accounts couched in the everyday psychological vocabulary with which most of us are conversant: the language of beliefs, desires, intentions, fears, hopes, and so on. Such accounts are the stuff of which novels and gossip are made. Although our best evidence for what people think is often what they say, much of our capacity to read the thoughts of others may also be nonverbal, involving the ability to tell moods and intentions immediately by various bodily cues – an ability we may not be conscious that we have.

Although we have an important stake in the success of our folk psychological attributions and explanations, and while social life as we know it would

be impossible without folk psychology, folk psychology also has obvious shortcomings (Churchland, 1981). Our accounts of one another's behavior are often sketchy, unsystematic, or of merely local utility. Moreover, they leave out whole ranges of abnormal mental phenomena such as autism or Capgras syndrome. We have no folk explanation for how we are able to perceive and navigate our way through a three-dimensional space cluttered with objects, how we integrate what we see with what we hear and touch, how we are able to learn language, how we recognize faces and categories, how our memory works, how we reason and make decisions, and so on. The explanations of these varied mental capacities lie far beyond folk psychology's province. If we want to understand the mind, then we need to find better ways to investigate its structure and function. The sciences of the mind have developed in response to this need.

Science aims at systematic understanding of the world, and psychology is the science that takes mental phenomena in general as its domain. This definition has not always been uncontroversially accepted. Behaviorists such as Watson (1913) and Skinner (1965) held that the only proper subject matter for psychology was the domain of observable behavior, in part on the grounds that minds were mysterious and inaccessible to third-person methods of investigation. Few today take this position. Mental states and processes may not be directly observable, but they can be inferred by a variety of converging techniques. Cognitive psychology in particular typically proceeds by positing such inferred states. Many of these states such as occurrent perceptions and thoughts are accessible via introspection with varying degrees of accuracy, but many are entirely unconscious.

"Phenomena" is a cover term for the body of noteworthy natural regularities to be found in the objects, events, processes, activities, and capacities that a science concerns itself with.[1] Objects can include such things as whole organisms (white rats, the sea slug *Aplysia californica*), artificial behaving systems (a trained neural network, an autonomous mobile robot), or their parts (the brain, particular brain structures such as the hippocampus or the supplementary motor area, a particular control structure in a computer). Here the relevant phenomena are reliable patterns of organization or behavior in these objects – for example, the predictable laminar organization and connectivity patterns in the neocortex. Events and processes include any changes

[1] This usage follows Hacking (1983). See also Bogen and Woodward (1988).

undergone by these objects: the myelination of the frontal lobes in normal development, a rat's learning to run a water maze, a child acquiring the lexicon of her first language, an undergraduate carrying out a motor task in response to a visual stimulus, a patient with dementia retrieving a memory of an event from his teenage years. Activities and capacities include any functions that an object can reliably carry out. Normal humans have the capacity to rapidly estimate quantity, to selectively attend to parts of a complex visual array, to judge which of two events is more likely, to generate expectations about the movement of simple physical objects in their environment, to attribute emotional states to others, and so on.

Mental phenomena encompass attention, learning and memory, concept acquisition and categorization, language acquisition, perception (both accurate and illusory), and emotions and moods, among others. We won't try to be exhaustive. Traditional distinctions among types of mental states have been made along the following lines. Some mental states involve concepts in their formation, expression, and function. These are the types of states associated with higher cognition and knowledge (from which "cognitive" derives its name). Such states include beliefs, desires, intentions, hopes, and plain old thoughts in general. Other sorts of states, such as sensory states, do not necessarily involve concepts in their activation. One can smell a rose without knowing that it is a rose one smells. One can hear a C-sharp on the piano without knowing that it is a C-sharp one hears. Emotions such as fear, love, and anger also form a distinctive class of mental states. Finally, there are moods: general overall feelings of excitement, happiness, sadness, mania, and depression.

Is there anything that all mental phenomena have in common? This is controversial, but one proposal is that they are all *representational*.[2] The higher cognitive states that involve concepts clearly involve representations that can fit into propositional attitudes and generate knowledge of various facts and states of affairs. Sensory states do not necessarily involve the activation of concepts, but they are still a type of representation on at least some views. They represent the presence of a physically perceptible property and the causal interaction of that property with a sensory system of the body. The sweet taste of sugar represents the interaction of the sugar molecules with

[2] We discuss the issue of how to distinguish mental phenomena in greater depth in Section 5.4.4.

the taste receptors in the mouth, for instance. Even moods have been portrayed as representations of general chemical states or changes in the body.

One goal of the sciences is to describe, clarify, and organize these phenomena. Consider the changes that the past 50 years have wrought in our understanding of the cognitive capacities of infants and young children, for example. At some point, normal children become able to understand and interpret the behavior of others in terms of their beliefs, intentions, and desires. In a pioneering study, Wimmer and Perner (1983) showed that four-year-olds are able to correctly predict how characters with false beliefs will act, whereas younger children are unable to do so. In one of their now-classic tasks, the child watches one puppet place a piece of candy in a certain location and then leave the room. The other puppet, which was present when the candy was hidden, now moves it to a new hidden location. The first puppet then returns, and the child is asked either where she will look for the candy or where she thinks the candy is. Passing this so-called false belief task involves correctly saying that she will look in the original location, rather in the actual location, since she will be guided not by the candy's actual location, but by her erroneous beliefs about it. Here the phenomenon of interest is the alleged shift from failure to success in this particular test (and related variants). This result was widely interpreted as showing that some components of "theory of mind" – those connected with the attribution of beliefs – are not yet in place prior to age four.[3]

Surprisingly, though, in recent years it has been shown that even 15-month-olds can respond in a way that seems to display understanding of false beliefs (Onishi & Baillargeon, 2005). These infants will look longer at a scene depicting a character searching in a place that she could not know an object is located (because she had earlier seen it hidden elsewhere) than at a scene in which she searched for it in the place where she should expect it to be. Looking time in infants is often taken to be an indicator of surprise or violation of expectancy, an interpretation confirmed by studies across many different stimuli and domains. Thus the 15-month-olds in this study don't seem to expect the characters to have information about the true state of the world; this strongly suggests that they naturally attribute something like false beliefs. Moreover, 16-month-olds will even act on this understanding, trying to help out individuals who are attempting to act on false beliefs by

[3] For much more on theory of mind, see Chapter 8.

pointing to the correct location of a hidden toy (Buttelmann, Carpenter, & Tomasello, 2009).

This case illustrates two points. First, what the phenomena are in psychology, as in other sciences, is often nonobvious. That is, one cannot, in general, simply look and see that a certain pattern or regularity exists. Experiment and measurement are essential for the production of many interesting psychological phenomena. Second, phenomena are almost always tied closely to experimental tasks or paradigms. The phenomenon of three-year-olds failing the false belief task and four-year-olds passing it depends greatly on *which* false belief task one uses. If we agree to call the nonverbal Onishi and Baillargeon paradigm a false belief task, we need to explain the seeming contradiction between the phenomena, perhaps in terms of the differing requirements of the tasks (Bloom & German, 2000). Individuating phenomena is intimately tied to individuating tasks and experimental methods.

To see this, consider the Stroop effect. In his classic paper, Stroop (1935) performed three experiments, the first two of which are the most well known. In experiment 1, he asked participants to read color names printed in a variety of differently colored inks. The names were given in a 10 × 10 grid, and no name was ever paired with the color of ink that it named. The control condition required reading the same names printed in black ink. Subtracting the time to read the experimental versus the control cards, Stroop found that on average it took slightly longer to read the color names printed in differently colored ink, but this difference was not significant. In experiment 2, he required participants to name the color of the ink in the experimental condition, rather than reading the color name. In the control condition, words were replaced with colored squares. Here the difference in reading times was striking: participants were 74% slower to name the ink color when it conflicted with the color name versus simply naming the color from a sample. Conflicting lexical information interferes with color naming.

Although this is the canonical "Stroop effect," the term has been broadened over time to include a range of related phenomena. Stroop-like tasks have been carried out using pictures or numbers versus words, using auditory rather than visual materials, using nonverbal response measures, and so on. Further manipulations have involved varying the time at which the conflicting stimulus is presented (e.g., showing the color sample before the word), and the effect persists. Wherever responding to one kind of information interferes asymmetrically with responding to another that is simultaneously

presented, we have a Stroop-like phenomenon. Much of the literature on the effect has focused on delineating the precise sorts of stimuli, tasks, and populations that display the effect (MacLeod, 1991). But the effect itself is elusive outside the context of these experimental manipulations – certainly it is not a straightforwardly observable behavioral regularity on a par with wincing in response to being kicked. More esoteric phenomena may be reliant on even more sophisticated experimental setups for their elicitation.

In these cases, what psychologists are primarily aiming to do is to *characterize* the phenomena. This may require deploying new experimental paradigms, modifying the parameters of old paradigms, or refining techniques of data collection and analysis. The phenomena themselves are dependent on these techniques of investigation for their existence. Producing and measuring these phenomena involve discovering how various parts of the psychological domain behave when placed in relatively artificial circumstances, under the assumption that this will be importantly revealing about their normal structure and function. This is perhaps the biggest advantage scientific psychology has over its folk counterpart, which tends to be resolutely nonexperimental.

But beyond producing and describing phenomena – that is, saying *what* happens in the world – psychology also aims to explain *how* and *why* they are produced. Where we are dealing with genuine, robust phenomena, we assume, as an initial hypothesis at least, that they are not merely accidental. There ought to be some reason why they exist and take the particular form that they do. It is sometimes maintained that what is distinctive about scientific theorizing, as opposed to other ways of reasoning about the world, is that it involves positing and testing explanations. As we have seen, this can't be the whole story, because making and refining ways in which we might better describe the world are themselves major parts of the scientific enterprise. But the psychological phenomena we discover often turn out to be novel or surprising. Hence better descriptions of the phenomena naturally tend to pull us toward generating explanations for their existence.

1.2 Explanations in psychology

We shouldn't assume that all sciences will deploy the same explanatory strategies. What works to explain geological or astronomical phenomena may not work for psychological phenomena. So we begin by considering four sample

cases of psychological explanation. We should note that these explanations are to varying degrees contested, but the present issue is what they can tell us about the structure of explanations in psychology, rather than whether they are strictly accurate.

1.2.1 Case 1: Psychophysics

Some of the earliest systematic psychological research in the nineteenth century concerned psychophysical phenomena, in particular how the properties of sensations depend on and vary with the properties of the physical stimulus that produces them. Light, sound waves, pressure, temperature, and other ambient energy sources interact with sensory receptors and their associated processing systems to give rise to sensations, and this relationship is presumably systematic rather than random. To uncover this hidden order, early psychophysicists had to solve three problems simultaneously: (1) how to devise empirical strategies for measuring sensations, (2) how to quantify the ways in which those sensations covaried with stimulus conditions, and, finally, (3) how to explain those covariations.

Fechner (1860), following the work of Weber (1834), hit on the method of using "just noticeable differences" (jnd's) to measure units of sensation. A stimulus in some sensory modality (e.g., a patch of light, a tone) is increased in intensity until the perceiver judges that there is a detectable change in the quality of her sensations. The measure of a jnd in physical terms is the difference between the initial and final stimulus magnitude. By increasing stimulus intensity until the next jnd was reached, Fechner could plot the intervals at which a detectable change in a sensation occurred against the stimulus that caused the change.

After laboriously mapping stimulus–sensation pairs in various modalities, Fechner proposed a logarithmic law to capture their relationship formally. Fechner's law states:

$$S = k \log(I)$$

where S is the perceived magnitude of the sensation (e.g., the brightness of a light or the loudness of a sound), I is the intensity of the physical stimulus, and k is an empirically determined constant. Because this is a logarithmic

law, geometric increases in stimulus intensity will correspond to arithmetic increases in the strength of sensations.

Although Fechner's law delivers predictions that conform with much of the data, it also fails in some notable cases. Stevens (1957) took a different experimental approach. Rather than constructing scales using jnd's, he asked participants to directly estimate magnitudes of various stimuli using arbitrary numerical values. So an initial stimulus would be given a numerical value, and then later stimuli were given values relative to it, where all of the numerical assignments were freely chosen by the participants. He also asked them to directly estimate stimulus ratios, such as when one stimulus seemed to be twice as intense as another. Using these methods, he showed that the perceived intensity of some stimuli departed from Fechner's law. He concluded that Fechner's assumption that all jnd's are of equal size was to blame for the discrepancy and proposed as a replacement for Fechner's law the power law (now known as Stevens' law):

$$S = kI^a$$

where S and I are perceived magnitude and physical intensity, k is a constant, and a is an exponent that differs for various sensory modalities and perceivable quantities. The power law predicts that across all quantities and modalities, equal stimulus ratios correspond to equal sensory ratios, and, depending on the exponent, perceived magnitudes may increase more quickly or more slowly than the increase in stimulus intensity.

Stevens (1975, pp. 17–19) gave an elegant argument for why we should expect sensory systems in general to obey a power law. He noted that as we move around and sense the environment, the absolute magnitudes we perceive will vary: the visual angle subtended by the wall of a house changes as one approaches it; the intensity of speech sounds varies as one approaches or recedes. What is important in these cases is not the differences in the stimulus, but the constants, which are given by the ratios that the elements of the stimulus bear to one another. A power law is well suited to capture this, because equal ratios of stimulus intensity correspond to equal ratios of sensory magnitude.

Stevens' law provides a generally better fit for participants' judgments about magnitudes and therefore captures the phenomena of stimulus–sensation relations better than Fechner's law, although it, too, is only

approximate.[4] However, both laws provide the same sort of explanation for the relationship between the two: in each case, the laws show that these relationships are not arbitrary, but instead conform to a general formula, which can be expressed by a relatively simple equation. The laws explain the phenomena by showing how they can all be systematically related in a simple, unified fashion. Once we have the law in hand, we are in a position to make predictions about the relationship between unmeasured magnitudes, to the effect that they will probably conform to the regularity set out in the law (even if the precise form of the regularity requires empirically determining the values of k and a).

1.2.2 Case 2: Classical conditioning

Any organism that is to survive for long in an environment with potentially changing conditions needs some way of learning about the structure of events in its environment. Few creatures lead such simple lives that they can be born "knowing" all they will need to survive. The investigation of learning in animals (and later humans) started with the work of Pavlov, Skinner, Hull, and other behaviorists. Given their aversion to mentalistic talk, they tended to think of learning as a measurable change in the observable behavior of a creature in response to some physical stimulus or other. The simplest style of learning is classical (Pavlovian) conditioning. In classical conditioning, we begin with an organism that reliably produces a determinate type of response to a determinate type of stimulus – for example, flinching in response to a mild shock, or blinking in response to a puff of air. The stimulus here is called the unconditioned stimulus (US), and the response the unconditioned response (UR). In a typical experiment, the US is paired with a novel, neutral stimulus (e.g., a flash of light or a tone) for a training period; this is referred to as the conditioned stimulus (CS). After time, under the right training conditions, the CS becomes associated with the US, so that the CS is capable of producing the response by itself; when this occurs, it is called the conditioned response (CR).

There were a number of early attempts to formulate descriptions of how conditioning takes place (Bush & Mosteller, 1951; Hull, 1943). These descriptions take the form of learning rules that predict how the strength of

[4] For useful discussion on the history and logic of various psychophysical scaling procedures, see Shepard (1978) and Gescheider (1988).

associations among CS and US will change over time under different training regimes. One of the most well-known and best empirically validated learning rules was the "delta rule" presented by Rescorla and Wagner (1972). Formally, the rule says:[5]

$$\Delta A_{ij} = \alpha_i \beta_j (\lambda_j - \Sigma_i A_{ij})$$

To grasp what this means, suppose we are on training trial n, and we want to know what the associative strengths will be at the next stage $n + 1$. Let i stand for the CS and j stand for the US. Then A_{ij} is the strength of the association between i and j, and ΔA_{ij} is the change in the strength of that association as a result of training. The terms α_i and β_j are free parameters that determine the rate at which learning can take place involving the CS and US. The term λ_j is the maximum associative strength that the US can support. Finally, $\Sigma_i A_{ij}$ is the sum of the strength of all of the active CSs that are present during trial n. This is needed because some learning paradigms involve presenting multiple CSs at the same time during training.

The essence of the Rescorla–Wagner rule is to reduce the "surprisingness" of a US. If a CS (i) is not associated strongly with a US (j), then (assuming no other CSs are present), the parenthetical term of the rule will be large, and so the strength of the association between i and j will be correspondingly adjusted. Over time, as its association with the CS increases, the surprisingness of the US decreases, and so less change in strength takes place.

The Rescorla–Wagner rule is one of the most extensively studied learning rules in psychology, and it has some significant virtues: it unifies a large range of phenomena by bringing them under a single, relatively simple formal description; it explains previously discovered phenomena; and it generates surprising and often-confirmed predictions about new phenomena. To get the flavor of this, consider some of its successes: (1) The rule explains why acquisition curves show less change over time, for the reason given in the previous paragraph. (2) Extinction is the loss of response to a CS when it is presented without its paired US. The model explains this by positing that

[5] Gallistel (1990, Chapter 12) gives an excellent critical discussion of the assumptions underlying the R-W rule and its predecessors. He notes that the R-W rule is cast in terms of associative strengths rather than directly observable response probabilities, which represents a significant change of emphasis over earlier behavioristic rules. For a review of some important behavioral findings concerning conditioning, see Rescorla (1988).

during nonreinforced trials, the λ_j term goes to zero and the β_j term is lower than for acquisition, so association strengths will gradually decrease. (3) The rule explains the phenomenon of blocking, which occurs when CS A is paired in pretraining with US, followed by training in which the conjunction of CS A and CS B is paired with US. The result is that pretrained organisms show less association between B and US than do those that lack pretraining; in this case, B is said to be blocked by A. (4) The rule also explains overshadowing, which occurs when A and B are presented simultaneously. In this case, reinforcing AB results in B having less associative strength than if it were reinforced without A. Both overshadowing and blocking were significant challenges to earlier learning rules (Kamin, 1969). The rule's further empirical successes are too numerous to mention here, but see Miller, Barnet, and Grahame (1995) for more examples, as well as cases in which its predictions are not confirmed.

1.2.3 Case 3: Visual attention

Visual perception normally presents us with a world of separate and relatively enduring objects and events. The brown wooden chair appears separate from the black coat draped over it, and the gray cat appears separate from the orange couch across which she walks. But this division of the world into objects with determinate properties is not obviously given just by the incoming light array itself. It requires some mechanisms of processing and interpretation in order to be extracted. Based on an extensive series of experiments, Treisman (1988) proposed an influential cognitive model of how stable perceptions of objects and their properties are produced that give attention a central role.

In Treisman's model, visual processing takes place in a series of "layers." These layers represent the properties that can be represented by the visual system. The layers are internally organized like "maps" of the properties that the visual system can extract from the low-level signals passed on from the retina and other early stages of visual processing. One map simply encodes locations in visual space and records for each location whether a visual feature is present or absent there. This master location map does not, however, specify *what* features are at which locations – it encodes only locations, presences, and absences. A hierarchy of further maps encodes the possible features that can be detected in the visual scene. Color is one dimension along which objects can vary, so one map encodes possible color values an object may have

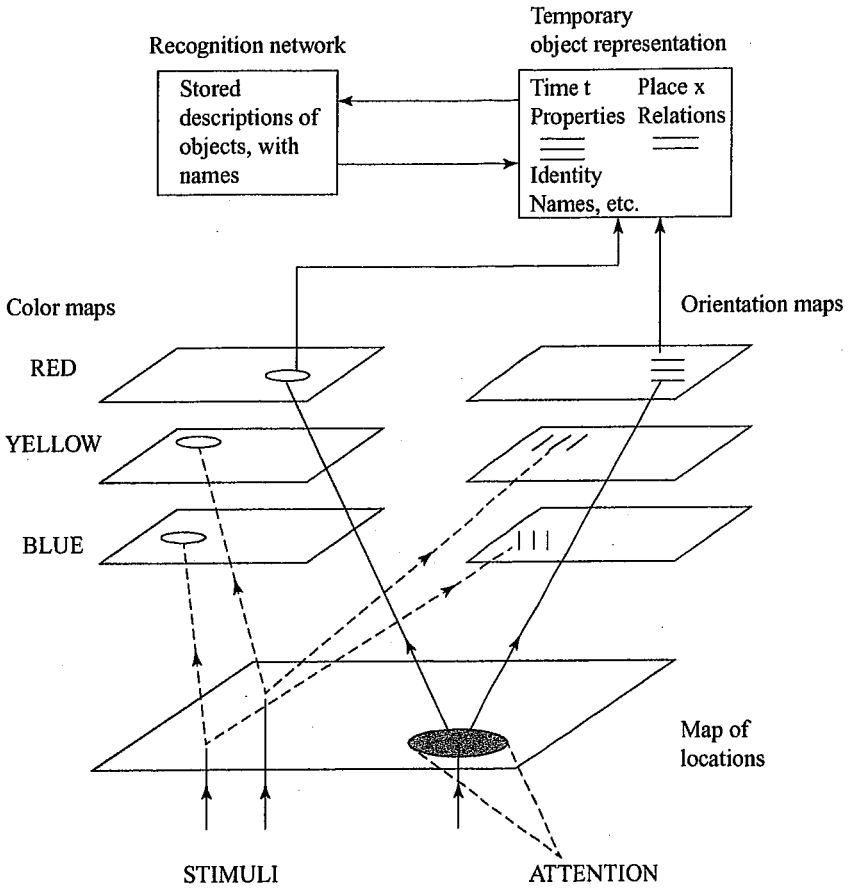

Figure 1.1 (from Treisman, 1988) A model of visual object perception and attention. Objects are represented by features located within maps. An object's location is encoded in a master spatial map, while the color and shape of the object's parts are encoded in separate maps. Attention integrates all of these features into an object file that is then used to identify the object based on what is stored in memory.

(red, orange, blue, etc.). Orientation is another (vertical, horizontal), as are size, motion, etc. The visual qualities that define how a perceived object is represented are distributed across this set of maps.

Attention is the "glue" that binds these separate features together into a unified perceptual representation of an object. When we focus our attention on a region of space, the features that are associated with that region at a time are jointly activated and packaged into an "object file" representation. This representation indicates that there is something at location l that has

features $f_1, f_2, \ldots, f_n$ at time t. Object files represent objects in terms of their visually detectable characteristics and can be passed on to further systems for elaboration and processing. For instance, an object file that describes a small black fuzzy object moving from left to right across the visual field might be classified as a *cat*, thus making available a stored body of information about cat behavior for the guidance of immediate action. Without attention, however, there is strictly speaking no perception of spatial locations as containing objects with determinate sets of properties.

This model explains a number of surprising phenomena: (1) Searching for objects defined by conjunctive features is a serial process, taking more time when there are more distractor items present; disjunctively defined objects, on the other hand, do not require more time to be located, even when the number of distractors increases. (2) Participants experience illusory conjunctions of features in conditions of divided attention, such as misperceiving a black "X" as a green "X" when they were adjacent to one another. (3) Participants could not reliably identify *conjunctions* of features (e.g., being a red "O" or a blue "X") without accurately finding their location, whereas they could reliably identify individual features even when they could not localize them accurately. (4) When attention is cued to a spatial location, identification of conjunctive targets is facilitated, whereas there is little effect on targets defined by a single feature; moreover, invalid location cues disproportionately affect conjunction targets. These and many other results are summarized in Treisman (1988). Taken together, they suggest that the underlying architecture of object perception depends on an attentional binding mechanism similar to the one Treisman outlines.

1.2.4 Case 4: Reading and dyslexia

Once we have achieved proficiency at reading, it seems phenomenologically simple and immediate, like other skilled performances. We perceive complex letter forms; group them into words; access their meaning, phonological characteristics, and syntactic properties; and then speak them aloud. But the cognitive substructures underlying this performance are complex. Evidence from acquired dyslexias (disorders of reading) has been especially important in providing insight into the structure of this underlying system.

Classifying the dyslexias themselves is no simple task, but a few basic categories are well established. In assessing patients' impairment, three major

stimulus categories are typically discussed: regular words (those whose pronunciation conforms to a set of rules mapping orthography onto phonology), irregular words (those whose pronunciation must be learned one word at a time), and pronounceable nonwords (strings that can be spoken aloud according to the phonological rules of the language). Surface dyslexia, initially described by Marshall and Newcombe (1973), involves selective impairment in reading irregular words ("sword," "island") versus regular words ("bed," "rest"). Many of these errors involve overregularization: "steak" might be pronounced as "steek," while "speak" would be pronounced normally. Pronounceable nonwords (e.g., "smeak," "datch") are also read normally – that is, as the rules mapping letters onto sounds would predict for normal speakers. McCarthy and Warrington (1986) present a case study of surface dyslexia, focusing on patient KT. He was able to read both regular words and nonwords, but failed to consistently pronounce irregular words correctly; the maximum accuracy he attained on high-frequency irregular words was 47%. Phonological dyslexia, on the other hand, involves selective impairment in reading pronounceable nonwords versus matched words. There is generally no difference between regular and irregular words. One such patient is WB, whose disorders were described by Funnell (1983). WB was unable to correctly pronounce any of 20 nonwords, but was able to pronounce correctly 85% of the 712 words he was presented with, which included nouns, adjectives, verbs, and functor words. Although there was some effect of frequency on his pronunciation of words, nonwords were unpronounceable even when they were simple monosyllables.

Surface and phonological dyslexia present a pattern of dissociations that suggest that normal reading is not a unitary cognitive faculty. In the former, there is impairment of irregular words relative to regular words and nonwords; in the latter, there is impairment of nonwords relative to regular and irregular words. Although these dissociations are rarely perfect – there is some preserved function in KT's case, and WB is somewhat impaired on infrequent words – they suggest that reading is explained by a set of connected systems that can be selectively impaired. The classic model to explain these dissociations is the dual-route model of reading presented first by Marshall and Newcombe (1973) and revised subsequently by Morton and Patterson (1980), Patterson and Morton (1985), and Coltheart, Curtis, Atkins, and Haller (1993). On this model, reading involves an initial stage of visual analysis, during which visual features are scanned and anything resembling a letter is extracted. These representations of letter strings may then be passed to two

distinct pathways. One pathway runs through the lexicon: the internal dictionary in which the semantic, syntactic, and phonological properties of words are stored. A representation of some letter string can access the lexicon only if it matches the visual form of some known word. If it does, then its meaning and phonological properties are retrieved, and the sound pattern it matches is passed to the articulation system, which generates the act of speaking the word.

Aside from this lexical pathway, there is also a pathway that does not involve the lexicon, but rather pronounces strings by applying a set of grapheme–phoneme correspondence (GPC) rules. These rules generate phonemic output for any string in accord with the normal pronunciation rules of the language. Hence they can produce appropriate output for regular words and pronounceable nonwords, but they cannot generate correct pronunciation for irregular strings. It is the presence of both the lexical and GPC routes to reading that give the dual-route model its name.

From the architecture of the model, it should be clear how surface and phonological dyslexia are to be explained – indeed, this is bound to be the case, because the model was developed and refined in part to account for those very phenomena. In surface dyslexia, the lexical reading route is damaged, but the GPC route is intact. This accounts for these patients' ability to read regular words and nonwords – the GPC route can produce the right output for both cases. It also explains the overregularization errors on irregular words, because the GPC route is only capable of presenting regular outputs to letter strings. Phonological dyslexia involves damage to the GPC route with a generally intact lexical route. Hence both regular and irregular words can be pronounced, so long as they have entries in the mental lexicon; however, nonwords that do not resemble any known words cannot be pronounced, because the rule system that could interpret them is unavailable.[6]

1.3 Laws and mechanisms

For several decades of the twentieth century, one conception of scientific explanation reigned virtually unchallenged. This was the idea that explanation in science essentially involves appealing to laws of nature.

[6] The dual-route model has been modified extensively over the years to account for other types of dyslexia, particularly "deep" dyslexia and nonsemantic dyslexia; there have also been single-route models that have attempted to capture the same phenomena (Seidenberg & McClelland, 1989).

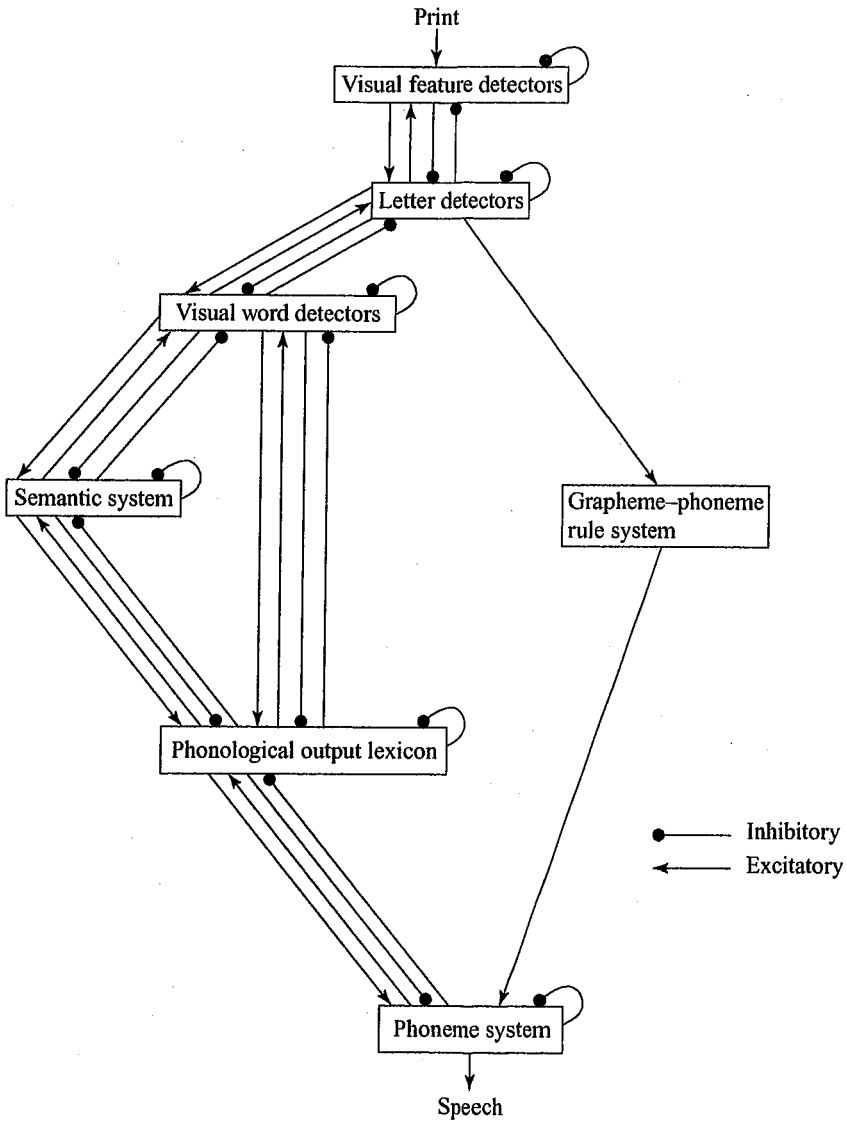

Figure 1.2 (from Coltheart, Curtis, Atkins, & Haller, 1993) An outline of the dual-route model of reading. The model takes visual words as inputs and produces spoken words as outputs. The main components are a lexical system for identifying whole words (left branch) and a nonlexical system for applying general rules of pronunciation to sound out words (right branch). The two routes explain why reading disorders are not all-or-nothing but instead have specific patterns of dissociation.

This conception is known as the covering-law (CL) view of explanation. The long history of this idea is not our present concern, but its outline is easy enough to sketch (see Hempel & Oppenheim, 1948; Hempel, 1965, for original statements of the view; and Salmon, 1989, for a review).

The CL view has three main components. First, scientific explanations are essentially deductively valid arguments in which the existence of the explanandum (the phenomenon to be explained) follows from the explanans (the things that do the explaining). So we explain a particular event such as a monkey's strengthened association between stimulus and response on a learning task, a neuron's firing an action potential, or the occurrence of a solar eclipse at a particular place and time by deducing the existence of that event from a set of premises. This captures the idea that there is a relationship between explanation and prediction: to explain a phenomenon is to have been in the position to predict its occurrence, given foreknowledge of the appropriate facts.

The second component is the requirement that among the premises there must be at least one law of nature. A law of nature is understood to be a generalization linking the occurrence of one event to the occurrence of another (or linking one property to another). To borrow Hempel and Oppenheim's (1948) example, how do we explain the fact that the mercury level in a glass thermometer will rise when it is immersed in boiling water? We can deduce that the mercury rises from the laws of heat conduction and thermal expansion, along with a description of the thermometer itself, the temperature of the water, and other antecedent conditions. These laws themselves describe how events of one sort (e.g., applications of heat to an object) lead to events of other sorts (conduction of heat throughout the object, expansion of the heated object, etc.). This illustrates how a particular phenomenon could have been predicted if only one had knowledge of the appropriate prior conditions and the laws governing how entities in those conditions behave in general.

The third component is that the statements included in the explanans must all be true. Given that explanations take the form of deductively valid arguments, this guarantees that the statement of the explanandum will also be true. The broader idea behind this component is that good explanations should not make essential use of false claims.

This pattern of explanation can be extended to general laws as well as particular events. Although Hempel and Oppenheim (1948) did not take this step, Nagel (1961) proposed that laws and theories at one level could be deduced

from – and hence, in his terms, *reduced to* – laws of a lower-level science. In his famous example, the Boyle–Charles law (that the temperature of an ideal gas is proportional to the product of its pressure and the volume of its container) can be deduced from the laws of statistical mechanics (relating pressure and volume to mean kinetic energy) plus a set of supplementary statements he called "bridge principles." Bridge principles (or bridge laws) relate the theoretical terms used in the laws at one level to those used in the laws at another level. These are required because the vocabulary of one theory typically contains terms not contained by that of another. So, in this case, if we add a bridge principle relating mean kinetic energy (a term used in statistical mechanics) to temperature (a term used in thermodynamics), we can deduce, with a few supplementary assumptions, that the Boyle–Charles law holds under certain specified boundary conditions. Hence we can see how at least some thermodynamic laws are explained in terms of underlying statistical mechanical laws.

Not every case in which one set of laws can be deduced from another is clearly a case of reduction: Galileo's laws of falling bodies can be deduced from Newton's laws of motion and gravitation in conjunction with the facts concerning the Earth's mass and radius, but this is less a case of reduction and more a case of showing these laws to be an instance of more general ones. But whether or not all such cases are reductive, Nagel's model shows that both particular and general phenomena can be explained by subsuming the explanandum under some law-involving explanans. We will return to the issue of reductionism in later chapters when we discuss the relationship between psychological and neuroscientific phenomena.

In fleshing out the CL view, we need to say something about what laws themselves are. This question has proven extremely recalcitrant, and philosophers have not converged on a common analysis. For our purposes, we will (to a first approximation) take laws to be true counterfactual supporting generalizations. Saying that laws support counterfactuals means that laws have modal force; they specify not just how things happen to be in the actual world, but also how they would have to be in relevantly similar worlds. So Coulomb's law states that the electrostatic force between two charged bodies is proportional to the product of their individual charges divided by the square of the distance between them. This law makes true other claims about how particular charged bodies *would* behave under different circumstances – for example, if the magnitude of their charges, or the distance between them, were

increased or decreased in various ways. Compare this to the true generalization that all the milk in Dan's refrigerator is past its drink-by date. This does not similarly entail the truth of the corresponding counterfactual: it isn't true that if this bottle of unexpired milk were in Dan's fridge now, it would be past its drink-by date. It is merely accidental that his milk is all expired, whereas it isn't accidental that charged bodies obey Coulomb's law. Although the distinction between accidental and so-called lawlike generalizations is difficult to draw precisely, some such distinction in terms of counterfactual force is presupposed by the account of laws we are assuming here.

Finally, laws may be either strict or hedged with ceteris paribus conditions. Strict laws are those that hold without exception: there is no case in which the antecedent of the law is satisfied but its consequent is not. Hedged laws, on the other hand, are those that obtain only under certain conditions – they have force, all things being equal, but they may have exceptions. Philosophers of science have typically supposed that strict laws, if there are any, are to be found only in basic physics, whereas the various special sciences that deal with nonbasic phenomena are more likely to contain ceteris paribus laws. As with the notion of law itself, explaining what it is for a law to hold ceteris paribus has proven extremely controversial. One prominent idea (due to Fodor, 1968) is that nonbasic laws are typically implemented by complex lower-level structures and processes. The law will hold only in those cases when these implementing structures are operating correctly or without interference – that is, when conditions are, in some open-ended and difficult-to-specify way, "normal." Others have replied that there are no true ceteris paribus laws and that, where we seem to have one, there is in fact a concealed strict law operating that just needs to be spelled out in further detail. We discuss the status of ceteris paribus laws further in the next section.

In recent years, an alternative view of explanation has been developed as a rival to the CL view. This new challenger is the mechanistic view of explanation developed by Bechtel (2008), Bechtel and Richardson (1993), Craver (2007; Machamer, Darden, & Craver, 2000), Glennan (1996, 2002), and Woodward (2002a).

The mechanistic view takes as its starting point the idea that in investigating many physical systems, especially biological systems, we are interested in explaining how they come to possess the capacities that they do, or how they are able to carry out the functions that they do. The lungs are responsible

for enabling us to breathe; what sort of physical facts about lungs makes them able to do this? The hippocampus is implicated in our ability to lay down new memories; what facts about its structure and function make this possible? Pyramidal cells and other neurons produce action potentials: how? These capacities may belong either to entire organisms, as when we ask how humans are able to perceive three-dimensional forms in space, or to their parts at many different levels of organization, as when we ask about how area V5 contributes to motion perception and how the hippocampus contributes to laying down new memories.

A mechanism can be thought of as an organized structure that executes some function or produces some phenomenon in virtue of containing a set of constituent parts or entities that are organized so that they interact with one another and carry out their characteristic operations and processes (Machamer, Darden, & Craver, 2000, p. 3; Woodward, 2002a, S375). This definition has several components. First, mechanisms are always mechanisms for something. There is some function they carry out, some characteristic effect they produce, or in general something that it is their purpose to do. Mechanisms, in Bechtel's (2008, p. 13) terms, are essentially tied to phenomena; therefore we can talk about the mechanisms of photosynthesis, episodic memory, action potentials, and so on. These can be schematized, following Craver (2007, p. 7), as "S Ψing," where S is some entity and "Ψ" is the exercise of some capacity by S, or some activity of S. Mechanisms may also simultaneously be mechanisms for the production of multiple phenomena.

Second, mechanisms are organized structures containing various constituent entities. These entities might be lipid bimembranes, various sorts of voltage-gated proteins, and masses of different types of ions, as in the case of the mechanisms responsible for producing action potentials in neurons. Or they might be larger structures, such as the divisions of the hippocampus into the dentate gyrus, CA1, CA3, the subiculum, and so on. Each of these parts constitutes a causally important part of the overall mechanism, and the mechanism itself depends on these parts being put together in precisely the right spatial, temporal, and causal sequence. Mechanisms are not just bags of parts – they are devices whose ability to carry out their function depends on the parts interacting in the right way. This might be a simple linear flow of control, as in an assembly-line model, or it might be more complex, involving cycles of activity, feedback loops, and more complex ways of modulating activity.

Third, the constituent parts of a mechanism are typically active. They themselves have causal roles to play in bringing about the activity of the mechanism as a whole. The proteins embedded in the cell membrane of a neuron are not passive entities in moving ions across the membrane; many of them play an active role in transport, such as the Na^+ channel, which rotates outward when the cell depolarizes, causing the channel to open and permit ions to flow outward. When the cell's potential reaches a further threshold value, a "ball-and-chain" structure swings into place, closing the channel. The active, organized operations of such component parts explain how the cell membrane produces the characteristic shape of the action potential. This raises a further important point about mechanisms: although some may be active only when their triggering conditions are met and are largely inert otherwise, others may be continuously endogenously active, integrating inputs from the outside into their ongoing operations, as in Bechtel's example of the fermentation cycle in yeast (2008, pp. 201–204).

Mechanistic explanation begins with a target explanandum or phenomenon – say, the ability of some entity to produce some function. These phenomena are explained by showing how the structural makeup of the entity in question enables it to carry out its function. This essentially involves displaying the causal sequence of events carried out by the component parts of the mechanism. It is characteristic of producing mechanistic explanations that one employs various heuristics to discover these components; two of the most important of these are localization and decomposition (Bechtel & Richardson, 1993). Decomposition can be either structural or functional: one might either figure out the natural parts of a mechanism (e.g., isolating the different parts of the cell through electron microscopy) or figure out the functional subcomponents that explain the mechanism's performance (e.g., figuring out the sequence of chemical transformations that must take place to produce a particular substance). Localization involves associating operations or functions with particular structures – for example, assigning the role of carrying out the citric acid cycle to the mitochondrion (Bechtel & Abrahamson, 2005, pp. 432–436).

The structure of mechanistic explanation differs from of CL explanation in a number of ways, of which we will note only two. First, mechanistic explanation is typically local, in the sense that it focuses on some phenomenon associated with a particular kind of entity. Only neurons produce action potentials, and the citric acid cycle takes place either in mitochondria (in eukaryotes) or

in the cytoplasm (in prokaryotes). The phenomena that are subject to mechanistic explanation are also typically both "fragile" and historically contingent. Fragility refers to the fact that mechanisms only operate with normal inputs and against a background of appropriate conditions. They are historically contingent in the sense that they are produced by processes such as natural selection that continuously adjust their components and performance. Genuine laws, according to some (e.g., Woodward, 2002a), are usually taken to have "wide scope" – they cover a range of different kinds of physical systems. Moreover, they hold independent of historical contingencies, at least in many paradigmatic cases such as Maxwell's electromagnetic laws. The norms of CL explanation have to do with discovering regularities that unify a maximal range of phenomena, whereas maximal coverage is not necessarily a norm of mechanistic explanation.

Second, the phenomena targeted by each explanatory strategy differ. The CL view in its classical formulation aims to explain the occurrence of particular events and can be extended to explain general regularities at higher levels. The canonical form of these explanations is that of a deductive argument. Mechanistic explanations aim to capture phenomena such as the fact that S can Ψ. They don't aim at explaining particular events per se, and they aim at explaining regularities only insofar as the explanations focus on particular systems, their capacities, and the effects they generate. The mechanistic view is also relatively unconcerned with prediction, at least of particular events. Many mechanistic systems may be so complex that it is difficult to predict how they will behave even under normal conditions.[7]

The CL view and the mechanistic view are hardly the only available perspectives on scientific explanation, but they are the two that have been most widely discussed in the context of psychology. With this background in place, then, we are finally ready to pose the question of which perspective best captures the norms of psychological explanation.

[7] However, fully characterizing a mechanism will involve being able to predict under what conditions its activity will be initiated or inhibited, and how its functioning is likely to be affected by "knocking out" various components – for example, producing a lesion in a certain brain region, or blocking neurotransmitter uptake at a certain synapse. This is not the same as predicting what the output of the mechanism will be, because this is already given by the description of the explanandum phenomenon itself.

1.4 Are there laws or mechanisms in psychology?

In keeping with the CL view's historical dominance in philosophy of science, many philosophers have argued that what makes psychology a science is just what makes *anything* a science, namely the fact that its core theoretical tenets are bodies of laws. Science aims to construct predictively and explanatorily adequate theories of the world, and what else are theories but sets of interlocking laws? So, for example, Jaegwon Kim (1993, p. 194) says:

> The question whether there are, or can be, psychological laws is one of considerable interest. If it can be shown that there can be no such laws, a nomothetic science of psychology will have been shown to be impossible. The qualifier 'nomothetic' is redundant: science is supposed to be nomothetic. Discovery, or at least pursuit, of laws is thought to be constitutive of the very nature of science so that where there are no laws there can be no science, and where we have reason to believe there are none we have no business pretending to be doing science.

The view could hardly be stated more boldly. Note that Kim himself claims to be doing no more than expressing accepted wisdom about science. In a similar vein, Jerry Fodor (1968) has influentially argued that theories in the special sciences (those outside of basic physics) are composed of bodies of laws, although these laws are autonomous in the sense that they do not reduce to the laws of any underlying science.[8]

Historically, many psychologists seem to have agreed with this perspective. Hence, from the early days of scientific psychology, we see attempts to state psychological laws explicitly. Our first two cases of psychological explanation, involving Fechner's and Stevens' laws and the Rescorla–Wagner learning rule, were chosen to illustrate this point. If they are true (perhaps within specific boundary conditions), then they could be used to explain the occurrence of particular psychological events. For instance, if there are laws relating the occurrence of sensations to later cognitive states, such as the formation of perceptual judgments or making of perceptual discriminations, then Stevens' law, in conjunction with such laws, would give

[8] Fodor in fact hedges this claim somewhat, making only the conditional claim that his view follows only if sciences consist of bodies of laws; he also refers to "the equally murky notions of *law* and *theory*."

rise to generalizations connecting stimuli and judgments. If this system of laws were sufficiently elaborate, it would amount to an outline of all possible causal paths through the cognitive system, from stimulus to behavior. Such is the form of an idealized psychological theory on the CL view.

But how commonly do such laws occur? Cory Wright (personal communication) has produced a Top 10 list of the most frequently cited "laws" in the psychological literature, along with the date they were first formulated. In descending order, the list runs:

1. Weber's Law (1834)
2. Stevens' Power Law (1957)
3. Matching Law (1961)
4. Thorndike's Law of Effect (1911)
5. Fechner's Law (1860)
6. Fitt's Law (1954)
7. Yerkes–Dodson Law (1908)
8. All-or-None Law (1871)
9. Emmert's Law (1881)
10. Bloch's Law (1885)

A notable fact about this list is that it peters out around the middle of the twentieth century. A further fact is that there is a distressing paucity of laws to be found. If science requires laws, psychology would appear to be rather infirm, even accounting for its relative youth.

Philosophers of psychology have sometimes proposed informal additions to this list. So, for example, we have the following: for any p and any q, if one believes p and believes that if p then q, then – barring confusion, distraction, and so on – one believes q (Churchland, 1981). This is supposed to be a predictive principle of folk psychology governing how people will form new beliefs. In a similar vein, we have the "belief–desire law": if one desires q and believes that if one does p, then q, then one will generally do p. Other candidates proposed by Fodor (1994, p. 3) include "that the Moon looks largest when it's on the horizon; that the Müller-Lyer figures are seen as differing in length; that all natural languages contain nouns."

Perhaps reflecting the fact that no psychologists seem to regard these as "laws," however, Fodor hedges and calls them "lawlike." This raises the possibility that psychologists *do* discover laws, but don't *call* them "laws." Indeed, although the psychological literature is law-poor, it is rich in what are called

"effects." Examples of effects are the already-mentioned Stroop effect; the McGurk effect, in which visual perception of speech affects auditory perception of phonemes; and the primacy and recency effects in short-term memory, in which the earliest and latest items in a serial recall task are retrieved more frequently than those in the middle. Cummins (2000) argues, however, that it is a mistake to think of these effects as being the elusive laws we seek. Although they are perfectly respectable true counterfactual-supporting generalizations, he claims that they are not laws, but rather explananda – that is, they are what we have been calling phenomena.[9]

The first point to make here is that whether something is a phenomenon is relative to a context of inquiry, namely a context in which it constitutes something to be explained. But in another context the same thing may itself serve to explain something else. Einstein explained the photoelectric effect, but we can also appeal to the effect in explaining why the hull of a spacecraft develops a positive charge when sunlight hits it. Cummins says of the latter cases that these do not involve the effect explaining anything, but instead just *are* the effect itself. But the photoelectric effect doesn't mention anything about spacecraft (or night-vision goggles, or image sensors, or solar cells, or any other devices that exploit the effect in their functioning); it's a general phenomenon involving the release of electrons following the absorption of photons. Further particular and general phenomena, as well as the operation of many mechanisms, can be explained perfectly well by appealing to the effect.

A second criticism of the idea that there are laws in psychology is that they are bound to be at best "laws *in situ*," that is, "laws that hold of a special kind of system because of its peculiar constitution and organization . . . Laws *in situ* specify effects – regular behavioral patterns characteristic of a specific kind of mechanism" (Cummins, 2000, p. 121). We have already noted that mechanistic explanation is inherently local; the idea here is that, given that psychology (like all other special sciences) deals with only a restricted range of entities and systems, it cannot be anything like laws as traditionally conceived, for traditional laws are wide-scope, not restricted in their application conditions.[10] Laws in situ, then, are not worthy of the name.

[9] This point is also made by Hacking (1983), who notes that the term "effect" also functions to pick out phenomena in physics.

[10] Note that if this criticism holds, it holds with equal force against other special sciences, such as geology, astrophysics, botany, and chemistry; anything that is not fundamental physics will be law-poor, if Cummins' argument goes through.

This is related to a criticism pressed by many against the very idea of ceteris paribus (cp) laws (Earman & Roberts, 1999; Earman, Roberts, & Smith, 2002; Schiffer, 1991). Recall that cp laws are nonstrict: the occurrence of their antecedents is not always nomically sufficient for the occurrence of their consequents. But if these laws are to be nonvacuous, we need some way of filling in these conditions to make them precise and testable; otherwise, we lose any predictive force they might have. And this we have no way of doing. In the case of folk psychological laws such as the belief–desire law, there are indefinitely many reasons why one may not act in a way that leads to getting what one desires. There are indefinitely many reasons why one may not believe even obvious consequences of things one already believes (perhaps it is too painful, or one simply isn't trying hard enough, or...). The same worries apply to the laws of scientific psychology: there may be indefinitely many stimulus conditions in which Stevens' law fails to hold, and saying that it holds except when it doesn't is profoundly unhelpful.

There are two possibilities when faced with this challenge. The first is to try to spell out substantive conditions on being a cp law that meet the normative standards of psychology; Fodor (1991) and Pietroski and Rey (1995) pursue this route. This normally involves saying what kinds of antecedent conditions are needed to "complete" a cp law by making its antecedent genuinely nomically sufficient. The second is to abandon the effort at stating such conditions and explain the status of special science laws in other terms. Woodward (2002b) takes this tack; we briefly sketch his approach here.

Although Woodward doubts that there is any way to fill in the conditions in cp laws to make them genuinely nomically sufficient, such statements may still be causally explanatory insofar as they express facts about what sorts of experimental manipulations – what he calls *interventions* – bring about certain sorts of effects. One paradigmatic sort of intervention in science is randomized trials, where we can directly compare the difference between the presence of one putative causal factor and that of another. If the presence of one factor leads to an effect with a greater frequency than the absence of that factor, if this difference is statistically significant and we have controlled so far as possible for all other factors, we can tentatively conclude that we may have located a causal generalization. There are also a host of quasi-experimental procedures for discovering such relationships, many of which are staples of psychological methodology. So Woodward says:

> It seems to me that if one wants to understand how generalizations that fall short of the standards for strict lawhood can nonetheless be tested and confirmed, it is far more profitable to focus directly on the rich literature on problems of causal inference in the special sciences that now exists, rather than on the problem of providing an account of the truth conditions for ceteris paribus laws. (2002b, p. 320)

He illustrates this point by reference to Hebbian learning (a process by which neural circuits strengthen their connections as a result of firing together). Although we may not know precisely what circuits obey this rule, or under what conditions they do so, we can still appeal to the generalization that neurons exhibit Hebbian learning because we can show that under certain interventions, neurons do strengthen their connections in the way that the generalization would predict (and fail to do so without such interventions). The same could be said for other putative cp laws in psychology. If one wants to refrain from using the term "law" to describe these statements, it will do just as well for our purposes to call them experimentally confirmable lawlike generalizations that back causal inferences, generate individual predictions, and support counterfactuals.

This brings us back to Cummins' point that psychological laws are only laws in situ. No one expects there to be (in his terms) a *Principia Psychologica* – an "axiomatic system, self-consciously imitating Euclidean geometry" (p. 121) from which all psychological phenomena could be derived. But even for historically contingent and mechanistically fragile systems such as living, cognizing things, there may be robust causal generalizations that can be discovered via systematic manipulations. Modest laws such as these are nonfundamental and hence not fully general, as the laws of physics presumably are. But they are the best contenders for laws in psychology.

All of this, though, is supposing that there really *are* laws in psychology, even of a modest sort. Another possibility is that psychological explanation just isn't law-based at all. This is the preferred interpretation of mechanists. On the strongest interpretation, this view claims that the canonical form of psychological explanation is to start with some function or capacity to be explained, then decompose the system into smaller interconnected functional subsystems that carry out the subsidiary operations required to carry out the larger function as a whole. These subsystems can be further decomposed in turn, and so on, until we reach a point where the mechanistic story

ultimately bottoms out – perhaps at the point where we have associated primitive psychological functions with neurobiological mechanisms. This "homuncular functionalist" approach has been championed by Cummins (1975), Dennett (1978), and Lycan (1981). It also seems to lie behind much classic "box and arrows" type modeling in psychology. In such diagrams, one finds boxes labeled with the functions they are supposed to carry out, and various lines connecting them to show the flow of information among subsystems, as well as relations of control, inhibition, and so on.

The dual-route model of normal reading discussed earlier provides a nice example of such functional decomposition. There are pathways for information flow, and boxes for carrying out functions such as visual analysis of images into graphemes, mapping of graphemes onto phonemes, word identification, and lexical retrieval. The precise sequence of operations carried out within each box is rarely explicitly specified, but the existence of distinct functional subsystems is attested to by the partial dissociations observed in lesion patients; this accords with the mechanist strategy of decomposition and localization. The same could be said of Treisman's model of visual attention. Although not "boxological," it does contain several separate constituents, namely the representations of space and various perceivable visual features, as well as a set of control structures, including the mechanisms of attention and binding, which produce representations of unified visual objects as their output. The existence and function of these parts are attested by the experiments describing how people behave in divided attention conditions, how they perceive unattended stimuli, and so on. Presumably both of these preliminary sketches could be filled out into more detailed mechanistic accounts; the elaborate "subway map" model of the macaque visual cortex developed by van Essen and DeYoe (1994) provides a guide to how complex such decompositions may become.

Often in psychology, especially where it interfaces with neuroscience, we do find mechanistic explanations, or sketches thereof. Computational models of cognitive functioning can be regarded as one species of mechanistic explanation, and computational modeling is a common tool in understanding precisely how a physical system might carry out a certain function (Polk & Seifert, 2002).

A final point on the relationship between laws and mechanisms: We've been discussing these two paradigms as if they were adversarial. A more optimistic proposal is that they are complementary. For one thing, psychological

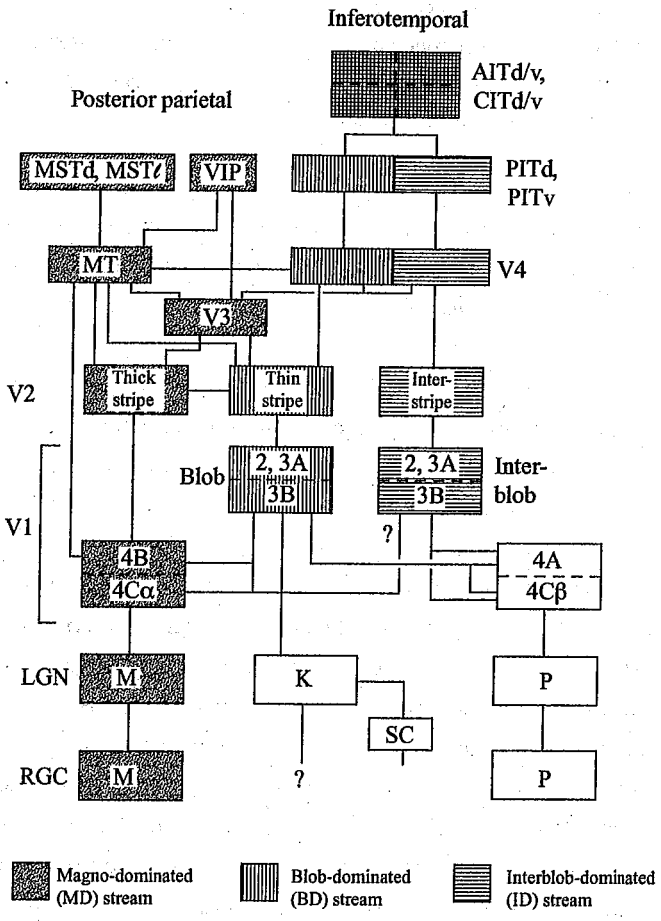

Figure 1.3 (from van Essen & DeYoe, 1994) Subcortical and cortical visual processing streams isolated in the brain of the macaque monkey. Subcortical processing is divided among the magnocellular (M), parvocellular (P), and koniocellular (K) streams, which originate from different populations of retinal ganglion cells. In the cortex, visual processing divides into portions dominated by magnocellular inputs and those that take inputs from so-called blob cells and interblob regions of V1. These inputs are then passed on to higher visual areas for further processing.

laws may entail the presence of corresponding mechanisms. This point has been emphasized by Fodor (1990), who says:

> Nonbasic laws *rely on* mediating mechanisms which they do not, however, *articulate* (sometimes because the mechanisms aren't known; sometimes because As can cause Bs in many different ways, so the same law has a variety

> of implementations). Ceteris paribus clauses can have the effect of existentially quantifying over these mechanisms so that 'As cause Bs ceteris paribus' can mean something like 'There exists an intervening mechanism such that when it's intact, As cause Bs.' (p. 155)

This point is correct and important. In nonfundamental sciences where there are laws, there are also mechanisms that implement them. Fundamental laws, by contrast, require no such mechanisms. So nomic explanations in psychology are backed by the promise of mechanistic explanations; certainly this must be true for psychophysical laws, which depend on the mechanisms embodied in our sensory systems, and similarly for laws governing learning, which may involve an array of more complex cognitive mechanisms.

Interestingly, mechanisms themselves may characteristically give rise to corresponding laws, in the modest sense of "law" employed here.[11] Consider: Where we have a mechanism, we have a structure that reliably produces a causal sequence running from its input (initiation conditions) to its output (termination conditions). There may also be effects of the normal functioning of the mechanism that are produced endogenously. So the normal visual system contains systems for producing representations of the relative size, color, distance, and so on, of perceived objects; however, these mechanisms, when functioning normally, also give rise to laws of vision that characterize how objects will be perceived under varying conditions, such as under changes in distance, illumination, or nearby contrasting objects. The Hering illusion provides a nice example – straight lines will reliably appear curved against a background of lines radiating from a central point. Both normal vision and visual illusions involve causal generalizations of this sort; indeed, reliable visual illusions can provide important hints about the rules the visual system follows in constructing representations of what is seen (Hoffman, 1998). Where these satisfy the conditions of being manipulable by interventions, we can regard them as stating rough causal laws that are subject to refinement by later experimental manipulations.

[11] Some mechanists, such as Glennan (1996, p. 52), have proposed that mechanisms themselves *rely* on causal laws to explain their operations. Woodward (2002a) has challenged this employment of the concept of a law, but it nevertheless seems true that the activities and operations of many mechanistic components are best accounted for in terms of lawlike generalizations: consider the role played by laws that describe the passive diffusion of ions, or chemical laws of bonding, in explaining the mechanisms of action potentials.

We are generally pluralists about psychological explanation. Some normatively adequate explanations may involve getting maximally precise descriptions of the causal generalizations that govern some part of the cognitive system. Ultimately, these causal connections will reveal the ways in which we can intervene on and manipulate a system in order to produce a particular outcome. Others may involve delving into the mechanisms that explain how we come to possess some capacity or other. These, in turn, will reveal how these causal levers function to bring about their effects. And there are almost certainly other explanatory strategies at work as well – for example, explaining how we come to have a capacity in etiological or evolutionary terms, rather than explaining how it functions at a particular time. Our present goal has just been to lay out some possible ways of interpreting and assessing research in psychology that aim not just at producing and refining phenomena, but also at explaining them.

1.5 Conclusions

Perhaps nothing could be closer to us than our own minds. But this intimacy does not always bring understanding. We are, in a sense, *too* close to our own minds to truly grasp their workings. Turning our attention inward, we may believe that we can trace the dynamic flow of our own thoughts and perceptions, pushed this way and that by the springs of desire and emotion, ultimately being channeled into actions. And we all become able to report our inner life in cogent sentences for others to consider, freezing and packaging these shifting experiences in a publicly inspectable form. We explain and justify our actions to ourselves and others, and these explanations often take the form of causal statements. At least some grasp of our minds' operations seems built into these practices of everyday life.

Yet, as we have seen, these too-familiar contours conceal surprises and mysteries. In the right sort of experimental situation, subjected to just the right carefully contrived conditions, the mind behaves in unexpected ways, revealing phenomena that are invisible both to introspection and to casual outward examination. Just as the physical and biological sciences needed to go beyond merely observing the natural world in order to uncover its deeper complexity, so understanding the mind requires intervention and systematic observation rather than just spectatorship. And making these phenomena clear is the first step toward explaining how they come about by uncovering the causal

regularities and mechanisms that produce them. Scientific psychology gives us a form of self-understanding that only becomes possible once we step backward and take this somewhat detached, objective stance toward ourselves. This is one of the paradoxes of psychology: that the understanding of our inner lives that it promises is available only once we begin to step outside of them.

An Introduction to the Philosophy of Psychology

DANIEL A. WEISKOPF and
FRED ADAMS

CAMBRIDGE
UNIVERSITY PRESS