Explanatory Models for Psychiatric Illness

Kenneth S. Kendler, M.D.

How can we best develop explanatory models for psychiatric disorders? Because causal factors have an impact on psychiatric illness both at micro levels and macro levels, both within and outside of the individual, and involving processes best understood from biological, psychological, and sociocultural perspectives, traditional models of science that strive for single broadly applicable explanatory laws are ill suited for our field. Such models are based on the incorrect assumption that psychiatric illnesses can be understood from a single perspective. A more appropriate scientific model for psychiatry emphasizes the understanding of mechanisms, an approach that fits naturally with a multicausal framework and provides a realistic paradigm for scientific progress, that is, understanding mechanisms through decomposition and reassembly. Simple subunits of complicated mechanisms can be usefully studied in isolation. Reassembling these constituent parts into a functioning whole, which is straightforward for simple additive mechanisms, will be far more challenging in psychiatry where causal networks contain multiple nonlinear interactions and causal loops. Our field has long struggled with the interrelationship between biological and psychological explanatory perspectives. Building from the seminal work of the neuronal modeler and philosopher David Marr, the author suggests that biology will implement but not replace psychology within our explanatory systems. The iterative process of interactions between biology and psychology needed to achieve this implementation will deepen our understanding of both classes of processes.

This essay addresses two fundamental questions about explanatory models for psychiatric disorders. In the first section, I propose a central role for multilevel mechanisms. I show how progress can be made using the approach of decomposition and reassembly despite the complexity and nonadditive nature of the etiological processes involved. In the second section, I address how to optimally interrelate biological and psychological explanatory perspectives. While the first section deals with relating parts to wholes in the context of intricate etiological mechanisms, the second part struggles with understanding the relationship between two distinct perspectives on the same basic phenomenon.

Levels of Explanation

First Principles

Rather than adopting a single explanatory perspective, as is often advocated in traditional theories of science, etiological models for psychiatric disorders need to be pluralistic or multilevel (1, 2). A range of compelling evidence indicates that these disorders involve causal processes that act both at micro levels and macro levels, that act within and outside of the individual, and that involve processes best understood from biological, psychological, and sociocultural perspectives.

Traditional models of science view the discovery of fundamental laws as the ultimate goal. That is, science should seek to explain vast details of the workings of our universe from a few basic principles, such as Newton's laws of motion and gravity. The traditional model of science sees explanation as emanating from such fundamental principles outward into the workings of the observed world.

Although deeply influential in 20th-century science, this model was developed from physics and does not easily apply to the biological and social sciences relevant to psychiatry. Indeed, a fundamental implication of this model of science, namely, that all real causal processes should be understood from one perspective and one set of laws, has been counterproductive in the field of mental health and has indirectly encouraged the rise of two perspectives that I argue have been counterproductive: “hard reductionism” ("all psychiatric illness is best explained solely in terms of molecular neuroscience") and “hard emergentism” (e.g., “all psychiatric illness is best explained solely in terms of specified mental or social mechanisms and cannot be deduced from biology"). Emergent properties of a system are properties that cannot be predicted from individual elements but arise only when the elements act together as a group.

A central tenet of this essay is that psychiatry should move away from this law-based model of science to one...
that focuses on mechanisms. Such a move—from laws to mechanisms as the fundamental explanatory goal of our science—produces a more coherent and practical conceptual framework. In particular, while the integration of biological, psychological, and social elements into causal processes was a tortured one using law-based models of science, it flows easily from a multilevel mechanistic approach. This approach to psychiatric illness is a conceptually rigorous descendant of a distinguished lineage of earlier integrationist accounts (e.g., references 3, 4).

What do I mean by a mechanistic approach? The differences between a mechanistic, reductionist, and emergent approach can be illustrated with a simplified story: Our task is to understand a home heating system. The hard reductionist identifies the thermostat and furnace as the fundamental features of the system—those that drive the basic processes. He takes them apart and explains their key working parts through well-understood engineering principles. That, he argues, “is all there is to it.” The hard emergentist graphs temperature fluctuations in the home—the key output variable—and discovers important circadian rhythms and predicts them with a complex statistical model. That, he concludes, “is all there is to it.” The mechanistically oriented researcher analyzes the home as a complex multilevel system including thermostat, furnace, ducts, heat gain and loss, degree of insulation, weather patterns, heat-generating appliances, and human activity. She develops a complex working model based on an understanding of the many component parts and their interactions. That, she summarizes, “provides the best explanation.”

In suggesting that we adopt a mechanistic approach to explanation in psychiatry, I mean that explanation requires the explication of causal mechanisms and the understanding of how those mechanisms are actually instantiated in the world. Our task is to clarify the mechanisms that underlie and have an impact on central mind/brain processes such as mood, perception, belief formation, and hedonic processes so that we can understand the causal processes whereby they become disordered in psychiatric illness.

Mechanistic models occupy a middle ground between the views of hard reduction and hard emergence. As Bechtel succinctly summarizes: “The decomposition required by mechanistic explanation is reductionist, but the recognition that parts and operations must be organized into an appropriate whole provides for a robust sense of a higher level of organization” (5, p. 130).

Understanding mechanisms requires a reductionist descent into the nitty-gritty of the world to figure out how things actually work. But in biological systems, events are always situated within contexts and causal processes are typically multilayered. Mechanistic explanations therefore require the integration of multiple organizational levels. For simple mechanisms—like your house thermostat—this can be a relatively easy job. More complex mechanisms can be much more challenging, but the basic principle still holds. An adequate scientific account must conceptualize how the whole thing works together, and this will always mean more than simply clarifying the operation of the lowest level of the mechanism. By “levels” I mean pieces of a mechanism that exhibit a part-whole (or component-system) relationship. Because molecules make up a membrane, neurons make up a circuit, higher-order neural systems (in a still mysterious way) make up an individual, and individuals make up a society, we can usefully talk of causal mechanisms with lower and higher levels. Ultimately we face the task of figuring out how the entire system works. But how do we get there?

**Decomposition**

This process typically begins with a “mechanistic sketch,” one that describes how the system might possibly work. The sketch is then filled in, characteristically in stages, to form an increasingly comprehensive mechanistic model. That is, science moves from “how possibly” to “how actually” descriptions of the mechanism (6).

The main workhorse in the traditional scientific approach to multilevel systems has been decomposition. First, the scientist tries to disassemble the system, breaking it down into its constituent parts (7, 8). Second, the scientist tries to understand each part in turn, starting with the simplest and working toward the most complex. Sometimes an additional cycle of decomposition is required, as components often have subcomponents. The final phase then is one of integration—the understanding of how all of the parts work together to produce the complex mechanism. Such an approach can be used to understand mechanisms as diverse as your car engine, the metabolism of a particular carbohydrate in a cell, and auditory perception in an owl.

The ease of decomposition varies in different kinds of systems. One conceptually important class of systems is “easily decomposable” (7). Such systems demonstrate “aggregativity” (9), meaning that their constituent parts interrelate with each other in a simple additive manner (10). Each module is “intersubstitutable,” meaning that it has a discrete intrinsic function that can be understood when separated from the entire system and studied on its own or when placed in different systems—that is, actions of the parts are not context dependent. Easily decomposable systems lack “causal loops.” Such loops can be either intralevel—when a component’s actions can alter its own properties or those of nearby elements—or interlevel. The most important kind of interlevel causal loop is where the output of the entire system feeds back to its own constituent components. This is called recursive, or top-down, causality.

**Alcohol Dependence as an Example**

To illustrate and concretize the principles outlined above, I here develop a much simplified explanatory sketch of a mechanistic understanding of alcohol dependence. I chose this syndrome because it well illustrates multilevel causal processes. The first task in understand-
ing a mechanism is to localize different components. Where do the causal pieces lie? Empirically supported risk factors for alcohol dependence occur on at least four broad levels: biological/genetic, psychological, social, and cultural/economic.

**Biological/genetic.** Risk for alcohol dependence may be altered by prenatal exposure to alcohol (11). Aggregate genetic factors strongly influence liability to alcohol dependence (12). We are learning that genetic factors can have an impact in many places on the pathway to alcohol dependence, ranging from specific effects on alcohol metabolism (13) and brain systems that interact directly and indirectly with ethanol (e.g., γ-aminobutyric acid [GABA], glutamate, and opioid systems) (14) to broader liabilities to the abuse of all forms of psychoactive drugs (15) and an even broader disposition toward externalizing behaviors (16, 17).

**Psychological.** A range of psychological constructs also affect risk for alcohol dependence, including several personality traits, such as neuroticism, impulsivity, and extraversion (18, 19), and several dimensions of alcohol expectancies (20, 21).

**Social.** As confirmed by twin studies (which show adolescent alcohol use to be strongly influenced by environmental factors (22, 23)), alcohol consumption and risk for alcohol dependence are robustly predicted by social factors, such as peer substance use, drug availability, and social class (18, 24).

**Cultural/economic.** Cultural, religious, and economic factors affect risk for alcohol dependence. Culture influences the forms of ethanol commonly consumed (25), the acceptability of public drunkenness (26), and the appropriateness of drinking by men versus women (which influences the vastly different ratios of alcohol dependence in men and women across cultures) (27). Rates of alcohol dependence often rise with the breakdown of traditional cultural beliefs and practices in migrant and native populations (19, 28). In the United States, religious beliefs influence both alcohol consumption and the risk for progression to alcoholism (24). Levels of taxation of alcoholic beverages and statutes controlling the sizes of alcoholic beverage containers permitted for sale both have an impact on the frequency of alcohol-related problems (29, 30).

Given these component parts, we must understand the nature of their interactions. How aggregative are they? It will be sufficient for our purposes to look only at that part of the picture involving genetic effects. At a biochemical level, interaction in risk for alcohol dependence has been seen for variants in genes at different stages of the ethanol metabolic pathway (31). Using twin designs, genetic effects on risk for drinking or alcohol dependence have been shown to vary as a function of religious beliefs (32), marital status (33, 34), and social environment (35, 36). Thus, the impact of genetic risk factors for alcohol dependence fails the additivity and “intersubstitution” assumptions. Their effects are dependent on both biochemical and psychosocial contexts.

Next, we examine the evidence for “causal loops” in the etiology of alcohol dependence. Genes strongly influence the initial response to ethanol (37). At one extreme, individuals with a variant of aldehyde dehydrogenase metabolize acetaldehyde so slowly that they develop a dysphoric flushing reaction after significant ethanol consumption (38). This genetic effect substantially reduces the chances that such individuals will repeatedly reexpose themselves to the large doses of ethanol needed to develop dependence (38, 39). At the other extreme, individuals who genetically have reduced sensitivity to ethanol’s effects are more likely to drink frequently and have an elevated risk of developing alcohol dependence (37, 40). So genes influence subjective ethanol effects, which influence alcohol expectations, which in turn loop out into the environment, influencing consumption patterns, which in turn affect risk of alcohol dependence.

Exposure to ethanol produces physiological tolerance both from increased metabolic rates and decreased CNS sensitivity (41). This can produce a positive feedback loop in which early phases of heavy drinking permit an individual to better “hold their liquor,” which in turn encourages yet greater consumption.

Impulsive, risk-taking adolescents seek out similar peers who provide support for and access to further antisocial and drug-taking behaviors (42, 43). Genetic factors influence this process (44, 45). So genetically influenced temperament causes individuals to select themselves into high-risk environments, which feed back on their risk for alcohol dependence by providing easy access to ethanol and encouragement for its excessive use. As one wag has put it, for us humans, who go out into the world to actively create our environments, our “brain has feet.”

Finally, at the “highest” level, a top-down causal loop is reflected in “reverse” (or “aversive”) cultural transmission for drinking behavior. The rate of abstention from alcohol is increased in the offspring of heavy-drinking parents (46, 47). Individuals at high genetic risk for alcohol dependence see the syndrome’s ravages in a parent, are aware of their own high risk, and consciously decide to abstain from ethanol consumption, thereby eliminating their risk for dependence.

**Examples of Nonaggregative Properties in Other Psychiatric Disorders**

Similar nonaggregative properties are common in the mechanisms that lead to other psychiatric disorders. For example, individuals differ in sensitivity to the pathogenic effects of adversity as a result of their prior experiences (48), genetic constitution (49–51), personality (52, 53), and social class (54). The impact of genes on liability to psychiatric illness varies as a function of environmental exposure (55) and other genetic loci (56). Depending
on a range of background factors, the same stressful experience can result in sensitization or habituation (57). That is, the impact of many risk factors for psychiatric illness is context dependent.

Disorders other than alcohol dependence also demonstrate robust causal loops, especially of the top-down variety. For example, an individual with high levels of the personality trait of neuroticism—strongly associated with risk for major depression (58)—is more prone to conflictual interpersonal relationships, reduced levels of social support, and increased rates of stressful life events, all of which increase risk for depression (59). The fearful child, after a single mildly traumatic experience with a neighborhood dog, avoids further contact with dogs, thereby preventing the habituation of the initial fear response. Feedback loops may also involve expectational sets. Anxiety-prone individuals selectively perceive danger (60), which in turn can increase symptoms. In a lovely illustration of top-down control, panic patients were randomized into groups who were told or were not told that they could control the level of inhaled CO2-enriched air. Although both groups received the same CO2 concentration, those with the illusion of control reported fewer and less intense panic symptoms (61).

Implications of Nondecomposability

What are the main lessons from this abbreviated explanatory sketch for alcohol dependence and brief review of relevant examples from other areas of psychopathology? First, hard reduction will not work because of the nonaggregativity and causal loops. The explanatory properties of these mechanisms are not reducible to any single level, molecular or otherwise. Second, although this nondecomposability greatly complicates our search to understand explanatory mechanisms in psychiatry, cynicism and pessimism are as premature and unwarranted as is zealous oversimplification. Bechtel documents how scientists, with care and persistence, have made major advances in understanding nondecomposable complex biological systems (5, 7, 62). Causal loops are not irrevocable barriers to detailed scientific understanding, as is well illustrated by the ability of early 20th-century biochemistry to clarify the citric acid cycle (62).

The initial phase of these successful approaches has always been to find subareas of local decomposability—relatively simple subsystems that could be profitably studied in isolation. This approach allows local causal processes to be clarified while ignoring other parts of the system. Despite the rising call for “translational research,” simple decomposition remains a critical first strategy toward approaching the etiology of psychiatric disorders. The naive emergentism that opines that the system is so complex and interrelated that we cannot possibly study any part in isolation is just plain wrong. But it is no less wrong than the equally misinformed idea—often professed by the hard reductionists—that all we have to do is study the parts in isolation and a detailed explanation will fall into place because the parts simply fit together. It will not and they do not.

Indeed, hard reductionists have typically argued that increased understanding will bring greater simplicity and reductive power—that the more we understand about the basic biology of psychiatric illness, the simpler and more potent will be our causal predictions. By contrast, I agree with Wimsatt when he writes, “The degree and kinds of emergence postulated of system proprieties should tend to increase with increasingly detailed specification of the internal structure and environmental relations of the systems in the model” (63, p. 287). That is, the more details we learn about the etiology of psychiatric disorders, the greater will be the number and importance of cross-modular interactions and causal loops. Etiological pathways for psychiatric disorders will be “deeply recursive,” moving many times between levels and forming what Wimsatt has evocatively called a “causal thicket” (64).

Scientists who brave this process—the stitching together of the initially disjoint subsystems—will likely experience what Craver has called “explanatory oscillation” (6) as they move iteratively back and forth across levels. As a model for what we hope to develop in psychiatry, we might consider how this integrative (or “stitching”) process has worked in the clarification of the mechanisms of memory where the phenominal and neural decompositions of memory were mutually informative and synergistically interactive over time (5).

The Nesting of Biological and Psychological Explanatory Perspectives

A complete picture of psychiatric illness must confront another even subtler dilemma faced by many other sciences (65)—how to integrate distinct perspectives on the same underlying process. For our field, the most prominent perspectival dilemma is that of how brain- versus mind-based approaches will interrelate in explanations of psychiatric disorders.

My approach to this question begins with the work of Marr (66), who proposed that a biologically complex information-processing system like the mammalian visual system can be realized (or understood) from three complementary perspectives. These three perspectives are as follows, with Marr’s original description of them in quotes and my rephrasing in italics (66, p. 25):

1. **Computational theory:** “What is the goal of the computation … and what is the logic of the strategy by which it can be carried out?” *What is the task this mind/brain system is designed to accomplish?*

2. **Representation and algorithm:** “How can this computational theory be implemented? In particular, what is the representation for the input and the output and what is the algorithm for the transformation?” *What functional processes are required to accomplish the task?*
3. **Hardware implementation**: “How can the representation and algorithm be realized physically?” *How are those processes actually implemented in brain "wetware"?*

The key feature of Marr’s approach—which provides a hierarchy of explanation—is that the biology (“hardware implementation” in his terminology) is understood in the context of functional explanations that articulate the goals of the system. Note that Marr’s implementation perspective focuses on the biological means by which a mechanism is executed while the representational and computational perspectives examine content. (To emphasize, Marr is proposing different perspectives on the same psychobiological process—here the mammalian visual system. This is in contrast to the first part of this essay, which examined different parts of a single broad mechanism.)

To try to build models of brain function from the bottom up, Marr suggests, is hopeless. If you took such an approach and began at the level of individual molecular processes as they occur within neurons and tried to work up from there to higher functions such as perception or motor behavior, let alone mood or “reality testing,” you would not be able to see the forest for the trees. Rather, you also need a top-down perspective, beginning with the task that the neural machinery was designed to execute.

Marr’s levels were designed for neural systems that process information. However, at a deeper level, his approach implements the old physiological distinction between understanding structure (hardware/brain) and function (processes/algorithms). These two complementary approaches have often been used iteratively in the scientific approach toward understanding complex biological systems and have, for example, been central to the development of cell biology (62).

Marr’s three perspectives could be used unaltered for guiding our approach toward understanding the mechanisms underlying certain psychiatric symptoms, such as auditory hallucinations or biased threat perception. Although many psychiatric problems are not currently amenable to an information-processing perspective, Marr’s underlying logic is nonetheless valid for psychiatry across the board. Obtaining a complete explanation of psychiatric disorders will require detailed understanding from a biological perspective. But this will not emerge from the bottom up—wherein biology would replace psychology—as predicted by the hard reductionists. Rather, it will happen by supplementing such strategies with top-down approaches, which allow biological explanations to be pursued and understood in the context of prior models articulated using psychological constructs.

Let me come at this key concept—that biological explanations need to be understood top-down in a context defined by mental constructs—from three additional, interrelated vantage points. First, both cognitive and evolutionary psychology advocate a “reverse-engineering” approach to brain/mind functioning (67, 68). The brain contains, this view suggests, many different neural sub-systems, all of which evolved to accomplish distinct tasks. Such tasks range from the relatively simple—such as maintaining an appropriate respiratory rate—to the extremely complex, such as the perception of meaning in human speech.

Imagine coming upon an old machine full of gears, sprockets, springs, and levers. What would be required to “explain” the workings of this machine? The concept of reverse-engineering provides a simple answer. To “explain” it, you must understand what its purpose is. Once you know the purpose, you can, with patience and ingenuity, “reverse-engineer” what the components of the machine are doing. So understanding the function of a machine forms the framework for an explanation of how it works. While the human brain is not a machine designed by a human, it is, functionally, a machine designed by evolution. So once we decide to try to understand what the brain is doing, we find that we cannot proceed without considering higher levels of analysis such as cognition, emotion, and perception.

How can we conceptualize what different parts of our brains are supposed to do? We can do that only in the language of function, which means using psychological constructs. Neurochemical terms will not work. So a reverse-engineering approach to understanding how brains make minds also suggests that biological explanations for the disturbed brain/mind systems that underlie psychiatric disorders have to be placed within the context of psychological processes.

The second perspective on this problem contrasts two different ways in which a more abstract (or higher-order) theory can relate to a more basic (or lower-order) theory. One possible form of that relationship is *replacement*, another term for hard-core reduction. Consider the physical concepts of temperature and mean kinetic energy. Once you know the mean kinetic energy of a gas, you learn nothing more by knowing its temperature. The higher-order construct (temperature) becomes redundant and is replaced by the more basic construct (mean kinetic energy). The second possible form of this relationship is *implementation*. Here, the more basic theory provides the mechanistic details of how the functions proposed by the more abstract theory actually get accomplished. In this case, the higher- and lower-order theories work together to provide a complete explanation. The lower-order theory does not replace the higher-order theory.

Which of these two relationships best describes how psychology and biology will interrelate in the explanation of psychiatric disorders? I argue that implementation is the more accurate depiction and fits well within the mechanistic framework here advocated. Following Gold and Stoljar (69), we can gain some confidence about the correct answer to this question by examining recent neurobiological research. For this purpose, they examine the work of Kandel on learning in *Aplysia* species (69) and conclude...
that his research program is best understood as an example of implementation, not replacement—that is, his work is best seen as “the fleshing out of a psychological story in neurobiological terms” (69, p. 822). Along the same lines, Hatfield (70) and Bechtel (5) have shown, respectively, that for three areas of visual perception (binocular single vision, stereopsis, and color vision) and for memory, neuroscientists have progressed by figuring out how the brain has implemented processes first worked out by psychologists.

Our third perspective on the relationship between biological explanations and mental constructs can be best illustrated with a hypothetical story: A research team shows definitively that gene X is associated with schizophrenia. The results are widely replicated. This research team then shows that gene X produces protein Y. A large definitive study shows that protein Y is abnormal in schizophrenia. This too is replicated, at which point they call a press conference to declare that they have “solved the riddle” of schizophrenia. Amid the triumphalist rhetoric, a young psychiatric resident raises her hand and asks, “But how does this abnormality in protein Y lead to the characteristic symptoms of schizophrenia—delusions, hallucinations, thought disorder, and negative symptoms?” The lead scientist, a bit stunned by the questions, replies, “We have no idea.”

A more comprehensive explanation of a psychiatric disorder must include an understanding of the production of the key symptoms and signs underlying that disorder. Parts of these explanations will have to be framed in psychological terms. Genes and molecules will surely prove to be critical causes of schizophrenia and thus will explain important things about the illness. But alone, they cannot explain it completely.

To say this in another way, psychology frames questions about how biological processes implement psychological functions. Moreover, as we understand the brain processes, we need to “back-translate” the biology into an understanding—in psychological terms—of the key psychopathological constructs under investigation (e.g., sad mood, drug craving, hallucinations, and compulsions). Merely showing a strong odds ratio between a particular genetic or molecular variant and illness is not enough. As Hatfield writes, “Researchers seek to understand the microactivities of neurons by asking how they contribute to one or another more global brain function, psychologically described” (70, p. 257).

Thus, eschewing hard reduction or hard emergence, we have the most to hope from a perspective in which biological explanations will sit within and implement “wetware” functions that are articulated in the language of psychology. This will not be a one-step procedure, as psychologists will not always “get it right.” An iterative relationship between psychology and biology—where initial psychological constructs are better defined and subdivided by initial biological findings, which in turn help clarify the biology—will be needed to reach a more complete understanding. In short, biological and psychological perspectives will coevolve.

These arguments do not imply that useful insights cannot come from the hard reductive or emergent perspectives. Indeed, effective therapies can be developed from basic biological research (such as associated genes) without having any idea of how the gene variant produces symptoms. Furthermore, important approaches to treatment, such as cognitive-behavioral therapies, have emerged from psychological constructs that contained no biology. However, these perspectives will leave us with only part of the picture.

**Summary**

A comprehensive etiological understanding of psychiatric disorders will require the integration of multiple explanatory perspectives. Law-based theories of science derived largely from physics, in which explanation arises from a few simple laws, poorly match the nature of the problems confronting psychiatry. Instead, I advocate a mechanistic approach—where the chief goal is to understand the mechanisms that derail the key mind/brain functions that are disordered in psychiatric illness. Simple methods of decomposition will not work, because the causal networks underlying psychiatric illness are not aggregate and contain multiple nonlinear interactions and causal loops. However, as in other areas of biology and neuroscience, progress can be made through the study of subsystems of local decomposability followed by the challenging task of integration.

Biology will not replace psychology within our explanatory systems. Rather we will slowly clarify, through progress in neuroscience, how the brain implements psychological functions. That iterative process will deepen our understanding of both biological and psychological processes.

**References**

2. Schaffner KF: Psychiatry and molecular biology: reductionistic approaches to schizophrenia, in Philosophical Perspectives on


61. Sanderson WC, Rapee RM, Barlow DH: The influence of an illusion of control on panic attacks induced via inhalation of 5.5% carbon dioxide-enriched air. Arch Gen Psychiatry 1989; 46:157–162
64. Wimsatt WC: The ontology of complex systems: levels of organization, perspectives, and causal thickets, in Biology and Society: Reflections on Methodology. Edited by Matthen M, Ware RX. Can J Philosophy 1994; 20(suppl):207–274