# Neuroscience, Explanation, and the Problem of Free Will

William T. Newsome

Department of Neurobiology and
Howard Hughes Medical Institute
Stanford University

Michael Gazzaniga, a leading brain scientist who essentially invented the modern field of cognitive neuroscience, tackles perennial (and increasingly urgent) issues of free will and moral responsibility and how modern neuroscience influences our conceptions of both. Gazzaniga states that his "goal will be to challenge the very concept of free will while leaving intact the idea of personal responsibility." Why? Because, he argues, free will is a scientifically outmoded concept, yet personal responsibility survives because it is defined at a higher phenomenal level—that of social systems—whose governing rules care not a whit whether actions of individuals are to any extent "free".

In developing his argument, Gazzaniga visits several important touchstones in the history of scientific ideas with a perspective that is often both illuminating and entertaining. Perplexingly, however, the overall line of argument seems to waver and change direction at a critical point concerning the causal status of high-level brain states that correspond to mental states such as beliefs, values, goals and memories. My objectives in this commentary are to identify points of consonance and dissonance between my own views and Gazzaniga's, and to offer perspectives that may facilitate a constructive resolution.

## Three principles: indeterminacy, levels of organization, and high-level causation

Gazzaniga and I agree on three basic principles for thinking carefully about the brain and how it mediates mental life. First, indeterminacy and unpredictability are built into the world and into cognition itself at fundamental level. Second, recourse to multiple levels of organization is essential for scientific explanation in neuroscience and other fields as well. Third, novel causal powers are inherent in highly organized systems, including the brain. Let's consider these in order.

Gazzaniga begins, appropriately, with the question of determinism. This issue seems to pose the greatest threat to the notion of human freedom, raising the specter, in Gazzaniga's words, "that we are all simply pawns in the brain's elaborate chess game." Gazzaniga begins his counter-argument by pointing out that the bottom-up, deterministic view of 19$^{th}$ century physics was shaken to its core by the quantum mechanical revolution of the 20$^{th}$ century, which introduced the doctrine of probabilism at the most fundamental level of physics. Quintessentially quantum mechanical events such as photon absorption in the eye or skin can affect whether we detect a predator (in dim light) or develop melanoma, thus

drastically altering the course of real-world, macroscopic events. A second revolution in the physical sciences—the advent of chaos theory—exposed hard limits to deterministic prediction of future events, even if the system in question remains deterministic in principle. And of course chaotic phenomena are by no means the only limit to predictability in physical systems. Gazzaniga might have noted that neuroscientists have long grappled with stochasticity in the timing of electrical "spikes" emitted by cortical neurons, forcing investigators to retreat to higher level constructs such as "average firing rate" to identify secure signaling mechanisms. He might also have noted that cognitive neuroscience is currently undergoing its own revolution in probabilism as it recognizes that Bayesian principles are incorporated into wide array of perceptual, decision-making and motor processes[1].

Even more surprisingly, perhaps, the brain seems to incorporate deliberately an element of randomness into its decision-making processes. The neural mechanisms that generate choices during resource acquisition, for example, seem to reflect an added "bonus" for probabilistic exploration of new environments and new alternatives, just in case the grass in fact turns out to be greener on the other side[2]. From an evolutionary point of view, occasional random choices perform the same creative function as occasional random mutations in the genome—they allow exploration of a much larger space of possibilities than would be encountered by simple deterministic processes.[3] Our world, including human cognition, is shot through and through with probabilism.

Gazzaniga next takes up levels of organization in nature. As matter becomes organized into systems of increasing complexity—especially in the living world—qualitatively new phenomena come into existence that cannot be understood with reference to lower levels of organization alone. For example, a predator-prey relationship is sensible only from a perspective that—at a minimum—presumes the existence of organisms, understands the necessity of energy acquisition for survival of organisms, and is informed about how particular species acquire energy resources. It is impossible to derive these conceptual resources from consideration of molecules alone. The important lesson is that scientific understanding of particular phenomena (predator-prey relationships) does not involve *elimination* or *replacement* of high-level entities and processes (organisms, energy extraction) with lower-level entities and processes (molecules, chemical reactions).

---

[1] D.C. Knill and W. Richards, Editors, Perception as Bayesian Inference. Cambridge University Press, 1996. M. Oaksford and N. Chater, Rationality in an Uncertain World: Essays on the Cognitive Science of Human Reasoning, Psychology Press, Ltd.: East Sussex, UK. 1998. A. Yuille and D. Kersten, "Vision as Bayesian Inference: Analysis by Synthesis?" *Trends in Cognitive Science*. 10:301-308, 2006. N. Chater and M. Oaksford, Editors, The Probabilistic Mind: Prospects for Bayesian Cognitive Science. Oxford University Press: Oxford, UK. 2008.

[2] G.S. Corrado, et al, "Linear-nonlinear-Poisson models of primate choice" *Journal of the Experimental Analysis of Behavior*, 84:581-617, 2005. B. Lau and P.W. Glimcher, "Dynamic response-by-response models of matching behavior in rhesus monkeys" *Journal of the Experimental Analysis of Behavior*, 84:555-579, 2005. N.D. Daw, et al., "Cortical substrates for exploratory decisions in humans" *Nature* 441:876-879, 2006.

[3] P.W. Glimcher. "Indeterminacy in brain and behavior" Annual Review of Psychology 26:25-56, 2005.

Rather, as Carl Craver argues incisively in his book, Explaining the Brain,[4] the power and beauty of reductionist neuroscience (and all reductionist biology, I believe) is to elucidate the physical mechanisms that *link* multiple phenomenal levels, which together comprise a unified whole.

This point hardly can be overemphasized in light of the steady stream of media stories in which neuroscientists announce that traditional explanatory constructs for human behavior such as beliefs, values, goals, and choices are in reality "nothing but" brain activations, neural circuit computations, collections of action potentials, neurochemical modulation, expression of genetic predispositions, or...(name your own favorite!). This "replacement" enterprise—which many neuroscientists seem to endorse—appears to be rooted in the traditional "covering law" model of scientific reduction which emerged from physics in the mid-20th century and asserts, roughly, that reduction is successful when high-level concepts or entities can be replaced by low-level entities, and all high-level laws and regularities can be derived from low-level laws[5].

Several weaknesses of the covering law model have become apparent over the last few decades of research in philosophy of science, but to my mind the most glaring is the end-game poverty of successful reduction under this model. As I write this commentary, I am sitting in a library full of books, computers and students working intensely on their various projects. If I were smart enough and had sufficient computational resources (both are pipe dreams!), I could in principle accomplish the deepest and most complete scientific explanation of the library by writing a quantum mechanical wave equation that describes (probabilistically) the motions of all atoms in the library for, say, the next 20 minutes. The problem, of course, is that *replacement* of my standard understanding of a library by an explanation expressed in a wave equation would leave me incredibly impoverished. I would know *absolutely nothing* about persons, ideas, and learning, to say nothing of books, computers, desks and chairs—there are no terms in the wave equation for such things![6] Neuroscientists should understand that this is the ultimate goal we affirm if we embrace the replacement (or eliminative) reductionist agenda. No one's favorite level of study enjoys special status (systems, circuits, cells, molecules, genes); all are destined to give way to the severe austerity of the wave equation. If this description of nature does not ring true, or at best seems partial and incomplete, then perhaps we should think harder about

---

[4] C.F. Craver. Explaining the Brain. Oxford University Press: Oxford, UK. 2007.

[5] *Ibid*., chapter XXX.

[6] Erwin Schroedinger, originator of the famous wave equation, was well aware of this poverty. From Nature and the Greeks, Cambridge University Press, 1954:

> "The scientific picture of the world around me is very deficient. It gives me a lot of factual information; puts all of our experience in a magnificently consistent order, but is ghastly silent about all and sundry that is really near to our heart, that really matters to us. It cannot tell a word about red and blue, bitter and sweet, physical pain and physical delight; feelings of delight and sorrow. It knows nothing of beautiful and ugly, good or bad, God and eternity. Science sometimes pretends to answer questions in these domains, but the answers are very often so silly that we are not inclined to take them seriously."

the real goals of our science and about the "nothing-buttery" that all-too-often infects our rhetoric, especially in the public domain. My own view is that Craver gets it right when he argues against a replacement model of neuroscientific understanding and for a "mosaic unity" that emerges from understanding the mechanisms that link levels of function within the nervous system.

The third principle that Gazzaniga and I share is that higher levels of organization (e.g. biological *systems*) possess causal efficacy that mere sums-of-parts do not. To my mind, this point is not even controversial. A lion has the ability to kill me; a pile of lion parts does not. A computer can perform a Fourier decomposition of a complex signal; a bucket of transistors cannot. The key ingredient that endows a collection of parts with causal efficacy is *organization*. Components, whether cells or transistors, acquire new power when they are organized into a mechanism that performs a function. The new causal power does not lie in the physics of the components, which typically does not change when the components are organized together; the secret is in the organization *per se* which exploits the physics to accomplish a functional goal. From this point of view, physics *constrains* but does not *determine* function. A set of chips organized into a computer cannot do anything that violates physics or Kirchhoff's circuit rules, but in the end, the behavior of the system is determined by *circuit design* (and there could have been many) not by physics alone.

The same principle applies to minds, seen as organized, high-level states of the nervous system. Like other organized systems, minds create possibilities that do not exist in their absence. Magellan's circumnavigation of the globe was dependent, in part, on a *belief* that the earth is round. My navigation to the grocery store is dependent, in part, on my *desire* to buy food. Thus, mental states, such as beliefs and desires, are critical actors in the causal story of behavior. Minds matter.

Gazzaniga and I seem to agree on these three principles—indeterminacy, multiple levels of organization, and causal efficacy of higher levels—which will frame the following discussion of freedom.

**Asking the right question about "freedom"**

I agree with Gazzaniga on two key points about what "freedom" is and is not. First, freedom does *not* imply an absence of causation. Increasingly, modern neuroscience is teaching us that our cognitive processes and mental experience are deeply rooted in the biology of the brain. Our beliefs, decisions, emotions and aspirations do not exist in a separate realm that somehow manages to communicate with the brain to instruct behavior; rather our mental states and processes emerge directly from the causal nexus of brain states and processes. I agree with Gazzaniga that we must abandon notions of "freedom" that imply independence of mental life from the brain.

But mind-brain independence, I think, is a poor way to define freedom in the first place. The critical issue is not whether the mind operates, in some sense, independently of the brain; the critical issue is whether high-level states of the brain that embody mental states

such as beliefs, decisions, emotions and aspirations play a causal role in the production of behavior. Thus Gazzaniga poses exactly the right question when he asks, "whether or not mental events like beliefs can be in the flow of events determining ultimate action." This question lies at the heart of any meaningful conception of free will or responsible action. I believe that my mental states have causes (I would be worried if they didn't!)—the key issue is *what counts as a cause*? For me, the essence of freedom is that my actions are caused, at least in part, by *my* beliefs, *my* values, *my* memories, *my* choices, *my* aspirations. When I act (or refrain from acting) because of outside coercion, I am not free. When my choices and actions are constrained in ways inimical to my core values because of the larger social structure I live in, I am not free. If I act from subterranean prejudices or fears that I am not even aware of, I am not free. If I am afflicted by a disease like Alzheimer's that robs me of my memory and my ability to acquire new data and reason about my beliefs, I am not free. I am most free when my behavior originates in those propositions I consider to be true about the world, and those values and aspirations that I have selected to guide my journey through the flux of events. I readily admit, of course, that much of my behavior is *not* free. I am subject to all of the negative qualifiers above and more (except, so far as I know, neurological disease)—this is simply part of the human condition. Importantly, however, "free" and "unfree" are not either/or conditions; most of the time our choices and actions lie somewhere along a scale between these poles, influenced to some extent by both. I consider personal growth and maturity to be a life-long effort to move from the "unfree" side of that scale toward the "free".

The reason that neuroscience is perceived in some quarters as so pernicious now becomes apparent: some interpretations of neuroscientific discoveries seem to undermine any basis for distinguishing between free and unfree choices, or responsible and irresponsible action. The most devastating message coming from certain neuroscience and psychology quarters is that our beliefs, values, memories, choices and aspirations are in fact illusory. The "news" is that such high-level explanatory constructs are epiphenomenal narratives that we tell ourselves; the real work of generating behavior occurs at a deeper level where neural gears grind according to a calculus that has little if anything to do with what we experience as beliefs, values and aspirations. In the end, when we have achieved a true scientific understanding of the mechanisms that produce behavior (i.e. a proper reduction of the psychological to the neural), our folk-psychological constructs can be tossed. These messages from neuroscience and psychology, if correct, abolish the "essence of freedom" outlined above since our beliefs, etc, are not causal; they are in fact illusory.

Puzzlingly, Gazzaniga seems to endorse this implication when he says:

> "The interpreter finds cause and builds our story, our sense of self. It asks, for example, 'Who is in charge?' and in the end concludes, 'Well, looks like I am.' It is an illusion, of course, but it appears to be how it works."

Gazzaniga's dramatic experimental observation of left hemisphere confabulation in split-brain patient PS, described in his chapter, is a very important result. At the very least, it illustrates our ability to weave fiction as well as fact into a narrative interpretation of ongoing events, especially in pathological conditions such the split brain. But the fact that

the "interpreter" sometimes confabulates to explain its own behavior does not mean that it confabulates all of the time or even most of the time. In Gazzaniga's own experiment, for example, PS's left hemisphere stated perfectly accurately why his right hand selected the picture of the chicken. In my own corner of neuroscience (sensory perception) visual illusions are studied intensely because they shed light on underlying mechanisms of normal vision. But neuroscientists do not infer from the existence of visual illusions that all of vision is illusory! My view is that our interpreters, like our visual systems, can be generally in touch with reality. The interpreter says that our actions are caused by mental states such as beliefs, desires, and choices, and the interpreter is generally right, consistent with the third principle above that Gazzaniga and I both accept.

The problem seems to be that, having assembled key intellectual resources—indeterminacy, the explanatory relevance of multiple levels of organization, and the causal efficacy of higher levels—Gazzaniga fails to capitalize on these gains, opting instead for the more familiar language of determinism. We are informed that "neuroscience is happy to accept that human behavior is the product of a determined system," that "beliefs and mental states stay part of our determined system", and that liberation from the negative influence of "determined brain states" on conceptions of personal responsibility is to be sought in the social realm, not within neuroscience. Gazzaniga may be right about this, and I have almost certainly failed to appreciate certain nuances of his argument. Nevertheless, it seems that we might acquire more insight, even within the neuroscience itself, if we scratch a bit harder at the problem.

**Multiple realizability and the limits to reduction: two examples**

We considered briefly in a previous section the key role of organization in creating complex, high-level entities that can possess novel causal powers. I now employ two simple, nonbiological examples—a musical tune and a computer program—to explore the essential role of organization and the limits of reduction, in the hope that lessons will emerge that are helpful in thinking about brain states.

Consider first the lovely melody line of Beethoven's bagatelle, Fur Elise, which has been performed and enjoyed countless times since its initial composition. How might we reduce Fur Elise to a more fundamental level scientifically? For any given performance we might analyze the exact pattern of sound waves in the air of the concert hall. But this is an imprecise reduction since the exact pattern of sound will differ from point to point within the same hall, to say nothing of differing from one concert venue to the next in different performances. To avoid this problem, one might resort instead to a description of the vibrations of the piano strings—the physical source of the sound waves—as each note in struck during the performance. But this certainly would not comprise a general reduction of Fur Elise since the melody can be played on many different instruments including some, like a clarinet, that have no strings at all. Furthermore, if Fur Elise is played in a different key, entirely different strings are struck even during a piano performance. Reducing Fur Elise to specific sound waves or physical means of production clearly doesn't work. At this point one might take a step back, regroup, and approach the problem from a different angle. We might refer instead to the sequence of notes inscribed on the pages of music that

guide the performer, irrespective of what instrument is used. This is certainly a more general description, but even here we must be careful. After all, those notes might be displayed on a computer screen instead of a piece of paper, or they might simply be stored in the memory of the musician. Even at moments when Fur Elise is not being performed anywhere in the world (or hummed in anyone's mind!), it continues to exist as organized particles of metal on computer disks, or organized bumps and grooves in pieces of vinyl. And of course millions of individual copies of Fur Elise exist on computer disks, vinyl records and paper sheet music around the world. If we want to reduce Fur Elise to a physical system involving the position and momentum of atoms, which of these is best?

This thought experiment points, of course, to the larger question of what, exactly, identifies Fur Elise as Fur Elise? What remains constant about Fur Elise across a truly vast number of physical instantiations in space and time? The answer seems to be that Fur Elise is best and most generally described at an abstract level—that of its high-level organization: Fur Elise is a sequence of notes produced in particular harmonic and temporal relationship with each other. This is *not* to say that Fur Elise is completely independent of physical instantiation; if all existing physical instantiations were lost simultaneously, including those in the memories of the world's musicians, Fur Elise itself would be lost forever. But while Fur Elise depends for its existence on *some* physical instantiation, it is a fool's errand to try to reduce Fur Elise to any *specific* physical instantiation. The proper level of understanding of Fur Elise is at the level of its own intrinsic organization.

While Fur Elise is an example of an organizational entity that resists reduction to a specific physical instantiation, one might legitimately ask about those "novel causal powers" that are supposed to arise with increasingly complex levels of organization. Does Fur Elise possess such powers? It seems so since a performance of Fur Elise can elicit emotions as well as memories of previous performances in human listeners.

Perhaps a more compelling example of novel causal power in an organizational entity is a computer program. As I write these words, Microsoft Word is transforming my keystrokes into readable English text. Helpfully, it also checks my spelling and occasionally stores the new text onto the hard disk in the (frequent!) event that I forget to do it myself. Word, like Fur Elise, is an organizational entity that defies reduction to any specific physical system of atoms and their motions. Considering my personal computer alone, Word has one instantiation on the hard disk where the executable file is stored, but it has had hundreds of *different* instantiations in my computer's memory as I power the machine up to work anew each day. What is constant about Word across all of my work sessions is not the exact identity of the transistors involved during any given session; what is constant is the *pattern of organization* of the interacting transistors (whichever ones they happen to be), and ultimately, the computational logic embedded in those interactions. Scale this problem up to the hundreds of millions of computers that run Word around the world, and the folly of reducing Word to a specific physical system of atoms and their motions becomes glaringly obvious. Again, this is not to say that Word is independent of physical instantiation; Word, like Fur Elise, would cease to exist if all its physical instantiations disappeared simultaneously. But Word, like Fur Elise, is defined by its intrinsic logical organization, not by any one of its many different physical instantiations.

In contrast to Fur Elise, however, Word certainly performs *work*. It has causal efficacy in the real world. It assists (and frustrates!) tens of millions of users around the globe daily. To purchase Word today, I would fork over $125 or so at my local electronics store, a tiny fraction of which covers the cost of the physical CD. What I am really buying—the actual *product*—is the organized information on the disk (what we call intellectual property) and the work the information can do for me once installed on my computer. After installation, in fact, I can throw away the CD although I'd best hang on to the serial number.

Philosophers, I have learned, have a useful name for this property of organizational entities—*multiple realizability*, meaning that the entity's organizational and functional logic can be "realized" in many different physical instantiations[7]. In analyzing a multiply realizable entity, the central goal is to sift through the details that vary from one instantiation to the next and discern the core organizational structure and logic that define the entity and are critical to its function. The variable details provide important clues to what is and is not essential to the core organizational logic of the system, enabling us to maintain focus on the critical level of analysis. Uncritical reduction, especially of the *eliminative* sort, will surely lead us astray because we will fail to recognize when our analyses have descended from core organizational principles into a morass of irrelevant detail, no matter how *accurate* that detail might be.

**What does all of this have to do with the brain?**

A core conviction of neuroscientists, ably communicated by Gazzaniga, is that mental states and their contents (e.g. a *belief* that the earth is round) are instantiated in the connections and activation states of highly organized neural circuits within the brain. As Gazzaniga also indicates, the circuits that instantiate any particular belief, decision, or goal are likely to be multiple, highly distributed, dynamic, and participating simultaneously in the instantiation of other mental states as well (i.e. signals are multiplexed in the brain). The neuroscientific evidence on these matters is still rudimentary, so we must frame such assertions as convictions rather than facts, but the evidence in their favor is mounting steadily from year to year.

Recall, now, that our primary agenda is to address Gazzaniga's key question: "are mental events like beliefs efficacious in determining ultimate action?" If the causal sources of our behavior lie, at least in part, in our beliefs, values, choices and aspirations, then the central requirements for free and responsible action are at hand. So how should we think about, for example, "beliefs" or "choices" from a neuroscientific point of view? The three principles developed above—probabilism, multiple levels of organization, and the causal efficacy of higher levels—should guide our thinking.

The crucial point that emerges is that high-level mental states such as beliefs are, like Fur

---

[7] See, for example, K. Aizawa and C. Gillette. "Levels, individual variation, and massive multiple realization in neurobiology." Chapter 22 in The Oxford Handbook of Philosophy and Neuroscience, John Bickle, Ed. Oxford University Press, 2009.

Elise and Word, organizational entities that are multiply realizable within the brain. The exact cells, synapses and ion channels that are active for any specific instantiation of my belief that "the earth is round" are likely to vary substantially from one occurrence of the belief to the next. The ultimate key to understanding a "belief", as instantiated in the nervous system, is to identify the large-scale organizational regularities (both spatial and temporal) that correspond to the belief, without becoming too distracted by the variable activity of the low-level components. It is this organizational structure of the neural system—not the details of any specific instantiation—that defines mental states and endow them with the causal power. An airplane can fly across the continent; a collection of airplane parts cannot. In the same manner, organized, high-level states of the nervous system create possibilities for the future that do not exist in their absence. Massive buildings are constructed in our cities because of beliefs about their function and likely appeal to customers. Wars are begun, in part, because of fear, ambition, and beliefs about the probable outcome of the conflict. Beliefs matter! They are essential components in the causal story of human behavior.

Although we are currently far from a scientific understanding of beliefs, striking evidence for the notion of multiple realizability and the centrality of high-level organization can be found in recent analyses of simple neural circuits like the stomatogastric ganglion of the lobster, in which wildly varying distributions of ion channel types can support the same emergent rhythm of the circuit[8]. These principles are also implicit in recent dynamical systems analyses of neural activity, whose central goal is to discern from the welter of single-neuron-level signals the core states of the larger system in which single neuron are embedded ("hidden" or "latent" states) and the dynamics that govern transitions between the states[9].

**Three advantages: prediction, manipulability, and parsimony**

It seems plausible, then, that we have the intellectual tools at hand to come to a meaningful understanding of freedom and responsibility. Our world is not rigidly determined, and high-level states of the nervous system that correspond to our beliefs, values and aspirations are both real and causally efficacious in determining future events. Does this view of human cognition and behavior have any merits other than its congeniality to freedom and responsibility? I believe that the answer is "yes", and that the advantages are of inherent importance to science: prediction, manipulability, and parsimony.

Considering humans to have real mental states (instantiated in brain states) with causal efficacy has overwhelming advantages for *predicting* the future. As I write these words, there exists a particular collection of atoms in this library room called "Bill Newsome". In principle, an observer possessing a valid quantum mechanical wave equation for this room

---

[8] E. Marder. "Variability, compensation and modulation in neurons and circuits." *Proceedings of the National Academy of Sciences*, 108:15542-15548, 2011.

[9] See, for example, J.S. Kelso, Dynamic Patterns: The Self-Organization of Brain and Behavior, MIT Press, 1995; K.V. Shenoy, et al., "A dynamical systems view of motor preparation: Implications for neural prosthetic system design," *Progress in Brain Research*, 192, 33-58, 2011.

could make probabilistic predictions about the motion of the "Newsome-system" of atoms over, say, the next 20 minutes. The probability of any particular future outcome would be extremely small, of course, since the number of possibilities multiplies ferociously with each passing time increment. Nevertheless, this is the best performance that fundamental physics can hope for, quantum uncertainty being what it is. A different observer, however, working not from the reduced view of quantum mechanics but from a high-level theory of the human mind and behavior, might base her predictions about my future movements on the calendar in my iPhone. That observer would predict, with a very high probability of being correct, that I will leave the library in 15 minutes due to a prior commitment to my wife!

But prediction *per se* is not necessarily good evidence of the validity of a scientific theory. As Aristotle famously pointed out, the crowing of the rooster predicts the sunrise, but is not a cause of the sun's rising. Evidence closer to the scientist's heart lies in *manipulability*, which undergirds empirical investigation in most scientific laboratories. If a scientist can elicit a change in the outcome of an experiment by manipulating a particular variable while holding others constant, we become more convinced that we have a grip on a causal mechanism at work in the experimental system. Now consider once again the Newsome-system in the library room. If we want to alter the timing of Newsome's departure from the library, would direct manipulation of Newsome's beliefs accomplish the goal? Yes. News that my wife wants me home 10 minutes earlier than originally planned would change my beliefs about the world and thus achieve the desired outcome (as would news that the library is on fire!). Criteria of manipulability, in addition to prediction, then, argue for the validity of minds as real, causal entities.

Because real biological systems are typically multilevel, it is sometimes possible to affect the outcome of an experiment by intervening at different levels. To change Newsome's beliefs about when he needs to leave the library, we might in principle attempt a low-level manipulation—independent modulation of all 50 million (or so!) neurons in the brain that collectively instantiate Newsome's belief that he needs to leave in 15 minutes. On the other hand, our observer with a good theory of the human mind can achieve the same result with a single manipulation rather than 50 million—by telling Newsome that his wife needs him home 10 minutes earlier than originally anticipated. The two manipulations can in principle achieve the same outcome, demonstrating that causal efficacy resides at both high and low levels of the system. But in this particular example, the high-level manipulation gets at the causal levers of the system—the level of its own intrinsic organization and function—far more directly and *parsimoniously* than the low-level manipulation.

It is sometimes argued that our high-level descriptions and explanations (of organisms and other systems as well) are a *practical* necessity for humans to get along in a very complex world, but we should all understand that these high-level constructs do not describe real entities with real causal efficacy in the world. That exalted status is the exclusive domain of atoms and the fundamental forces of physics. But why should anyone buy into such a claim when the advantages of good high-level explanations—prediction, manipulability and parsimony—are manifestly those most valued by scientists in the first place? Beats the hell out of me!

**Summary**

Humans have free will, and thus responsibility, to the extent that our behavior and the choices we make are driven by our own beliefs, values, and aspirations, and to the extent that we are able to critically evaluate and modify our existing beliefs and values in light of new data derived from interaction with the world. According to this view, the key issue for free will is not whether our actions have causes (they do!), but rather what the causes are. The critical question is whether our beliefs, values and aspirations—the stuff of traditional notions of personhood, agency and freedom—are real entities with real causal efficacy in the world, or whether they are illusory constructs that we make up to describe our experience of a world whose causal determinants lie at much more fundamental level.

Many neuroscientists appear to subscribe to the latter point of view, leading to skepticism about our own ability to control our actions and effect change in the world. This conviction seems to be driven by a reductionist methodology (and ideology!) that is *eliminative* in the sense that it seeks to replace high-level constructs and processes with lower level explanations where *fundamental* truth is thought to lie. In contrast, I argue that mental states and processes, like many other complex processes in our world, are organizational entities instantiated in high-level neural systems within the brain, which resist explanation through eliminative reduction. Understanding organizational entities and processes *requires* engagement at multiple phenomenal levels and elucidation of the mechanisms that link phenomena at different levels. Causal relevance and efficacy are distributed across multiple levels, as we saw earlier when considering how to get Newsome to move from the library sooner than originally planned.

A more important real-world example of this is the recent finding that acute depression is more effectively treated by a *combination* of cognitive-behavioral therapy and antidepressant drugs than by either alone[10]. Cognitive-behavioral therapy is a quintessentially high-level intervention in which the explicit goal is to change the patient's belief structures and modes of interacting with the world. Pharmacological treatment, on the other hand, is a quintessentially low-level intervention in which the explicit goal is to manipulate the synaptic concentration of the neurotransmitter, serotonin. Both work, again telling us that causal efficacy is distributed across multiple levels of the system. The two work better together than either alone, telling us that we ignore multi-level explanation and causal efficacy at the peril of our patients and loved ones!

Michael Gazzaniga and I seem to agree on much of this material. We part ways significantly, perhaps, at three points: Gazzaniga seems convinced that, 1) a deterministic

---

[10] M.B. Keller, et al. "A comparison of nefazidone, the cognitive behavioral-analysis system of psychotherapy, and their combination for the treatment of chronic depression." *New England Journal of Medicine*, 342:1462-1470, 2000.

J. March, et al. "Fluoxetine, cognitive-behavioral therapy, and their combination for adolescents with depression." *Journal of the American Medical Association*, 292:807-820, 2004.

framework is most appropriate for neuroscientific explanations of cognition and behavior, 2) the subjective, high-level experience of personal control over ongoing events is frequently (mostly?) illusory, and 3) our notions of personal responsibility are unaffected by the non-existence of freedom in the traditional sense because responsibility is defined at a social level, not at a neuroscientific level. In contrast, 1) I place more emphasis on the increasing role of probabilistic accounts in cognitive science and neuroscience, 2) I suspect that high-level mental states and their causal efficacy, as understood by the "interpreter," are frequently (mostly?) accurate, and 3) I believe that a positive reinterpretation of "freedom" can be facilitated by a proper understanding of multi-level relations and the limitations of reductive analysis *within neuroscience itself.*

I would like to know whether these apparent differences in the high-level mental states of Gazzaniga and Newsome are real or illusory!