PHILOSOPHICAL PSYCHOLOGY, VOL. 16, NO. 2, 2003

Routledge
Taylor & Francis Group

Neural holism and free will

DANIEL A. LEVY

ABSTRACT *Both libertarian and compatibilist approaches have been unsuccessful in providing an acceptable account of free will. Recent developments in cognitive neuroscience, including the connectionist theory of mind and empirical findings regarding modularity and integration of brain functions, provide the basis for a new approach: neural holism. This approach locates free will in fully integrated behavior in which all of a person's beliefs and desires, implicitly represented in the brain, automatically contribute to an act. Deliberation, the experience of volition, and cognitive and behavioral shortcomings are easily understood under this model. Assigning moral praise and blame, often seen as grounded in the notion that a person has the ability to have done otherwise, will be shown to reflect instead important aspects of signaling in social interactions. Thus, important aspects of the traditional notion of free will can be accounted for within the proposed model, which has interesting implications for lifelong cognitive development.*

1. Introduction

Consider the following situation: one late fall afternoon, Professor Smith, incumbent in the chair of Moral Philosophy at No Souls College, is busily engaged in the composition of her current article. Suddenly, she finds herself considering the possibility of adjourning to the kitchen for a nice cup of hot cocoa. In the end, she demurs, deciding to press on with a few more paragraphs. Or does she?

In fact, there was no chance whatsoever that she would have taken a cocoa break at that moment. Unbeknownst to her, a tendency towards diligence, implanted in her as a small girl by an extensive series of punishments and reinforcements administered by loving parents and admiring teachers, makes it impossible that she yield to temptation under the conditions described. That tendency keeps her chained to her desk by bonds stronger than adamantine—only to release her a few minutes later when she feels a headache coming on, and thinks that it would be prudent to drink something nourishing so that she will be able to continue working.

This is a trivial, but representative, example of human behavior in which free will is thought to play a part. No matters of great moral moment are involved in this case, yet the putative decision-making process is articulated here just as thoroughly as when we are called upon to decide whether to sacrifice our lives to save that of our own child—or that of a complete stranger.

Daniel A. Levy, Department of Psychology, Hebrew University of Jerusalem, Mt. Scopus, Jerusalem 91905 ISRAEL, email: dlevyisi@yahoo.com

ISSN 0951-5089/print/ISSN 1465-394X/online/03/020205–24 © 2003 Taylor & Francis Ltd
DOI: 10.1080/09515080320103337

We all live with the intuitive sense that we make decisions between available options, and live self-directed lives. Yet such a view is in serious conflict with many other truths about the universe that we hold to be fundamentally true. Whether we are convinced of the grim reality of determinist causality, or believe that our world is fraught with ruthless indeterminism, little room is left for free will in the sense of having control of our actions.

It is my goal in this essay to present a different way of thinking about what it means to have free will. This presentation endorses determinism as the basic description and grounds of explanation of all physical processes [1], including biological processes in general and brain processes in particular, and accepts that determinism is incompatible with the notion that we ever have the ability to act otherwise than we do (which many free will theorists consider to be the necessary condition or minimal supposition of human free will). At first glance, it would seem that embracing these two principles renders free will an incoherent idea. Nevertheless, I hope to show how it is possible to accommodate many of our intuitions about free will within the proposed explanation, and even to go one better—to offer a program for the maximization of free will.

The main thrust of this essay will be in the direction of the ontological status of free will, not its moral implications. The question of free will is often phrased in terms of moral responsibility: how can we hold somebody accountable for something if they could not have done otherwise? This query sometimes takes on a desperate air: "We must find a way to ground moral responsibility by an account of free will, or else how will I know whom to blame or praise!" For me (and for many, I suspect), this is a secondary issue. The important questions are first-person ones: Am I a free agent? Is my experience of volition illusory? What am I to make of my subjective experiences of confronting situations and making choices?

Of course, attempts at answering such questions are not lacking in the literature. In this presentation I would like to propose a different approach to this classic problem, one based on some recent insights into the way our minds (and brains) work, derived from clinical, laboratory and theoretical studies of cognition—the approach of *cognitive neuroscience*. This enterprise is, I believe, of great philosophical relevance; it is an attempt to identify and understand scientifically the factors which produce our behaviors, color our perceptions, evoke our emotions, shape our personal memories, and determine that of which we are aware—in short: all those factors that make us whom we are as persons.

The following argument can be summarized thus: all of our beliefs and desires are either completely determined by antecedent physical conditions or the function of random developments, so the traditional notion of free will is indeed fundamentally incoherent. Rather, the human will is best understood as the integrated sum total of all a person's beliefs and desires, expressed in the innate or acquired mental structures for implementing those desires in the universe. Such a unitary will is conceivable only under a connectionist model of mind, which can support a real-time integration of all such factors in producing human action. Freedom of the will is to be sought not in independence from the influence of prior propensities and experiences, but in the practiced achievement of totally harmonized behavior

through the learned suppression of behavioral modularity—the attainment of what I will call *neural holism*. Phenomena such as deliberation, the experience of volition, and the frustrations we experience as a result of our cognitive and behavioral shortcomings are most easily understood under this model. The practice of assigning moral praise and blame, which is seen by some theorists as grounded in the notion that a person has the ability to have done otherwise than he or she did, will be shown to reflect instead important aspects of signaling in social interactions. Thus, I will claim that all of the important aspects of the traditional notion of free will can be accounted for within the proposed model, which takes brain realities seriously as a constraint on models of human thought and experience. Additionally, this model provides the basis for a program in which people can maximize their free will through self-knowledge.

2. The implications of cognitive neuroscience and connectionism for a theory of mind

Let me begin by introducing the cognitive paradigm within which I will be presenting my approach to the issue of free will. It is based on the model of mind called connectionism [2], and on some important recent findings of cognitive neuroscience research.

The last half-century has presented psychologists and biologists with a veritable explosion of knowledge about the mind and the brain. As in many scientific fields, however, theory has not kept up with the data. While fascinating facts about thought and behavior (both human and animal) continue to be amassed faster than any team of researchers can assimilate them, integrative models to account for these findings lag far behind. Cognitive scientists striving for systematic accounts of processes such as perception, memory, language, reasoning, emotion, and behavior have generally offered models of mind based on information processing frameworks. Charts and diagrams describing the flow of information invariably include "black boxes" such as "central executive processors," "attentional modulators," or "speech part analyzers," to which a host of cognitive competencies are attributed. However, their authors are generally unable to give an account of how the brain is supposed to accomplish the computational tasks imputed to it by the labels in those diagrams. Quite a few cognitive psychologists cheerfully maintain that their responsibility is only to provide an account of the processes in question, without reference to the brain substrate in which they are instantiated; those processes could equally be executed by any physical system, be it a brain, a computer, or a Chinese room. This approach culminated in what is known as the *representational-computational theory of mind* (Fodor & Pylyshyn, 1988), which, as its name implies, explains cognition as logical, rule-based computations performed on mental representations.

The shortcomings of this rarefied approach to cognitive processes have rapidly become apparent. For one thing, it simply could not account for the flexibility of human thought, especially of linguistic competence (which was its own home turf to begin with) [3]. Second, it could not explain realities such as the time-course of mental activity, or the brain's execution of basic tasks such as seeing and walking.

Third, many cognitive scientists felt a profound unease with a theory of mind that ignored the workings of the brain. Biological research began to provide valuable insights into how things like memory and learning could be understood on the neural level (Kandel & Pittenger, 1999). It was time to unpack the black boxes.

An alternative emerged in the form of connectionism. This approach offers an understanding of all mental processes in terms of stochastically governed neural dynamics (Churchland & Sejnowski, 1992; McClelland *et al.*, 1995; see also Goldblum, 2001). That is to say: the way our minds/brains operate is best understood not in terms of logical formalisms or syntactic processes, but in terms of the properties of patterns of activity of our neurons. These patterns are a function of the relative strengths of the connections between the neurons (and so the term "connectionism"). Importantly, these patterns are radically different in character from the semantic–logical descriptions that have hitherto been used by both laymen and scientists to describe cognition and behavior.

A key point in the connectionist model of the mind is that the entire neural ensemble comprising the brain is simultaneously the representation of information and the active system that uses it. Knowledge is implicit in the connectional structure of the brain rather than explicit in the states of its individual units (Rumelhart, 1990).

Another recent trend in the neurosciences is the growing awareness of the extent of interconnectivity in the brain. This interconnectivity is revealed both within and between functional systems. An example of within-system interconnectivity is provided by vision, the perceptual system to which the largest amount of brain is dedicated in humans and primates. It has also been the most extensively studied of the senses. Such studies repeatedly confirm the preponderance of two-way neural connections (feed-forward and feedback) between the majority of brain areas and systems implicated in visual perception (Olshausen *et al.*, 1993). Connectivity across cognitive systems was less well documented, and the idea of modularity (Fodor, 1983) has been an influential concept in cognitive psychology. It accords well with neuropsychological findings regarding the dissociations between various preserved and lost mental abilities following brain damage [4]. However, recent findings that early perceptual processes are modulated by non-modality-specific structures governing attention (Sheinberg & Logothetis, 1997; Knight, 1997), as well as those which seem to play a dominant role in controlling affect (LeDoux, 1996), make it unlikely that any higher-level brain system will meet Fodor's (1983) informational encapsulation criterion of modularity (Posner & Fan, in press). More and more, a picture emerges of a massively distributed process of cognition, in which almost every part of the brain can potentially influence not only the immediate activity but also the response profile of every other area of the brain, directly or indirectly.

Another recent series of findings relates to the relationship between perception and memory. The current regnant view in neuroscience is that while there are brain areas (such as the hippocampus, mamillary bodies, entorhinal cortex, etc.) that play important roles in the consolidation and retrieval of memory, the actual memory representations are located in the same brain areas involved in the initial perception

and acquisition of that information. Support for this contention comes from findings regarding the physiological activation of perceptual brain areas during memory retrieval tasks (Fletcher & Tyler, 2002).

The implications of these findings are far-reaching. First, under connectionist models, and in light of the findings about memory–perception integration, *there is no separation in the mind between process and content*. Gone is the notion of a mental essence independent of information about the universe. Our entire mind–brain is an integrated whole, and its principles of operation are physical (summing excitations and adjusting synaptic weightings) rather than formal and logical. One of the original characterizing aspects of this approach was its claim that mental activity is based on *parallel distributed processing* in the brain (Rumelhart & McClelland, 1986). For our purposes, it is the "distributed" quality that is of utmost importance. This distribution means that our persona is a reflection of the immense number of perceptual and behavioral dispositions encoded by neural structures throughout our brains (and other body loci). *Any* action we perform may be shaped by the interaction of *all* the components of our entire organism.

The second implication is that *the entire microstructure of thought is open to constant revision*. These revisions may result from various sources. The nature of people (indeed of all animals) is such that the brain is constantly receiving input from the environment, and these inputs may be one source of changes in the brain. Additionally, our brains engage in a great deal of spontaneous neural activity, which is equally capable of bringing about synaptic modification. These processes may result in modifications to all of the synaptic connections among all of the neural components of our brain.

3. Free will as self-integration

Now we can turn to a proposal about the nature of free will:

> Free will is the attribute of a person such that every action he or she does is a function of the interaction between the event environment and the sum total of his or her mental dispositions at the moment of that action.

By the *event environment* I mean all the *physical factors outside the person* that bear upon a particular action—everything from gravity to tyranny. This is a departure from most descriptions of free will. Generally, when a person is constrained by physical conditions or by other agents to act or desist from acting in a certain fashion, his or her free will is considered to be impaired. I have chosen to parse the "event horizon" differently for reasons of simple consistency, to sharpen the main point of this proposal. This is my reasoning: neither compatibilists nor Libertarians say that a person's free will is generally impaired by gravity. I suggest that we need to go a bit further. In principle, it makes no difference to the internal processes of the individual whether his ability to soar across a canyon is impaired by gravity, chains, or the grip of another person. Therefore, let us lump all of those factors under the label of the *event environment*, and think of free will as what is happening within the person in question. I have purposely included only physical constraints in

the event environment, since psychological coercion such as threats (e.g., "steal the jewel for me, or I will murder your child") may be significant for our moral attitudes towards a person's behavior, but in no way dictates (in an absolute sense) the action that person chooses to perform. Of course, such threats or promises will affect a person's choices, but in the end such choices are internally generated and not externally imposed [5]. As I mentioned earlier, our main concern is the ontological status of free will, and so the internal realities are the ones of interest.

As the reader may have noticed, the above phrase "the sum total of his or her mental dispositions" might, under some views (including mine), be restated as "the self," to which, after all, we wish to attribute free will. To unpack this notion a bit further: the self is a composite of the dispositions and abilities with which we were born, all the experiences we have ever had [6], and the interactions between them. This is an unabashedly materialist and determinist definition of self. In this simple form, independent of the philosophical exposition it requires, it also sounds quite trivial, in fact almost tautological, and not particularly promising as the key to the question of free will. But we need merely restate it in connectionist terms, and its great power in accounting for our experience of free will becomes immediately apparent.

The key is in the words "sum total." If an action we perform reflects the sum total of our mental dispositions, it reflects all of our beliefs, all of our desires, and all the strategies we have acquired for interacting with the universe. After all, our beliefs are the representation of the world in our neural networks that perceive the spatial and temporal conjunctions of sensory properties that for us constitute objects and environments, and store information about them in the non-propositional form of activation patterns. Our desires are our behavioral inclinations, similarly represented non-propositionally in the brain. Both beliefs and desires can be explicitly expressed following conscious introspection, but it is their implicit representations that affect our progress through life, not our (often fallible) formulations of them [7].

Under other materialist theories of mind, it is hard to imagine how freely willed actions, so defined, could be possible. How could a person bring to bear all of her beliefs and desires in making a finite decision about a possible action, if those beliefs were stated in propositional form, and if her desires needed to be articulated in terms of a potentially infinite test list of conditions to be satisfied? In the connectionist paradigm, and given the understanding of brain interconnectivity noted above, such decisions are possible. "All of our beliefs" and "all of our desires" are naturally represented/encoded in our neural structures in such a form as to enable them to directly affect decision processes, without requiring the mediation of conscious thought [8].

More and more of our mental life is being understood to occur in the realm of the subconscious, with only some end products reaching awareness, and then only sometimes (Dennett, 1991). Perception without awareness is evidenced by cognitive phenomena such as subliminal priming (Marcel, 1983; Merikle & Joordens, 1997), and by neuropsychological phenomena such as blindsight (Weiskrantz, 1997). Neither is consciousness of perceived stimuli required to produce emotional

responses, which, like cognition, involve unconscious processing mechanisms (LeDoux, 1996). Similarly, I propose that volition is a process that occurs subconsciously; only the products of that process are available for conscious report. This suggestion is not as radical as it might seem at first. Think about the mental monologue that might precede a "conscious decision." How often does such a stream of consciousness include a carefully formulated propositional expression of the considerations relevant to a real-life decision? The internal line of patter is invariably more along the lines of: "But … [mental image of person] … angry … tomorrow … football … [affective gut feeling of apprehension] … OK … Seymour?" or the like. This seems to me to be fairly convincing evidence that the subconscious processes of volition are where the action is.

I think that this perspective on mental processes changes the way we relate to the determinist view that, given conditions *x*, person *y* could not have done otherwise than he did. A particular set of antecedent conditions necessarily drives us to produce the response we make [9]. This is a function of our biological nature, and any non-dualist view of mind must accommodate this reality.

Shouldn't this make us despair—this realization that we are automatons with no powers of deliberation, no choices, relentlessly driven to perform mechanically, like a robot or computer? Not in the least. Under the proposed description, our responses (optimally) reflect all of our beliefs and desires, as represented in our neural states. Therefore, those responses are by definition our free will; we *cannot want to will otherwise*, for in doing so we would be false to our essence, to our most dearly held truths, to our strongest hopes, to our deepest desires. Doesn't that sound better?

But just one minute, Dr. Pangloss. If everything we do is necessarily an expression of our free will, we should never desire that our behavior be anything other than what we actually do! However, it is almost a universal human trait to be dissatisfied with some aspect of ourselves: we eat too much, erupt in fits of anger, are unable to concentrate on tasks, have our heads turned by attractive members of whichever sex we prefer, and the like; but all those acts are freely willed, by this definition. If all of these actions are freely willed, why do we experience conflict or ambivalence?

4. Taming the wild modules

The answer lies in what I shall call *residual modularity*. Despite the great extent of mind–brain integration, there are still many very significant mental processes executed by specialized perceptual and behavioral systems. The dependence of our cognition and action on these systems was forged by evolution, and the flow of information throughout the brain is seamless (and beneficial) under normal circumstances. The sleight-of-hand that usually characterizes their smooth functioning is revealed, though, in cases of brain damage, psychiatric illness, or even in the psychopathology of everyday life (not exactly in Freud's sense, but almost so). When the well-oiled meshing of perception and action is undermined, the composite nature of our minds is revealed.

The simplest form of residual modularity is *perceptual automaticity*. Simple visual illusions, experienced by normal individuals under normal conditions, demonstrate that our knowledge about the universe is not always successful in influencing our sensations. Consider the well-known Müller-Lyer illusion (Gregory, 1998). We *know* that the lines on the page are equally long, because we have just measured them with a ruler. Yet we continue to *see them as* being of different lengths because of the orientations of the arrowheads at their ends. Such perceptual "illusions" are present in all the sensory modalities: we locate a sound we hear through earphones directly above our heads, though we know that it is coming in through two discrete speakers immediately over our ears; or take the so-called cutaneous rabbit, in which three bunches of physical taps—at the wrist, elbow, and upper arm—result in the subject feeling a sequence of single taps along the arm separated by small distances (Geldard & Sherrick, 1972), and the like.

Such automaticities reflect modularity at a rather basic level of brain function. For example, the main work of sound localization is accomplished by sub-cortical brain areas that receive no top-down input from the cortex, and therefore cannot adjust their performance despite our conscious knowledge that we are wearing electronic headphones. Similarly, visual illusions, though rooted in cortical processing, reflect the encapsulation or impenetrability of the relevant visual areas to the particular top-down information flow that would serve to adjust our perception to accord with our broader understanding of our circumstances.

It is important to note that perceptual automaticities exist because our sensory systems have been formed during the long process of evolution to provide the optimal interaction with the natural environment. Astronauts have vestibular-system adjustment problems with weightlessness, as (until recently) gravity was a given of life. Earphones wreak havoc in our sound-localization abilities, because in nature sounds are just not presented simultaneously with the same amplitudes to both ears independently of the shapes of our outer ears. Additionally, we should remember that our perceptual strategies are quite adaptive, having proven their survival value not only through the course of human development, but across that of all our evolutionary ancestors that shared a given perceptual mechanism. They enable us to respond quickly to challenges and opportunities afforded by our environment—more quickly than lengthy, exhaustive analyses of the type performed by serial processing systems such as von Neumann-architecture computers. A creature concerned with making it through another day monkeys with these systems at its peril.

A slightly more complex form of residual modularity is *behavioral automaticity*. This includes many complex reflexes: we naturally exhibit freezing behavior when encountering fearful events or situations, and the body's automatic mechanism directing blood to our legs makes our face blanch and gives us the shivers. The facial reaction of disgust to offensive tastes or smells—upper lip curled to the side as the nose wrinkles slightly—is universal, and may reflect (as Darwin suggested) a primordial attempt to close the nostrils against a noxious odor or spit out a toxic food (Goleman, 1995). Changes in the environment automatically trigger in us an orienting reflex (Sokolov, 1963) that lifts our eyebrows, shifts our attention and sometimes our posture towards the source of a salient stimulus, and affects our heart

rate, respiration, skin conductance, and other physiological factors. Once again, the modular nature of these responses is indicated by the fact that they are not subject to direct conscious control.

Other behavioral automaticities are those everyday cases in which we proceed with complex motor processes without the constant input of conscious decision—and sometimes in spite of it. Let me offer one trivial example of a behavioral automaticity that shows how pervasive they are in our everyday lives. This morning, during those hectic minutes of getting the kids out to school, I attempted to prepare a cup of coffee while juggling the phone, the teachers' notices, and the sandwich making. I washed my favorite mug, and decided to dry it before preparing the coffee. I went over to pick up a towel, took out a jar of coffee from the cabinet and, instead of drying the cup, I found myself drying the jar, which was not wet at all. Sound familiar? (I used to think that it was only me who had such slips, but friends, professional colleagues, and the literature have reassured me that they are rather common.) In this case, I had mapped out a motor action plan (wiping with a towel) that was to be applied to a certain target object (the mug). Another object with which I was interacting (the coffee jar) had a great degree of salience for my overall goal at the time (having coffee), and captured the motor action I had planned. Another example of this, which many people report having, is the case of driving by a highway exit that one had planned to take. Certainly the driver is on some level conscious of the road, her car, other cars, her planned destination, etc. Yet the integration between higher-level goal and immediate action is disrupted. Those of us who occasionally experience such slips know how sheepish we feel when we find ourselves performing the unintended action. It is (figuratively) almost as if the behavior that we had contemplated had a mind of its own, and escaped our control. Were such behavioral automaticities to happen often, they would seriously compromise our autonomy as free actors. We are wise in trying to identify and prevent such slips of control from disrupting our planned actions.

Some of us do not have that option. A far more troubling form of behavioral automaticity is what I will call *modular tyranny*. This describes the patterns of behavior exhibited by some individuals with brain damage or psychiatric illnesses. Consider the case of those suffering from post-traumatic stress disorder (PTSD), which seems to result from maladaptive potentiation of the responsive circuits in the amygdala (LeDoux, 1996). They are driven to overt and covert fear responses by environmental stimuli that are filtered out and ignored by the rest of us as being harmless. Or take the case of Tourette's syndrome (movingly described in Sacks, 1985), in which patients pour out an uncontrollable stream of obscenities. Contrary to our intuition that this behavior reflects acquired responses best understood through psychoanalysis, research has revealed that it is a genetically transmitted condition somehow involving abnormal metabolism of the neurotransmitters dopamine and serotonin (Jankovic, 2001). Similarly modular is alien hand syndrome, a clinical disorder in which the patient's hand performs actions that are beyond her conscious control. The actions appear purposeful, but the patient claims the actions are involuntary (Ramachandran & Blakeslee, 1998). As another class of modular maladaptivity, consider psychiatric illnesses such as schizophrenia, bipolar mood

disorder (manic-depression) or obsessive-compulsive disorders, all of which can be present with a great deal of preserved cognitive function. We understand that a person suffering from obsessive-compulsive disorder, spending all day washing his hands and checking dozens of times that he remembered to lock the front door, cannot be thought of as having free will. His actions are mechanically dictated by stereotyped scripts, from which he cannot escape. Thus, obsessive-compulsive disorder is a malady of free will, because it prevents normal strategic planning and meta-control of behavior from overcoming compulsions. Cases in which dysfunctional behaviors produced by these illnesses are curbed by medication that acts on specific neuronal populations (as opposed to global effects of alcohol, anesthesia, or psychotropic drugs) reinforce our intuition that they involve potentially delineable disorders of mental function, even if they are not always localizable to specific brain areas.

The extent of our own susceptibility to control of our modules in everyday life is less obvious to us, but no less real. There are any number of behaviors that inhabit the middle kingdom between the simple automaticities that are innocuous (or even beneficial under many circumstances), and the neurological/psychiatric pathologies just mentioned. Consider road rage—gripping the wheel, tightening neck muscles, racing pulse, trembling and sweating. Think of situations in which the escalating frustration of a harassed parent with a recalcitrant child leads to screaming and corporal reactions that the parent often immediately regrets. These and many other complex reactions represent cases in which our responses seem to emanate from a part of our selves, which temporarily escapes our control. Many of them are characterized as being driven by our limbic system, a group of brain structures which is heavily involved in emotional cognition and responding (LeDoux, 1996; Goleman, 1995). Limbic system behavior can be described as a "shoot first, ask questions later" strategy: good at the OK Corral, not so good in the living room or in the boardroom.

All these cases of residual modularity—innocuous, pathological, and borderline—give us reason to wonder to what extent our behavior can be attributed to an integrated self possessing the attribute of free will, despite the evidence for brain and behavioral integration cited above. Note that malefactors inimical to free will inhabiting any of the "intuition pumps" (Dennett, 1984) that have been used to raise fears about our autonomy are small change relative to modular tyranny. For what threat could there be in nefarious neuroscientists, hidden hypnotists, and invisible jailers (more on this below), compared with *parts of our own minds* that wrench control of our actions away from our integrated selves?

The good news for free will is that (aside from cases of organic dysfunction) we are corrigible regarding our automaticities. By learning about the nature of our cognitively impenetrable modular systems, we can strategically plan our adventures in perception and action so as to maximize our control of them—to bring their functioning into line with our overall values. We can learn which perceptual information is most trustworthy; when it is worthwhile for us to accept the evolutionarily honed speed-accuracy tradeoff with which we are born—and when we should hold out for a better offer.

On the action side, we can be aware of and attempt to change non-adaptive behaviors that are intuitively clear to us (such as an acquired addiction to mood-altering substances), through strategic planning [10]. We can learn to skew our responses to take the automatic actions of our modules into account. We can also attempt to modify to our benefit learned tendencies of which we are initially unaware because of our subjectivity. Take for example a man who constantly bickers with his wife under conditions in which it is simply not justified by circumstances. He may become aware through psychotherapy of harmful acquired response patterns originating in childhood situations. He may then make a stronger strategic effort to be aware of the onset of those responses, and to curtail them, or reorganize his family dynamics to avoid the situations in which they are elicited. We can learn that our overly hearty appetite, which would have served us well during most of the feast-or-famine history of *homo sapiens* and ancestors, is less adaptive in countries of abundance, and that planning, rather than impulse, must govern our diet.

What is the basis of our ability to meta-control the course of our behavior? Neuropsychologically, strategic control of automatic behaviors is considered to require the activity of our frontal lobes, specifically of the orbitofrontal cortex. Consider the classic case of the nineteenth-century railroad engineer Phineas Gage, who sustained orbitofrontal lobe damage that led to a total change in his personality (Damasio, 1994). Before his accident he had been a conscientious and responsible individual; afterwards, despite showing no deficits in overall intellectual function, he became unreliable, callous and contentious. His inability to act in accordance with his preserved abstract knowledge of the potentially negative consequences of his choice—especially in the complex realm of interpersonal relations—left him a social and emotional cripple. His behavior, and the behavior of subsequently studied patients with similar brain damage, can be characterized as being woefully driven by immediate situational properties at the expense of long-term considerations. Such responses obtain despite those patients displaying clear explicit knowledge of their negative consequences. In the absence of orbitofrontal damage, we enjoy the capacity for behavioral meta-control, long-term planning, and personal integration as described above.

Interestingly, two well-regarded recent models—the *somatic marker hypothesis* of Damasio & colleagues (Bechara *et al.*, 2000) and Rolls & associates' reinforcement association account (Rolls, 2000)—emphasize that these frontal executive functions are based on acquired emotional value markers, subconsciously applied to potential courses of action, rather than on explicit propositional reasoning. So a great deal of our control of our modules is exercised automatically. There are likely to be many other brain systems that contribute to the subconscious and conscious control of modular behavior. Obviously, we would especially like to identify the brain basis of rational conscious thought! It is instructive, however, that the orbitofrontal cortex has been demonstrated to be a necessary, if not sufficient, component of the meta-control process [11].

As we see in daily life, there are profound differences between individuals in the ability to execute such strategic controls, in which we feel the strong hand of determinism at work. Genetic dispositions and life experiences dictate not only the

challenges we face, but our ability to meet those challenges as well. We have no reason to assume that people should be more alike in this than in their basketball-playing prowess.

5. Neural holism

Under the description just provided, free will turns out to be a project, not a trait. It turns out that the real issue is not free will vs. determinism. Free will turns out to be a quality that can characterize human beings even in a deterministic universe. The proper question is: how can we get more of it?

If our brains were totally integrated, then every act would indeed be a function of everything we are. But we have modules, which are rather feisty. We can only tame them by understanding them (and perhaps not even then). To the extent that we are able to use learned strategies to achieve integrated expression of self through all of our actions, we realize our maximum free-will potential. I will call this approach to free will *neural holism*, since it is founded on the neural basis of behavior and points towards the integration of brain-based beliefs and desires in the achievement of the kind of free will we wish to possess.

Espousing such a view requires a re-examination of many of the important insights and questions that have featured in the discussion of free will in recent years. In the following sections I will attempt to address some of those issues, with the hope that these comments will serve as groundwork for future articulation of the implications of the neural holism doctrine.

6. Being able to do otherwise

The first issue I address is often at the heart of incompatibilist views on the free will question. Does holding someone responsible for her actions require that she could have done otherwise than she did (Frankfurt, 1969)? Is this a fundamental part of the common human notion of free will, as some (van Inwagen, 1983; Kane, 1996; Ekstrom, 2000) have maintained? Are we really enamored of the ability to do otherwise than we do? I think not. As self-respecting persons, we invest our decisions with the authority of our confidence as rational beings and as competent participants in human society. Being told that we "could have done otherwise"—in *any* sense—simply makes light of this aspect of our self-concept without providing any benefits. How could we look ourselves in the mirror if we really believed that we could equally well have done what we decided—and its opposite?

I think that the contention that people deeply desire the ability to have been able to do otherwise is based on a different psychological reality. None of us like to be predictable. We consider it an affront to our personhood when others take for granted what our responses will be to a particular situation; we feel trivialized by such attitudes. There are good reasons for such feelings on our parts. Predictability makes any creature easy prey. It is adaptive for us to have evolved an aversion to such predictability, and to feel uncomfortable when we sense that our behaviors are transparent to the competition. This very real discomfort lies at the root of many

philosophers' convictions that "the ability to do otherwise" is a foundation of most people's intuitions about free will.

There is, however, another feeling that is central to our complex of intuitions about free will, agency, and autonomy that actually plays the role that has been attributed to the "ability to do otherwise." This is our desire for the capacity to be corrigible, to learn from our experience—so that the next time we find ourselves in a functionally equivalent situation (since we can never step into the same river twice), we can respond differently, should experience have taught us that our previous response was less than optimal. We are not surprised that we often respond automatically and ineffectually to new situations and unfamiliar environments (though we may be anguished by our ineptitude). What we want as autonomous individuals is the capacity to learn [12]. Fool me once, shame on you; fool me twice—shame on me. The "could have done otherwise" condition exists, in a sense, but it applies to other cases in which we will act again. In general, under the model offered here, free will is a quality *over time*, not in every choice or situation. Our free will is redeemed by our ability to use strategies, to think about a series of actions, to plan to do things differently next time we confront similar situations [13].

7. Neural holism and other compatibilist accounts of free action

The neural holism account of free will provides an important solution to the question of what it means to act freely. This becomes evident when comparing the doctrine we espouse with other compatibilist accounts of free will. Compatibilism in general explains free action as being characterized by the lack of external coercion of action or motive. Though an act might be causally deterministically derivative from the past, it is not controlled by anyone or anything external to the self, and is therefore free.

Such accounts, to be effective expositions of free will, require a careful definition of the self, relative to the process of choosing and acting. This turns out to be a rather thorny problem.

Let us start with a very simple definition of self as the sum of one's wants, and freedom as the absence of obstacles to the fulfillment of desire. Then unencumbered acting on desires is free action. The objection to this is obvious: desires themselves can result from *external manipulation* (e.g., guilt, social conditioning, brainwashing, thought implantation by an evil demon or a nefarious neuroscientist) or *internal conflict* between desires (e.g., addiction).

One classic compatibilist approach to solving this problem, proposed by Frankfurt (1971), involves "hierarchical" structures based on people's meta-cognitive abilities. These perform a ranking of desires, and making decisions to act only on some of them. Frankfurt suggests that "free action" is being able to act as one wants, and "free will" is being able to will what one wants. Our "higher-order volitions" dictating which lower-level desires will be satisfied is descriptive of free will. Those higher-order volitions are constitutive of the self [14].

One problem with such an account is that it suffers from infinite regress. For every second-order volition to want to have a first-order desire, there needs to be a

third-order volition to have that a second-order volition, and so on. Additionally, there is an identification problem. Where exactly, in all these escalating levels, does the self reside (Ekstrom, 2000)?

Neural holism is simply not prone to these problems [15]. The *identification* problem is not a challenge to neural holism, since it offers a non-hierarchical description of the values in accordance with which a person acts, and these define his or her self. This will always involve the sum total of our beliefs (i.e., any and every type and form of knowledge about the universe) and our desires (i.e., all of our dispositions to act in any situations). We need not and cannot cordon off a portion of our neural selves and identify that as our true selves, for all our component elements are potentially indispensable parts of what determines those beliefs and desires. These could be perceptual, motor, memory consolidation systems, or even the part of the hypothalamus that regulates our body temperature. In connectionist thinking, our mental states are neither rule-governed nor explicitly listed in a case-by-case format, but rather complexly coded in our total neural structure and synapse weights. There is no way to make these dispositions explicit short of enabling ourselves to act in real-world situations. This is another example of E. M. Forster ingenious observation about mental life: "How can I know what I mean until I see what I say?"

Additionally, neural holism has no *regress* problem. In any given behavior, a fully ramified "neural shakeout" gives integrated expression to whatever desires and biases a person has exactly to the extent that they are reflected in the person's mental (= neural) makeup. All levels of consideration are computed simultaneously, and the system settles into a solution that fulfills the maximal number of constraints, even though this solution may not translate into a semantic description that has any intuitive meaning for us. In a connectionist mind, there is no executive or overseer that is on a higher level than other component parts. All representations and processes are fully distributed over the whole brain.

Neural holism also provides a better account than Frankfurt's later (1992) proposal that "identification is constituted neatly by endorsing higher-level desire with which the person is satisfied … simply *having no interest* in making changes." That attempted reformulation still suffers from the problem that it requires experiencing consequences and endorsing the course of action on the basis of them—and that says nothing about the person's frame of mind when the decision was made. Neural holism's notion of the simultaneous contribution to behavior of all values represented in our neural structures neatly sidesteps that problem.

Another compatibilist-style approach to the challenge of defining self with reference to desire was provided by Watson (1975). He addressed this problem by weighing a definition of free action as one's actions that are in line with one's overall value scheme. An agent acts freely when acting in accordance with values; with "those principles and ends that he—in a cool and non-deceptive moment—articulates as definitive of the good, fulfilling, and defensible life." But Watson himself (1987) rejected this as too rationalistic—we occasionally seek thrill rather than choosing a more rational choice, and we cannot say that this is not our free will, so we need to think in terms of value, not only rational judgment. Watson does not

provide an account of how we determine exactly what a person values—but neural holism, as we have seen, is exactly that.

Not all compatibilist thinkers share Watson's reservations about the limitations of rationality. Double (1991, p. 34) criticizes Frankfurt on the grounds that he provides at best only a subjective criterion for determining which choices belong to agents, and goes on to say that, even if we have marked a particular choice as being the free choice of the agent, he is not *normatively* free because this does not establish the rationality of the choice. Rather, rationality is required as a factor in identifying a choice as being free. This account seems to me fraught with problems. What rationality is required for a choice to be one of free will—are a person's own pre-established standards of rationality sufficient, or are we to hold him to some external criteria of rationality? How detailed a conceptual analysis does a person need to do so that Double will award them the merit badge of having made a freely willed choice? And what if a person has undertaken extensive rational thought about the choice, but has made many errors of reasoning—will we deny them the quality of freedom that they would under other circumstances have been awarded? Double is being rationalistically chauvinist in a matter which should not be dependent on an agent's logical competence. As Tversky & Kahneman (1981) have shown in numerous cases, people make choices irrationally perhaps regularly and often; would we therefore *only on that basis* deny them free will, were we inclined to offer it to them otherwise? Again, our non-judgmental neural holism account is free of these complications, because we need to take no position regarding the rationality, benefit, or utility of any choice for the agent—all we require is that the choice reflects that person's beliefs and desires as instantiated in their neural state, and that is a freely willed action.

It is instructive to consider Double's criticism of Frankfurt and Watson's approaches (Double, 1991, p. 35) that a person acting on desires cannot definitively be said to be free because he may be brainwashed. This kind of criticism makes an implicit but mistaken distinction between processes asserted to have "unfair" control over a person's desires, such as hypnotism [16], and any other experiences a person may have had in life, including whatever education is offered in her or his culture. Our life experiences mold our responses just as definitively as unusual interventions such as hypnosis or brainwashing. But the intuitive sense in which we want free will cannot be that we wish to be independent of anything we have learned (through teaching or experience) in our lives.

Double writes that a self-critical attitude is necessary to achieve free will. But here, too, we confront the problem of infinite regress; this would require self-critical attitude about every belief and desire that a person has, and on what basis? Absolute rationality—can anyone claim to have achieved that?

Neural holism asserts that free will does not require that a person be rational—just the opposite. The process of decision-making is generally not rational, but stochastic. It also does not guarantee that the person will make the decision which is most just, unbiased, or beneficial for her or his own self-interests as judged by expert outside observers. Nevertheless, the behavior that the agent produces will very much be his or her own free will. Free will does not mean infallibility. Ideally,

we would like our responses to be maximally adaptive in optimizing our survival and affective well-being (we are designed to want that). But even less-than-optimal behaviors are still freely willed actions. This is certainly in accord with our intuitions about our behavior and our will.

8. Neural holism starves the bugbears

Dennett (1984) has taught us to beware of and to disable "intuition pumps," the convincing power of which rely on getting us to accept that very difficult conditions can be taken as premises, because they are logically possible. Those premises are then used to jump to far-reaching conclusions about general cases. The discussion of free will, says Dennett, has been populated by many such anxiety-raising intuition pumps, which he terms *bugbears*: nefarious neuroscientists, evil puppeteers, hidden hypnotists, and the like (referred to above). The contribution of these bugbears has been in the fears they raise about various deterministic scenarios: "But that would be like having invisible electrodes inserted in our brain, by means of which some other agent is controlling our every action!" and the like.

A side benefit of the neural holism approach is that the possibility of true external control of any person's behavior approaches impossibility. Recall that the processes governing the dynamic patterns characterizing an organism such as us are stochastic, not rational–logical. The algorithm for generating a totally accurate prediction of the behavior of the organism in the real world will only work on a one-to-one model encompassing every possible variable, down to the molecular level, that effects neural operation—in other words, a clone inhabiting a cloned environment.

What does this do for us? It makes us *effectively unpredictable beings*. We are therefore immune from mind control. This is not to say, of course, that other agents cannot change our behaviors. A simple blow with a large club will do that quite effectively. Rather, I mean that it is impossible for another agent to know with logical certainty what conditions and inputs will produce specific positive actions on our parts. Causing us to experience stupor or sluggishness is easy; but knowing exactly what inputs will, with perfect certainty, cause Professor Smith to marry Mr. Right is impossible. This realization should reinforce our confidence in our personal autonomy.

9. Deliberation

Many discussions of free will have focused on the experience of deliberation—the feeling that we are weighing alternatives, turning them over in our minds, and reaching a decision. Is this illusory? This would seem to be the case, if our choices are pre-determined by internal and external antecedent conditions. Are our intuitions then playing a cruel trick on us? That is what some determinists might tell you, and our sensibilities rebel against such an account [17].

Connectionist models provide a different way of understanding the process of deliberation. Neural network models for the solution of complex problems in

perception or action decisions can be thought of as constraint–satisfaction engines, in which each unit in the system (one or many neurons) represents a hypothesis of some kind relevant to the process at hand. The network solves the problem by satisfying as many constraints as possible using a best-match process in which inputs and outputs are compared recursively (in a step-by-step) fashion until the difference between inputs and desired outputs are minimized. Were all optimal solutions to perceptual and behavioral challenges immediately linearly accessible, life would be a lot simpler. However, it is a reality of network models that they sometimes settle into non-optimal solutions to problems, in which the match between current state and desired state is better than the initial state, but not optimal. This happens because proceeding from where the system is to the optimal state requires a series of steps, some of which are steps backwards, in which the difference between the current state and the target state becomes temporarily greater. In the absence of an executive controller overseeing the operation, how does the system break out of such impasses? One way is by "introducing noise" into the process, which forces the network back out of a more stable configuration into a less stable one, which has the advantage of allowing the next step to be in the direction of the solution possessing absolute optimality. Over the course of an unlimited number of iterations, the system will tend to settle into the most stable configuration, in which the mismatch between current and optimal state is absolutely minimal. The more iterations of this type, the greater the chance that the absolute minimum will be achieved.

This is the neural infrastructure of the process of deliberation. It represents the time needed for us to compute through the settling of our cognitive systems to the optimal solution, with the introduction of recursive "lateral thinking" to help us to avoid local minima (De Bono, 1973). The decision we reach will indeed be determined by our beliefs and desires, but is dependent on the amount of "time on task" we spend, as the cognitive psychologists say. In this sense we have reconciled hard determinism with our feeling of deliberation [18].

10. Free will and consciousness

Relevant to the experience of deliberation is a possible terminological move, the implications of which go far beyond the limitations of this article. In the earlier section about compatibilism, I mentioned both free will and free action, without specifying a distinction between those terms. Let me now suggest the following: perhaps it will be beneficial to set the meaning of "free will" to be the *subjective conscious experience* of volition, while "free action" is the *performance* of neurally holistic actions. A general account of consciousness as an emergent property of some neural systems (such as human brains) should incorporate an explanation of the relationship between the (usually) subconscious processes of action selection and the experience of volition. Working out this relationship is beyond the scope of the present paper [19].

11. A cognitive neuroscience note on moral judgments

It strikes me that the question of free will is often hijacked from where it belongs—in

its rightful place in the philosophy of mind—and held captive in discussions of ethics. The question of free will is generally phrased in terms of moral responsibility: how can we blame or praise somebody for something if they could not have done otherwise? I must confess that I am not terribly troubled by this ostensible problem. It is quite possible that praise and blame (as concepts) are illusory remnants of our primitive notions. They have not proven themselves to be particularly effective in creating a great society, and perhaps we would do better and produce more overall human happiness if we banished such notions from our discourse. Those of us who consider that the philosophical enterprise of ethics is simply impossible can still be concerned about free will.

Of course, philosophers concerned with morals and ethics will adduce the phenomenon of people making moral judgments about others as a real-world basis for the conviction that free will is required for our concept of how people should and do in fact behave. Let us consider, then, our judgment of others. We clearly do this all the time. It seems to be based on the belief that we can assign praise and blame, holding people morally responsible for their actions. But that description is already conceptually biased. In fact, what is happening is that we are constantly assessing the potential dangers and benefits that others hold for our survival. If we get the sense that another agent has a propensity for behaviors that stress a cooperative strategy, we praise them, which is one marker for them to perceive of our willingness to enter into a relationship of cooperation with them over time. Of course, the praise can also be unlinked with subsequent behavior over time. That's why we are so sensitive to others' lack of sincerity, and condemn and ostracize them for it. If we can't take people's verbal behaviors of praise and blame as indications of how we can expect them to behave in the future, and we are led to make strategic plans that assume their cooperativeness while in fact they are going to take a competitive and exploitative strategy, those statements are deceptive and detrimental to our survival. Of course in individual cases we can be deceived (and often are!), but the adaptivity of such a system is evident in the long run. People whose behavior matches their verbal signaling wind up receiving more helpful cooperation from others, which is beneficial for their own survival, while those who in the long run are noticed by others to be deceitful are ostracized and do not get beneficial cooperation. Our brains are wonderfully sensitive to a wealth of somatic cues of other people's attitudes and intentions towards us (perception of affect in face and voice, body language, etc.), and we use that ability automatically in our interactions with others.

Under this approach, the question posed in ethics of judging people based on their intentions or actions takes on new meaning. In moral judgments, we are said to assign less blame to someone if they had an evil intention than if they did an evil action; but why should that be the case? Behaviorally, we can understand that we gauge others' value to our survival not only based on their intentions to us, but on their ability to carry out those intentions. And *in the long run*, the most effective indicator of a person's future ability to carry out an action which might have positive or negative impact on our survival is their past behavior. A murderer has shown that she may have both the inclination and the ability to kill; we therefore assign to her more blame than to another person, of whose murderous intent we may be no less certain.

Note that this is not just a simple-minded lab scientist's response to the question of moral praise and blame; Strawson (1962), in one of the foundational papers of the philosophical discussion of this issue, takes pretty much the same position: that reactive attitudes are natural, not necessarily predicated on people having moral responsibility. Ekstrom (2000) responds that something's being natural doesn't make it justified. I would merely respond that we have no more need to justify such reactive attitudes than I do respiration or digestion. They are "appropriate" because they are adaptive behaviors.

12. Conclusion

Does our proposed understanding of free will address people's concerns about this issue? Gilbert Ryle (cited by Dennett, 1984), in commenting about fatalism, suggested that it is simply not a burning issue for most people. Perhaps this is so because, troubling as the implications of fatalism might be, there is nothing to be *done* about them. At a certain point, mere talk is tiring. Even if it is lacking in other virtues, the present account of free will offers a way in which people can affect how their will operates, by the acquisition of new understandings of their minds and its processes.

Writings about free will often reject particular views on the basis of their not sufficiently accounting for person-in-the-street intuitions about the matter. At the foundation of what I have proposed lies the conviction that whatever free will might be, it is a phenomenon of mind, and phenomena of mind must, in my physicalist conviction, be understood in relation to phenomena of the brain. Whatever that relationship turns out to be, it is vital to the scientific and philosophical enterprises that people's intuitions, many of which are products of naïve dualist views grounded in various religious and cultural traditions, be corrigible in light of empirical insights that were not otherwise intuitively accessible [20]. If such conceptual criticism is disallowed, I am indeed skeptical about the prospect of any progress in the free will debate. Hopefully, the weight of the evidence to which I have referred in this study will be sufficient to bring at least some people to reappraise some of their initial intuitions about free will.

We are told that the search for free will reflects our "unquenchable thirst for individuality and personhood … self-reflectedness … higher aspirations in human beings towards a *worth* for their existence that transcends transitory satisfactions" (Kane, 1996, p. 101), and that free will has implications for personal relations (viz., "love as freely given gift") and satisfaction of desire ("desires of one's own"), and its pursuit is a search for the dignity deriving from knowing that at least some aspects of the direction and outcome of my life owe themselves directly to me (Ekstrom, 2000). All of these valuable conditions can be fulfilled in a deterministic account in which our free actions are those which are a function of our entire system of beliefs and desires, even when of possession of those beliefs and desires are fully dictated by the antecedent conditions of our lives. We can only aspire to the fulfillment of what we are, not what we might possibly have been had circumstances been otherwise.

The maximization of free will is both a project for the individual and a project of human societies over the course of history. Humanity has classically gained new insights into our automaticities through psychological observation, and more recently we have added the tools of cognitive neuroscience to our investigative arsenal. Individuals must painstakingly apply the accumulated insights to their own lives, and adjust them to fit their individual, genetically determined dispositions and life-events. Neural holism requires that we accept our selves as products of our genetic heritage, environment, and personal experience, and those factors only; no hard core of impenetrable individuality, no transcendent essence independent of our place in space and time. Once we overcome our natural reluctance to do so, we will find that we can not only make peace with this understanding of our selfhood, but even to celebrate it, achieving an integrated personality and exercising our truly free will.

Notes

[1] Actually, I endorse a "weak" form of *probabilistic determinism*, which recognizes the existence of randomness on the quantum level. This randomness, however, has a law-like character in the aggregate, in that over the course of large numbers of events, patterns of distribution with general predictive value emerge; since this predictive value approaches absolute determinism as a limit, it provides an acceptable level of explanation for biological and therefore of psychological phenomena. In any event, the indeterminacy of quantum mechanics is no help in providing an account of free will, as it is hard to understand how we could feel empowered by the notion that our choices are functions of random subatomic micro-processes (see below for the critique of libertarianism that follows this line).

[2] More specifically, on those aspects of connectionism that attempt to model human thought, rather than those concerned with information processing in general. Connectionism as described here is sometimes called computational neuroscience or neural network theory.

[3] See Shanon (1993) for an overview of the representational-computational theories of mind and a discussion of their strengths and weaknesses.

[4] For example, take the condition of prosopagnosia (Moscovitch *et al.*, 1997; Bentin *et al.*, 1999), in which a patient with temporal lobe damage is impaired in the recognition of faces while performing normally in the identification of other objects—as in the man who mistook his wife for a hat (Sacks, 1985). Such cases can be taken as evidence for the modularity of face perception.

[5] Parsing the event environment in this way is reminiscent of Galen Strawson's (1986, ch. 2) first sense of self-determination; i.e. of actions, not of choices and decisions.

[6] This includes our episodic memories of life events, our semantic knowledge about the world, and our procedural memories: motor skills, scripts and action schema (Shachter, 1996).

[7] Clearly, much more needs to be said about the relationship between our intuitive understanding of beliefs and desires, and the neural activation patterns that are formed by our sensory experience and its ramification via conscious contemplation. The description I have presented here seems reasonable if we are talking about the belief that a there is a chair in the corner of the room. A somewhat more complex account of semantic information representation in neural networks seems required for the belief that justice always triumphs, or that the soul is immortal. I believe that such an account is possible, but I will not attempt to present one in the current article.

[8] Though such thought can be facilitative under certain circumstances; see below regarding deliberation.

[9] Reservations about the absolute nature of this determination might arise, fueled by our inability in practice to accurately predict the particular behaviors exhibited in specified circumstances. However, our inability to achieve predictive certainty is a function of the functional impossibility

of modeling all of the environmental and neural factors affecting a given action (see below); there is no reason to think that there is any meta-quantum indeterminism whatsoever in the environment–organism interaction itself (pace the libertarian community (e.g., Kane, 1996; Ekstrom; 2000), which has offered an ample range of suggestions about where the indeterminacy underlying free will must be located in order to provide a plausible account of the nature of free action, without providing a shred of evidence that such forms of indeterminism actually obtain). And as we have explained in [1] above, quantum indeterminacy is absolutely inadequate for grounding any meaningful form of agency or autonomy.

[10] Dennett (1984, p. 63), mentions the idea of using higher-order strategies to maximize self-control. Such strategies will only work, though, under the correct understanding of the ways our behaviors are produced by neural dynamics and the integration of modular processes (Goleman, 1995).

[11] So is the orbitofrontal cortex, then, a module for behavioral metacontrol? Not really, since it receives inputs and relies on constantly changing information represented in the amygdala, somatosensory/insular cortices and peripheral nervous system (Bechara *et al.*, 2000) and in temporal lobe higher-level perceptual systems (Rolls, 2000). So while damage to the orbitofrontal area may especially impair our good judgment, that area is part of a widely distributed brain system, not an encapsulated module.

[12] In fact, Dr. Smith's sister, Dr. Jones, often complains to her, "It drives me up the wall that you always think that you're right," to which Dr. Smith always replies, "What are you talking about? Of course I always think I'm right. If I didn't think I was right, I wouldn't say what I said; I'd say something else." Clearly, Dr. Jones is complaining about Dr. Smith's lack of humility. Dr. Smith, however, is not asserting her infallibility; she is making a statement about her naïve epistemology. She is simply claiming that her expressed opinion at any given time is the best showing she can make given the information at her disposal. She avers that she is open to being corrected by new evidence or insights.

[13] Walter (2001) similarly maintains that doing otherwise in a similar situation may suffice for free will. However, he suggests that chaos theory provides a model for how this might occur, in that in chaotic systems small differences in initial conditions can result in considerable long-term effects. It seems to me, though, that neural learning is a much stronger basis for free will over time than chaos.

[14] Aside from the regress and identification problems, there are also internal problems in the type of cases classically used to illustrate Frankfurt's account of first- and second-order desires. An addict desires the rush he gets from heroin. That euphoria (for better or for worse) is a biological reality, not an illusion. However, the problem of the addict (we think) is that this euphoria entails addiction. What is the distinction, though, between addiction and the persistence of desire? Is there a difference in principle between a user's desire for heroin and anyone else's desire for good food and a clean place to sleep? We all desire to get those things on a regular basis and suffer withdrawal symptoms when we are deprived of them. We can survive on gruel in hovels, if necessary, but might steal and kill to avoid having to do so. What about the medical downside of using the drug? The junkie who is aware of the medical perils of heroin and decides to continue using it is making a decision no different in principle than that of a junk-food lover who lives on burgers and fries despite having a dangerously high cholesterol level.

Frankfurt writes about the addict choosing to have the desire to be addicted. Do addicts in general decide that they want to be addicted or dependent? Some, perhaps, have behavioral dispositions towards abdication of responsibility and feel comfortable with a situation in which they can blame their actions on their addiction, claiming that it is beyond their control. In general, however, the second-order desire is not to be addicted, but to accept the price for the first-order desire.

One may argue that one of the effects of heroin is the general impairment of the mental ability to make a fully integrated choice about any future actions, even when not under the immediate effect of the drug (i.e., in "moments of lucidity"). However, the burden of proof that the willing addict's supposed second-order desires have a completely different referent than his first-order desires is on the hierarchical compatibilists.

[15] As far as the problem of our desires being controlled by external manipulations—the stochastic process aspect of neural holism deals with much of the external-control intuition pump (see below).

[16] Not that hypnosis really offers the kind of behavioral control that many philosophers have attributed to it (see Kirsch & Lynn (1995) and Crawford & Gruzelier (1992) for slightly more realistic views of what hypnosis can and cannot accomplish).

[17] Searle's (2001) problems with the epiphenomenal character of deliberation are partially dependent on his insistence on consciousness as a factor in free will. I believe that this account of deliberation more than adequately deals with his difficulties.

[18] I believe that my late father-in-law had a pre-theoretical inkling of this notion. Regarding important life decisions, he used to instruct his children: "Deliberate, deliberate, deliberate—and then decide on impulse." What a great strategy for letting implicit beliefs and desires achieve their full value in our process of making choices!

[19] Consciousness is another topic regarding which cognitive neuroscience has great potential to contribute to philosophical discussion. As property after property of mental function which were thought in the past to be ineluctable aspects of conscious thought turn out to be possible in the absence of consciousness (as in the present case of volitional behavior), the emergent-property approach seems more and more the way to proceed. The challenge of cognitive neuroscience will be to provide a description of the neural prerequisites for the emergence of consciousness and the operational conditions under which it obtains.

[20] A point about intuitions: the starting point of any discussion of free will must indeed be in people's pre-philosophical intuitions about their own feelings of free will and action. However, cognitive neuroscientists (and more generally, all psychologists) relate to intuitions on two levels: not only as ideas to be analyzed and tested, but also as data for the enterprise of understanding human thought. The image of the psychoanalyst in caricature comes to mind: "Tell me, Professor Smith … how long have you been having these delusions that you possess free will?" More seriously, we must ask: What can we learn about cognitive and affective processes from the fact hat so many people have the self-concept of being free agents? This is a non-trivial truth about the human condition, but its exploration is beyond the scope of the present article.

References

Bechara, A., Damasio, H., & Damasio, A. (2000). Emotion, decision making, and the orbitofrontal cortex. *Cerebral Cortex*, *10*, 295–307.

Bentin, S., Deouell, L.Y., & Soroker, N. (1999). Selective visual streaming in face recognition: evidence from developmental prosopagnosia. *Neuroreport*, *10*, 823–827.

Churchland, P.S. and Sejnowski, T.J. (1992). *The computational brain*. Cambridge, MA: MIT Press.

Crawford, H.J. and Gruzelier, J.H. (1992). A midstream view of the neuropsychophysiology of hypnosis: Recent research and future directions. In E. Fromm & M.R. Nash (Eds), *Contemporary hypnosis research* (pp. 227–266). New York: Guilford Press.

Damasio, A. (1994). *Descartes' error*: *Emotion, reason, and the human brain*. New York: G.P. Putnam's Sons.

De Bono, E. (1973). *Lateral thinking: creativity step by step*. New York: Harper & Row.

Dennett, D.C. (1984). *Elbow room: The varieties of free will worth wanting.* Cambridge, MA: MIT Press.

Dennett, D.C. (1991). *Consciousness explained.* Boston, MA: Little Brown.

Double, R. (1991). *The non-reality of free will*. New York: Oxford University Press.

Ekstrom, L.W. (2000). *Free will: A philosophical study*. Boulder, CO: Westview Press.

Fletcher, P. & Tyler, L. (2002). Neural correlates of human memory. *Nature Neuroscience*, *5*, 8–9.

Fodor, J.A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Fodor, J.A. & Pylyshyn, Z.W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3–71.

Frankfurt, H.G. (1969). Alternative possibilities and moral responsibility. *Journal of Philosophy*, *66*, 829–839.

FRANKFURT, H.G. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy*, *68*, 5–20.

FRANKFURT, H.G. (1992). The faintest passion. *Proceedings of the American Philosophical Association, 66*, 5–16.

GELDARD, F.A. & SHERRICK, C.E. (1972). The cutaneous "rabbit": A perceptual illusion. *Science*, *178*, 178–179.

GOLDBLUM, N. (2001). *The brain-shaped mind*. Cambridge: Cambridge University Press.

GOLEMAN, D. (1995). *Emotional intelligence*. New York: Bantam Books.

GREGORY, R.L. (1998). *Eye and brain: the psychology of seeing*. Oxford: Oxford University Press.

JANKOVIC, J. (2001). Tourette's syndrome. *New England Journal of Medicine*, *345*, 1184–1192.

KANDEL, E.R. & PITTENGER, C. (1999). The past, the future and the biology of memory storage. *Philosophical Transactions of the Royal Society of London. Biological Sciences, 354B*, 2027–2052.

KANE, R. (1996). *The significance of free will*. New York: Oxford University Press.

KIRSCH, I. & LYNN, S.J. (1995). Altered state of hypnosis: Changes in the theoretical landscape. *American Psychologist, 50*, 846–858.

KNIGHT, R.T. (1997). Distributed cortical network for visual attention. *Journal of Cognitive Neuroscience*, *9*, 75–91.

LEDOUX, J.E. (1996). *The emotional brain*. New York: Simon & Schuster.

MARCEL, A.J. (1983). Conscious and unconscious perception: An approach to the relations between phenomenal experience and perceptual processes. *Cognitive Psychology*, *15*, 238–300.

MCCLELLAND, J.L., MCNAUGHTON, B.L., & O'REILLY, R.C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*, 419–457.

MERIKLE, P.M. & JOORDENS, S. (1997). Parallels between perception without attention and perception without awareness. *Consciousness and Cognition*, *6*, 219–236.

MOSCOVITCH, M., WINOCUR, G., & BEHRMANN, M. (1997). What is special about face recognition? Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition. *Journal of Cognitive Neuroscience*, *9*, 555–604.

OLSHAUSEN, B.A., ANDERSON, C.H., & VAN ESSEN, D.C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *The Journal of Neuroscience*, *13*, 4700–4719.

POSNER, M.I. & FAN, J. (in press). Attention as an organ system.

RAMACHANDRAN, V.S. & BLAKESLEE, S. (1998). *Phantoms in the brain: Probing the mysteries of the human mind*. New York: William Morrow.

ROLLS, E.T. (2000). The orbitofrontal cortex and reward. *Cerebral Cortex*, *10*, 284–294.

RUMELHART, D. (1990). Brain style computation: Learning and generalization. In S.F. ZORNETZER, J.L. DAVIS, & C. LAU (Eds), *An introduction to neural and electronic networks*. San Diego, CA: Academic Press. 405–420.

RUMELHART, D. & MCCLELLAND, J. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, MA: Bradford Books.

SACKS, O.W. (1985). *The man who mistook his wife for a hat, and other clinical tales*. New York: Summitbooks.

SCHACTER, D.L. (1996). *Searching for memory: the brain, the mind, and the past*. New York: Basic Books.

SEARLE, J.R. (2001). Free will as a problem in neurobiology. *Philosophy*, *76*, 491–514.

SHANON, B. (1993). *The representational and the presentational: An essay on cognition and the study of mind.* New York: Harvester Wheatsheaf.

SHEINBERG, D.L. & LOGOTHETIS, N.K. (1997). The role of temporal cortical areas in perceptual organization. *Proceedings of the National Academy of Sciences*, *94*, 3408–3413.

SOKOLOV, E.N. (1963). *Perception and conditioned reflex*. Oxford: Pergamon.

STRAWSON, G. (1986). *Freedom and belief*. Oxford: Oxford University Press.

STRAWSON, P. (1962). Freedom and resentment. *Proceedings of the British Academy*, *28*, 1–25.

TVERSKY, A. & KAHNEMAN, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.

VAN INWAGEN, P. (1983). *An essay on free will*. Oxford: Clarendon Press.
WALTER, H. (2001). *Neurophilosophy of free will*. Cambridge, MA: MIT Press.
WATSON, G. (1975). Free agency. *Journal of Philosophy*, *72*, 205–220.
WATSON, G. (1987). Free action and free will. *Mind*, *96*, 145–172.
WEISKRANTZ, L. (1997). *Consciousness lost and found: A neuropsychological exploration*. New York: Oxford University Press.