The Journal of Mind and Behavior
Spring 2003, Volume 24, Number 2
Pages 197–228
ISSN 0271–0137

# Agent Causation, Functional Explanation, and Epiphenomenal Engines: Can Conscious Mental Events Be Causally Efficacious?

Stuart Silvers

*Clemson University*

Agent causation presupposes that actions are behaviors under the causal control of the agent's mental states, its beliefs and desires. Here the idea of conscious causation in causal explanations of actions is examined, specifically, actions said to be the result of conscious efforts. Causal–functionalist theories of consciousness purport to be naturalistic accounts of the causal efficacy of consciousness. Flanagan argues that his causal–functionalist (neural correlate) theory of consciousness satisfies naturalistic constraints on causation and that his causal efficacy thesis is compatible with results of Libet's (readiness potential) experiments on conscious causation. First, the notions of conscious effort and conscious causation are analyzed with respect to the project of naturalizing the mind, that is, the attempt to assimilate folk-psychological explanation (explanation by belief and desire) to the causal model of explanation in the natural sciences. It is argued that a serious obstacle for any naturalist program is that mental states are individuated by their (non-causal) semantic content, not the mechanistic, physical properties of their neural state instantiations. In particular, it is argued that explanation by reference to mental state content yields not a causal but an interpretive or rationalizing account of action in which the question of causal efficacy is irrelevant. Then, Flanagan's causal–functionalist theory of consciousness is critically assessed; specifically his interpretations of Libet's negative experimental results on the causal efficacy of consciousness are diagnosed and disputed. It is contended that Flanagan misinterprets the results of Libet's consciousness experiments and that his functionalist concept of consciousness fails to yield an adequate explanation of the alleged causal efficacy of consciousness. Finally, his thesis is countered with other experimental results that appear to favor an epiphenomenalist view over the causal efficacy account of consciousness.

---

Work on an earlier version of this paper was done with the support of a Fulbright Senior Scholar award, at the Department of Logic and Philosophy of Science, University of the Basque Country, San Sebastián, Spain. I express my thanks here for that support. This paper is a contribution to the current Research Project BFF2002-03842 of the Spanish Ministerio de Ciencia y Tecnologia. Requests for reprints should be sent to Stuart Silvers, Department of Philosophy and Religion, 206 Hardin Hall, Clemson University, Clemson, South Carolina 29634–0528. Email: sstuart@clemson.edu or silv8002@bellsouth.net

When playing tennis it is important to hit the ball with topspin. Players try to hit the ball deep into their opponent's side of the court while minimizing the risk of hitting it out. Hitting with topspin is the way to go. To do this, a player "brushes up" the back of the ball, swinging "low-to-high" with the tennis racquet while stroking through it. Tennis "hackers" know to accomplish this one must concentrate, that is, one must consciously try to cause the trajectory of the swing to be such that the racquet face vertically grazes the back of the ball while trying to propel the ball forward. The player makes a conscious effort to do this.[1] To help with this, tennis players (at my level, at least) talk to themselves, reminding themselves of the prescribed procedure. You often see professional players (with disgusted looks on their faces) make a shadow stroke with the prescribed trajectory, especially after failing to properly execute the actual stroke. They are consciously reminding themselves of how to do it. (The great Arthur Ashe left notes to himself on his chair so that at the changeover he could remind himself to bend his knees while making strokes, etc.) These cases (and my first-person report) are meant to illustrate the connection between conscious mental events and subsequent action. Indeed the notion of conscious effort seems tailor-made to bridge the gap between an agent's intention and a thwarted result, particularly when complex muscular coordination is involved. The connection between conscious effort and intended but failed action is elucidated by analyzing the norms associated with performing actions successfully (or competently or adequately, and their respective contraries). These factors contribute to explaining the agent's capacities, both actual and presumed, as well as impediments, both internal and environmental, which interfere with achieving the desired result. The issues involved are of genuine philosophical interest but they are not ones here addressed. Instead the concern is with explanatory models presupposed by the notion of conscious effort, models of mental causation. The influential functionalist thesis of consciousness will be discussed; in particular the focus is on functionalist theories of conscious mental causation as the source of a naturalistic account of agency. For this project the idea of a conscious effort is taken to be a paradigm of agency and the corresponding concept of agent causation. In this opening section, therefore, the nature of the connections between conscious effort, agency, and conscious causation is spelled out.

As should be clear from the "topspin" case, the notion of consciousness at issue here is phenomenal or experiential. It is the feature of some mental states that are said to be subjective, first-person reportable, and characterized typically with expressions such as, "It seems to me . . ." and "It appears to me . . . ."

[1]The ultimate goal, of course, is to do this automatically, to execute the stroke without the occurrent thought.

In calling the phenomena "experiential" the idea is to capture Nagel's (1974, p. 435) characterization of consciousness as "what it is like to be the subject of experience," that he claimed is knowable only by the subject.[2] Owen Flanagan's (1992) functionalist theory adopts the more moderate position in which phenomena reaching the threshold of experience can be properly characterized in terms of expressions like "It seems to me." Moreover, he claims that there is a plausible sense in which (selected) others can know what it seems like to you. However, the focus here is on Flanagan's causal–functionalist theory of consciousness because, as he has developed it over the last ten years, it is perhaps the most detailed exposition of this approach.

## Conscious Effort and Agent Causation

The opening remarks about conscious efforts referred to the intended exertions made consciously and reportably, as when one tries to do A in virtue of explicitly and occurrently wanting to do A. One explicitly and occurrently wants to impart topspin to the tennis ball while hitting it and believes that one will succeed if one brushes up the back of the tennis ball with the racquet. The import of conscious efforts is that it is the player, the agent, who, ceteris paribus brings about the desired result as a consequence of the conscious attention to the details of stroke-making. The agent is, in the ordinary sense, *causally* responsible for bringing about the result, for making it happen. In performing the action an agent makes an effort that is explained by referring to the causal efficacy of that effort. To use a first-person case, the explanation of why I, for example, hit with topspin refers to my conscious effort to brush up the back of the ball (i.e., impart topspin). There are, of course a surfeit of reasons why I fail, when I do. Perhaps even more interesting are the reasons why I succeed if I do. While it pains me to reveal it, in my case the occasional success is usually quite independent of my explicit and occurrent conscious efforts. (My partners often remark that even blind squirrels sometime stumble upon acorns!)

There are also unconscious efforts such as reflexes and other such conditioned and unconditioned behaviors that occur automatically in response to information that does not reach through to the threshold of experience or awareness. These are, by their very description, not actions in the full-blooded sense of falling within one's (apparent) conscious control. Indeed, it is fair to say that robust experimental data indicate that much if not most of

[2]There is serious disagreement about what might be called the grain of consciousness. Nagel and to a lesser extent Jackson (1982) and McGinn (1989a, 1989b, 1999) consider experiential consciousness as ineffable, a view that Flanagan (1992) calls "new mysterianism." Dennett (1991) and (to a lesser extent and for different reasons) Flanagan deny the idea of a kind of experience that transcends our conceptual and linguistic capacities.

behavior is under the control of non-conscious, i.e., not consciously experienced mental states (Lackner and Garrett, 1973 and more recently, Wegner, 2002, and Wilson, 2002). One of the issues is whether behaviors that have been learned or conditioned to be automatic are nevertheless conscious.[3]

Whether one's efforts succeed or fail, the explanations entailing some sort of mental causation are the stock and trade of folk psychology. They are explanations by belief and desire and exhibit all the familiar methodological short comings of mentalistic explanations; they are circular, admit of no independent criteria for beliefs and desires, require countless ceteris paribus clauses to accommodate countless exceptions, etc. Despite the list of debilitating methodological charges against folk psychological explanations, such explanations remain hard-working and remarkably reliable predictive accounts of typical human action. Life as we know it is unthinkable without them. It is precisely this deep entrenchment that makes the functionalizing of conscious mental states so appealing. Everything in our typical, everyday practices succumbs to rational interpretation. Most observed behavior can be made intelligible by interpreting it in light of the agent's beliefs and desires. Dennett's (1971) analysis of the "intentional stance" provides an account of provisionally ascribing rationality to organisms (or complex systems) to explain their behavior in light of the content of the beliefs and desires attributed to them. The ingredients by which actions become intelligible to us include explanation by belief and desire in combination with some sort of empathy and/or analogical inference from the observer's own introspections about what she would do in the observed circumstances. Dennett adopts an agnostic or instrumental stance toward the postulated mental states. However, from a more robust, realist perspective, for such explanations to fall within the scope of science, there needs to be a lawful account of the purported causal relations between conscious states of mind (the beliefs and desires) and consequent behavior.[4] This is the problem of agent causation.[5]

---

[3]In the discussion that follows, neither the question of whether consciousness is a property of states or persons nor the various theories about what makes a state or person conscious will be addressed. It is more useful for the analysis of agent causation and conscious effort to adopt James' (1890) idea of a threshold of experience such that one is consciously aware of whatever information crosses it. This is Flanagan's (1992) approach in distinguishing what he calls "informational" and "experiential sensitivity."

[4]"Purported" because the competing *hermeneutical* thesis denies the relevance of causal relations to the understanding of human action. Those who favor a hermeneutical approach argue that descriptions of causal relations cannot capture the *meaning* or *intelligibility* of action and such descriptions are, in any event, irrelevant to the project of understanding human action.

[5]Chisholm (1976, p. 53) opens his discussion of agency by quoting Aquinas on agent causality:

> Moreover, the particular and the individual are found in a more special and perfect way in rational substances, which have dominion over their own actions and which are not only acted upon, like

Explanation by agency has a lot to recommend it, as the foregoing remarks suggest. The aptness of the feelings of pride and disappointment toward oneself for what is typically taken to be one's actions depends upon the coherence (or intelligibility) of the notion of a conscious effort. The same must be said of the seemingly irrepressible practices of ascribing praise and blame to one's own actions and actions of others. Matters to be taken seriously are those that one is supposed to attend to consciously. It is fair to say that it seems to each of us that we do things; that we try consciously to achieve certain goals. One's individual identity and the self-reflective meaning one's life has, are tied inextricably to the conscious efforts one invests in one's various projects; it is the core of the view one has of oneself as an agent. Agency provides this solace. However, if the agent is also a natural object subject to nature's indifferent causal powers, then the notion of agency, and all the individual integrity that it entails, threatens to be nothing more than an illusion. Flanagan (1996) captures the idea poignantly.

> It matters that I am not just along on some ride that the cosmos, for some absurd reason, is taking. And yet if I am just an animal, if what I think and do is just the emergent product of what the world outside, my body, and its brain jointly produce, then it is hard to see what sense there is to the ideas that I am an agent, that I am self-productive, and that I create or co-create some of the meaning my life has. (p. 53)

To allay such anxieties, that he thinks are "reasonable but not rational," Flanagan (1992) exploits the resources of his functionalist theory of consciousness to explain agent causation in (constructive) naturalistic terms. He develops arguments for the causal efficacy of conscious mental states based in large part on the experimental work of Libet (1985, 1991). The argument developed here is that his characterization of conscious mental states strains the limits of naturalism and that Libet's results do not provide the base that a functionalist theory requires. Moreover, recent results of experiments by Bechara, Damasio, Damasio, and Tranel (1997) tend to confute the idea of a functional role for consciousness. But first some preliminary functionalist and naturalistic essentials are outlined.

---

> others, but which can act themselves. Actions belong to individuals. And thus among substances individuals of a rational nature have a special name: the name is *person*. (*Summa Theologica*, Part I, Question 29, Article II)

Aquinas' "only acted upon . . . others" are objects whose behaviors are explained by causal powers in nature. The naturalist project is to resolve the puzzle created by this dichotomy, to construct an explanatory bridge across what many claim is an unbridgeable gap. Naturalism must explain how we, as persons, can be both "like others," natural objects subject to the indifferent causal forces in nature and agents, "which can act themselves." The puzzle thus discloses two kinds of apparently incompatible explanations: natural cause explanation and agent cause.

## Agency, Mental Causation, and Naturalism

In light of distinctions drawn in the previous section to characterize the idea of a conscious effort, let us understand an action intuitively as something that an agent performs rather than something that happens to or befalls her. Action is thus behavior under an agent's causal control, specifically, under the causal control of the agent's conscious mental states. She is aware of and can issue introspectively based reports on the contents of her mental states. Such reports are not just fallible; a growing body of experimental evidence discloses that they are unreliable (Sternberg, 1996). The explanation of actions presupposes not only the causal efficacy of mental states but also what can be called the agent's efficacy in the causal processes that eventuate in its behavior. That is, to do the work required of a compelling concept of agency, one that satisfactorily grounds ascriptions of moral and epistemic responsibility, an agent has to be the irreducible causal source of her behavior. As Flanagan (1996) noted, mental causation is important to the image one has of oneself as an agent responsible for the choices that one believes one makes.

It is basic to a naturalistic view that the world's constituents are physical objects, their properties and relations. There is serious disagreement among naturalists about the relationship of naturalism to reductionism, namely, whether the properties of and relations among physical objects are fully explicable in purely physical terms. The debate is about the autonomy of psychological explanation and whether psychology qualifies as a "special science" (Fodor, 1974) comparable to biology and geology in its relation to physical laws. Kim (1993, 1998) argues that proposed schemes of reduction of psychology to neuroscience threaten the autonomy of psychology as a "special science."[6] In contrast, Botterill and Carruthers (1999) defend a version of naturalism that is compatible with the irreducibility of the "laws" of special sciences.[7] The discussion here is about some of the specific details of

---

[6]Kim's point is that despite its adherents advertising functionalism as a non-reductionist thesis and despite conceptual chicanery (substituting *naturalism* for *reductionism*) treating mental states as second order descriptors or designators fails to dislodge the fact that the causal powers ascribed to them (to make sense of agency and knowledge) resides in the physical (or neural) states in which they are realized. Kim is clear on this: "If we think of functionalizing as reduction, as I recommended, the problem of mental causation generalizes to supervenient properties that are not reducible to their base properties" (1998, p. 118).

[7]To the objection that irreducibility of "special science" laws would undermine even the most modest semblance of unity and integration of the sciences Botterill and Carruthers (1999) argue:

> On the contrary, we can continue to insist on token identities between special-science occurrences and physical events. That is, whenever a property is tokened which falls under some special-science law (of psychology, as it might be), we can (and surely should) require that the token be identical

the debate with respect to token psychological events and the neural activity that realizes them. Accordingly, naturalism about the mind is the thesis that mental states are (ultimately) neural state configurations in the sense that a token mental state, say, Ria's belief that a Dutch tennis player won the men's Wimbledon championship in 1996, is realized in (or instantiated by) some state or other of her brain. Naturalism thus locates agents in the causal flux. Consonant with this reading, Flanagan states, "By naturalism I simply mean the view that all phenomena are natural and subject to causal principles" (1996, p. 56). Flanagan's functionalism is the non-reductive (constructive) naturalist view that consciousness can be understood in terms of its positive contribution to an agent's success in executing actions and ultimately the survival of organisms in which consciousness is realized. Consciousness is, therefore, a real phenomenon that plays a causal–functional role in explanations of human action. Functionalism thus countenances the general idea of mental causation and argues that the resources of evolutionary theories in biology are sufficient to establish explanations of human action in terms of specifically conscious mental causes. In this way, it is the aim of functionalist accounts to reconstruct the intuitive idea of mental causation within the scope of generalizations and principles that are compatible with, but are not reducible to, the natural laws that constrain scientific explanation. The claim is that although the "laws" and principles used in psychological explanations do not contravene neurophysical law, such laws lack the conceptual resources to capture the explanatorily relevant generalizations. The naturalist's project is to assimilate agent explanations to scientific ones. The obstacles are, however, formidable. Consider the following case.

Typically an agent's voluntary action is explained in terms of some mysterious capacity of the agent to initiate a causal chain of events, which has no relevant causal antecedents. For example, Antonia, while driving, presses down on the directional signal indicator lever because she wants to execute a left turn safely and believes that signaling to do so increases her chances of making a safe left turn. The naturalist acknowledges that there are causal antecedents in all such cases. There are, however, different kinds of causal antecedents. There are the physical conditions involving Antonia's intact neuro-muscular system, her ability to drive, etc., and that she is otherwise uncoerced. Therefore, it is necessary to demarcate the kinds of causes that do not threaten the understanding of voluntary action but are compatible with and indeed make sense of the idea of the voluntary character of the ensuing action. This approach seems to bode well for the naturalist's project. Still,

---

with (be none other than) the tokening of some lower-level (and, ultimately, physical) property happening at the same time. So, belief in the irreducibility of the special sciences is at least consistent with our well-motivated physicalism. (p. 187)

this naturalist must ultimately confront the question of the source of the causal powers ascribed to voluntary action. The naturalist has to identify the source of agent causation. It has been forcefully argued that to infer that the absence of all coercion is sufficient for the performance of the action begs the metaphysical question of freedom entailed by agent causation (van Inwagen, 1983). For it is one thing to specify the causal conditions a mental state must satisfy to occupy the appropriate explanatory role in the production of behavior. It is another thing entirely to give an account of the mental state's causal power.

Naturalists characteristically account for the causal power of mental states in terms of the concept of *supervenience*. Supervenience is a dependence relation between the constituent properties of a thing and the non-constituent properties that emerge from their interaction; it is a relation between properties at distinct levels of interaction. Supervenience may be viewed as a relation between two sets of properties where one determines or anchors the other. Geological properties, for example, are said to supervene on (but are not reducible to) physical properties because two states with exactly the same physical properties will not differ in their geological properties. In such cases it is said that there are no geological facts independent of the physical facts and thus express the (metaphysical, but not explanatory) dependence of geology on physics. Naturalistic explanations in psychology have it that mental states supervene on brain states. Supervenience holds among mental states and brain states in the sense that two neurobiologically identical organisms will not differ psychologically. It is argued that supervenience explains the lawful co-variation of neural and mental states.[8] The apparent causal powers of mental states are in reality functions of the local causal relations among brain states. This helps to shield the notion of a mental cause from such charges as "crazy causation" (Fodor, 1987) and "action-at-a-distance (Block, 1990).[9] Crazy causation is Fodor's classification of any concept of causal relation that fails to supervene on some appropriate physical, i.e., neural, mechanism. This specific *individualist* version of naturalism holds that mental states are the causal products of the neural states on which they supervene. All causation then is *local* in the sense that causal relations obtain among events in virtue of their physical properties. Assuming this view that

[8]Philosophers disagree about whether the dependence should be logical or something weaker, such as physical lawfulness, to adequately explain mental events naturalistically.

[9]Fodor rejects any non-local, non-mechanical concept of cause because it mystifies the cause and effect relation.

> We abandon this (localism) principle at our peril; mind–brain supervenience (identity) is our only plausible account of how mental states could have the causal powers that they do have . . . . In the case of the behavioral consequences of the attitudes, it requires us to individuate them in ways that violate the commonsense taxonomy. So be it. (1987, p. 44)

mental states are realized in neural states, the individualist maintains that whatever causal relations are claimed to hold among mental states, such relations obtain solely in virtue of the physical properties of those underlying brain states. The charge of "crazy causation" is directed toward *externalist* theories of mental content (e.g., Burge's 1979, 1986) in which contextual and historical properties co-determine the content of mental state.

For the last generation or more the concept of supervenience has been the fulcrum in the explanatory machinery used by philosophers of mind to naturalize mentality. While philosophers of science had long exercised the same notion under the name of emergence, it seemed to some mind theorists that there is something special to supervenience that makes plausible the idea of a nonreductive theory of mind (Silvers, 1997) compatible with the general constraints of physicalism. This seemed to confer credence on the idea of naturalizing mental causation and, in turn, agency. Although the naturalizing labors continue unabated some of the initial enthusiasm has waned. In particular, Kim's (1993, 1998) relentless analysis of the supervenience concept has seriously challenged the plausibility of the notion of nonreductive physicalism. In the following section the issue is how certain features of this challenge bear on the concepts of mental causation and conscious efforts.

## Mental Content and Mental Causation

There is no "received" view of naturalism. Indeed, as just characterized, naturalists disagree about its compatibility with physicalism as well as how to elucidate the proposed compatibility. Whatever the outcome of this internecine conflict, insofar as naturalism relies on the supervenience relation, it looks like it is the (physical) brain states, not their mental content, that do the causal work. This is a problem for a naturalistic conception of mind because although mental-cause explanations are articulated in folk psychological language, i.e., the language of beliefs and desires, such explanations fail to meet physicalist standards of causation. And such failure, if that is what it is, makes a mystery of how non-physical mental states exert causal control over the production of behavior. This problem of mental causation is engendered by two wholly distinct modes of understanding and explanation. On the one hand, there are explanations of bodily motions that, under a certain description, realize a specified action. Such explanations of action are formulated with the concepts and vocabulary descriptive of everyday objects and events. More particularly, this latter form of explanation exploits the logical and semantic relations among such terms to rationalize or make intelligible the linkage between mental states and actions. On the other hand, explanations of brain state activity are articulated using the concepts and vocabulary of neuroscience where causal relations are described in terms of

the dynamical properties of neurons. The differences in explanatory domain reflect quite distinct explanatory goals and criteria. Causal models of explanation are, for example, assessed in terms of the accuracy of their predictions of frequency rates and degrees of neural activity. Belief-desire explanatory models aim at understanding actions by rationalizing them. The goal is to make the action intelligible via the meaning relations that obtain among the terms used to describe the contents of an agent's "states of mind." The idea is that predictive accuracy at the neurological level is irrelevant to the understanding of an agent's action because such models predict the wrong kind of phenomena.

This difference in explanatory aims creates the following dilemma for the project of naturalizing agency. If an agent's behavior is explained in the language of the dynamics of her nervous system, the descriptive vocabulary alludes to electrical and thermal properties of neurons, firing rates, levels of neural excitation and inhibition, synapses, axons, dendrites, trajectories of muscular motion, and the like. None of this, it would seem, speaks to the agent's action as described from her perspective, i.e., what she takes herself to be doing and thus fails to address the question of how to understand her behavior in light of her attitudes. Alternatively, if the agent's behavior is explained as *action*, the descriptive language identifies objects and events as she conceptualizes and interprets them. Here it is the vocabulary of beliefs and desires about everyday, middle-size objects of perception that describes the situation.

For example, in the former case, there is the explanation of what is taking place within Antonia's central nervous system that eventuates in the motion of her arm when, while driving, she signals that she intends to turn left. This account depicts the causal relations among the mechanical activities and structures of her neural and motor systems, the features of which are causally responsible for her arm's motions. Notice that an explanation involving reference to the same neuro-motor mechanisms and activities might be given when Antonia reaches for a pencil on her desk or pushes the Tab key on her computer's keyboard. The actions, though entirely different, are, for all intents and purposes, indistinguishable from the neurobiological perspective. In the latter case, the explanation is of the action of Antonia's signaling to turn left, which is something she can do with a countless number of different arm, wrist, hand, and finger motions. She can perform this action in a variety of ways. She can, for example, press down on the directional signal lever on the steering column with either of her hands, with one or more fingers or with her wrist or forearm or elbow. She can lower the driver's side window and extend her arm and any number of her fingers (which might also be the action of drying her nail polish), with or without cigarette, pencil, cell phone, or a countless number of other (unlikely) objects or just with her

fist.[10] The point is that there is no specifiable set of finite motions that constitutes signaling a left turn and thus no causal laws or causal explanations of it.

In contrast, the action of signaling a left turn involves a variety of normative, regulated social conditions and institutions. In short, there are rules that explain what it is and what it *means* to signal, and to *fail* to signal, changes in driving direction. And, to complete the hermeneutical explanatory circle, the normative-explanatory force of rules derives from the propositional or linguistic representation of them as rules, and thus from actions performed or executed in virtue of understanding such representations. In other words, the explanation of why an agent acts one way rather than another (in a given context) is *because* she understands the meanings of the propositional representations that describe the context. "Understanding" here refers to one's knowledge of the meaning relations among the terms comprising the representation. So, the explanation of Antonia's action (while driving) of signaling to turn left, is because she wants to turn left safely and believes that by pressing down on the directional signal lever, she will increase the likelihood that she will execute the turn safely.

Notice too that the explanation ascribes both *intentional content* and *causal efficacy* to Antonia's mental states. The intentional content is what Antonia's mental states are about, that she wants *to turn left safely* and she believes that *signaling increases the likelihood of safe turning* and thus it makes rational sense that Antonia signals a left turn. The folk psychological explanation has it that her belief and her desire ceteris paribus cause Antonia to activate the left turn signal in her car. The movement of Antonia's arm is the action of signaling in virtue of its being under the causal control of her relevant mental states.

This extended example is to illustrate the idea of the *content individuating properties* of mental states. Content individuating properties are used in formulating the familiar idea that beliefs (desires, hopes, fears etc.) are distinguished from one another by the differences in their content. Antonia's two beliefs are individuated in terms of the differences in the two distinct contents of her belief states, namely, *that pressing down on the directional signal lever activates lights on the left of the car*, and *that pressing up on the lever activates the lights on the right side*. So much is (folk-psychologically) obvious. The problem is that because naturalized agency locates the causal powers of the agent's mental state in the agent's brain states, the mental states themselves as identified by their *content*, seem to be rendered causally inert. Antonia causally *initiates* the set of motions, beginning with the downward pressing of

[10]Upon arriving in South Carolina, my wife, who is European, hypothesized that there was a local tax on the use of directional signals while driving to explain why so few drivers signaled. It is now agreed that there are more plausible and more distressing hypotheses!

the lever that constitutes signaling to turn left, in virtue of the content of her beliefs and desires. She presses down rather than up because of her beliefs and desires about activating the left turn signal, not the right turn signal. However, if from a neurobiological perspective, the only difference between the mental states implicated in turning left and turning right is their intentional content, then the explanation, for all its plausibility, is not causal. The reason is that, as McGinn (1989a) emphasizes, the content of mental states is not a mechanistic feature of the world. Content is not mechanistic because the content of the mental state is determined by the semantic relations it bears to other mental content.[11] It is the network of such semantic relations that determines mental content. Thus the content of your belief is distinct from the vehicle, the brain state, that bears it because the (individuating) properties of mental state content are manifestly different from the properties that distinguish brain states. The latter are, in McGinn's term *mechanistic*, whereas the former are not — they are semantic in character. They are two distinct kinds of properties.

As a result, insofar as agency presupposes conscious mental causation, it remains mysterious. Brain state explanations, as noted, are formulated in the languages of brain science that expresses the law-like generalization of neurobiology, neurophysiology, and neurochemistry. In contrast, agent explanations are framed in terms that are related via meanings, not laws. Since the meaning properties of the terms used to describe the intentional content of mental states are of a wholly different order and do not reduce to the mechanical properties of the underlying brain states, the problem for naturalists is to account for the claimed causal efficacy of mental states in agent explanations. In short, the problem is to explain the local causal efficacy that naturalized agency requires in terms of properties that do not supervene on brain mechanisms.[12]

Still, there is the need to explain the rationality of the agent's action by subsuming it under the scope of what she believes and wants, and these are identified with the propositional contents of the relevant attitudes. One appealing remedy for the dilemma is to explain action by the content individuating properties of mental states and to show that mental state content does play a causal–functional role. If a plausible case of this kind can be made it would support the thesis of a functional explanation (and theory) of naturalized free agency.

---

[11]For example, your belief that Tom is a brother entails your belief that Tom is a male sibling. Therefore, "My brother is an only child" is not a logically possible content of a belief that you or anyone can have.

[12]The apparent failure of reductionist accounts of phenomenal consciousness has spawned a new crop of absolute-emergence theories of consciousness. Some seem to endorse McGinn's (1996) idiosyncratic naturalism that he describes as "wise incomprehension."

When content individuating properties reach the threshold of experience they are, as has been argued, just those with the precious features that make sense of the idea of conscious effort. Recalling the topspin case, the explanation of the effort to brush up the back of the ball with the tennis racquet is because one wants to hit it with topspin (and believes that brushing up imparts topspin, etc.). If successful, such a functional theory should help avoid a nettlesome dilemma, namely, having to decide whether to indulge the concept of cause to fit the idea of psychological kinds that agency requires or to adjust the mental kinds to meet constraints on (local) causation. Can it be both ways?[13] Can there be a concept of agency that is both consciously rational and naturalized?[14]

If conscious mental states are to fulfill a bone fide functional role in the explanation of action, empirical evidence must bear out that such consciousness improves the performance of action, that is, that consciousness enhances effort. The expectation is that ceteris paribus one executes topspin tennis strokes better when one consciously attends to the effort than when one does not. Flanagan (1992, 1996) attempts to salvage the intuitively plausible idea of conscious mental causation and to make the naturalistic view of the world safe for free agency. He draws upon results of experiments in psychology to argue against epiphenomenalism, which he contends imperils the view one has of oneself as an agent. This epiphenomenalist threat is forbidding because, to recall Flanagan's poignant remarks above, ". . . [If] what I think and do is just the emergent product of what the world outside my body, and my brain jointly produce, then it is hard to see what sense there is to the idea(s) that I am an agent" (1996, p. 53). Below the main points in his argument are spelled out and then followed by an examination of his interpretation of the experimental evidence that he claims underwrites a functionalist account of conscious causation.

## Epiphenomenalist Anxiety

The distinction has been stressed between conscious and non-conscious mental causation to delineate precisely the idea of conscious effort. Flanagan (1992) explains a system's non-conscious mental processes distinguishing between "informational sensitivity" and "experiential sensitivity." One is informationally sensitive to a vast array of (cognitively processed) stimuli that fail to reach the threshold of consciousness. Subdividing consciousness any further is, he argues, linguistically confusing and irrelevant to the expla-

[13]Everyone's goal is to establish that the causal and the intelligibility theses are not (must not be!) incompatible.

[14]See footnote 9 for Fodor's decision.

nation of action (cf. Block, 1995). Essential reference is made to such nonconscious mental phenomena in psychological explanations of discrimination, categorization, reaction to environmental stimuli, cognitive integration of information, internal monitoring, and the like. Such cognitive operations are executed without one's being conscious of them. It is a commonplace that there is no awareness of the sensory information processing that occurs continuously in the nervous systems.[15] In mentalist vocabulary these kinds of phenomena figure in the descriptions of the various *propositional attitudes*. In psychology such latent capacities are viewed as dispositions. There is also the distinction between dispositional beliefs and occurrent beliefs; between Antonia's dispositional belief that her computer has 128 MB of memory and her occurrent belief that there are cracker crumbs on the keyboard. If conscious beliefs and desires are to fulfill the causal role assigned to them it is essential to identify their experiential properties.

To this end, Flanagan argues that occurrent beliefs differ from standing beliefs in two critical ways. Occurrent beliefs are manifest and locally causal in the production of behavior and they are also phenomenally conscious. That is, there is not only awareness of occurrent beliefs declaratively in the sense of the ability to report them; this declarative awareness also has an experiential quality. Dispositional or standing beliefs per se lack that experiential feature in virtue of their subpersonal nature. In short, occurrent mental states come about with an experiential quality that contributes to the ability to distinguish one attitude from another (for example, belief from hopes, desires, expectations, etc.) as well as differentiating between disparate mental state content. Beliefs have, he says, for lack of a better term, a "beliefy" feel that differs experientially from desires, expectations, hopes, and the countless other distinguishably different propositional attitudes. The explanation "is that there really is a certain way it feels to believe something and that way of feeling is different from the way it feels to desire something" (1992, p. 68). Flanagan endorses the idea that occurrent beliefs have a "beliefy" quality and wonders if being in a brain state with a beliefy feel causes one's confident belief assertions. Goldman also suggests "a phenomenological model for the attitudes" (1993, p. 364).[16] He argues that occurrent propositional attitude states have a qualitative character. "'Abstract' or 'conceptual' thought often occupies awareness or consciousness, even if it is phenomenologically 'thinner' than modality-specific experience" (p. 365). Moreover, "someone who had never experienced certain propositional atti-

[15]In contrast, Chalmer's (1996, p. xii) "hard problem of consciousness" is the mystery of "Why is all (sic) this processing accompanied by an experienced inner life."

[16]Quotations are from the paper reprinted in his 1993 anthology.

tudes, for example, doubt or disappointment, would learn new things on first undergoing these experiences" (p. 365).[17] With respect to the first occurrence of these mental states, Goldman concludes, "there is something that it is like to have these attitudes" (p. 365). In addition, he stresses that there is an introspectively discriminable property of attitude strength that people report in terms of the confidence in their beliefs and intensity of their desires. Indeed, mentalistic vocabulary is rife with this kind of phenomenological weighting.

Such claims are, however, more like introspective reports than arguments. For all its heuristic value, phenomenology is notoriously treacherous evidential terrain. On the one hand, it is not obvious that occurrent beliefs have a discriminable, phenomenal quality. In the case of "analytic" truths, e.g., the belief that two plus two is four or that bachelors are unmarried, the phenomenological claim that these have some discernable qualitative feature seems quite unfounded.[18] On the other hand, with respect to the experienced degree of belief confidence ("attitude strength"), there is experimental data indicating that, "the eyewitness accuracy–confidence relationship is weak under laboratory conditions and functionally useless in forensically representative settings" (Wells and Murray, 1984, p. 165). Indeed, experimental data suggest something of an *inverse* correlation between belief confidence and accuracy (Loftus, 1979).

Despite these skeptical concerns, let us suppose that occurrent beliefs are in some way qualitative experiential phenomena. There remains an even more serious question about their status in the causal chain that produces action. That question concerns the causal efficacy of phenomenal experience: "How do phenomenally experienced, non-physical occurrent beliefs cause the physical processes that realize human action?" There is the classical physicalist view that denies the causal efficacy of experience, engendering either eliminativism or epiphenomenalism (see Kim, 1998). This is a no win situation for functionalism. The reason is that whereas eliminativism disavows consciousness as an artifact of an empirically false "folk" theory of mind, epiphenomenalism renders it causally inert.[19] A more modest, nonre-

---

[17]Similarly, Jackendoff (1987) extracts from what he regards as the otherwise absurd behaviorist idea that thought is mere subvocal speech the "phenomenological insight" that thoughts are experienced as a series of linguistic images. "That is, although linguistic images cannot be identical to thought, they constitute our experiential evidence that thought is taking place" (p. 288).

[18]In his comments on an earlier, lecture version of this paper, Tyler Burge reported to me that he experienced no identifiable phenomena feel with his beliefs involving logical truths.

[19]See P.M. Churchland (1995) for the latest version. He has subsequently modified his radical position to a more reductionist version.

ductive physicalism thesis ascribes a causal functional role to experience. If occurrent beliefs are part of the local causation of action, and beliefs are experiential phenomena or involve experience essentially, then it would follow that either action is explained by Fodor's crazy causation or that phenomenal experience does have a causal–functional role in guiding behavior. Flanagan opts for the latter alternative and adopts the resources of neo-Darwinian evolutionary biology to underwrite the argument that phenomenal experience is not excluded from cognitively causal transactions.[20] His thesis is, therefore, that phenomenal experience does have a significant causal–functional role in the production and explanation of action.

## Suspicious Phenomenal Functionalism

Since functionalism acknowledges conscious states, it rejects both eliminativism and reductionism, leaving epiphenomenalism as a live option. To counter what he calls the epiphenomenalist suspicion Flanagan (1992) bases his functionalist account of phenomenal experience on Libet's (1985) well-known "readiness potential" experiments.[21] Libet connected subjects to electroencephalographs that measure the brain's readiness potentials in the somatasensory cortex and to electromyographs that register hand muscle activity. Stimulating certain sections of the somatosensory cortex yields sensations in corresponding sections of the body. The subjects were instructed to flex their right hands spontaneously whenever they wished. The experiment's design called for the subjects "to pay close introspective attention to the instant of the onset of the urge, desire, or decision to perform each such act and to the corresponding position of a revolving spot on a clock face (indicating 'clock time'). The subject is also instructed to allow such acts to arise 'spontaneously,' without deliberately planning or paying attention to the 'prospect' of acting in advance" (Libet, 1985, p. 530). The instruction to "pay close introspective attention to the instant of the onset" would seem to correspond to the characterization of reportable, declarative consciousness. Presumably, subjects could report their conscious detection of the urge to flex and mark this awareness-event via the revolving spot on the clock face. Flanagan notes that in the cases where there was no preplanning, and where there was conscious preplanning ("of getting ready to spontaneously flex a few seconds before they flexed") the hypothesis that "the readiness poten-

[20]Flanagan's evolutionary approach distinguishes consciousness as being selected *for* in contrast with dreaming which he regards as an exaptation or free-rider. See his *Dreaming Souls* (2000).

[21]See Velmans (1991) and peer commentary for an informative exchange.

tials precede conscious intention or urge which precedes muscle movement, was confirmed" (1992, p. 136). Libet (1985) reports,

> Onsets of readiness potentials regularly begin at several hundred ms before reported time for awareness of any intention to act in the case of the acts performed ad lib. It would appear, therefore, that some neuronal activity associated with the eventual performance of the act occurs before any (recallable) conscious initiation or intervention. This leads to the conclusion that cerebral initiation of the kind studied . . . can and usually does begin unconsciously. (p. 536)

In light of these findings, Libet's own interpretive discussion is instructive for he questions the idea of a functional role for consciousness. "If the brain can initiate a voluntary (sic) act before the appearance of conscious intention, that is, if the initiation of the specific performance of the act is by unconscious processes, is there any role for conscious function?" His hypothesis is: "Conscious control can be exerted before the final motor outflow to select or control volitional outcome. The volitional process, initiated unconsciously, can either be consciously permitted to proceed to consummation in the motor act or be consciously 'vetoed'" (pp. 536–537).[22] This seems to confer support on the functionalist view.

However, Libet's comments on veto power are misleading for he refers to volitional processes that are initiated unconsciously. The worry here has to do with the relationship between processes that are at once both volitional and unconsciously initiated. For volitional processes are typically those identified with agency and agent causation. Unconsciously initiated processes, however, are those said to be not under the agent's conscious *causal* control. In the language of the free will issue, volitional processes are, on pain of a regress of causes, ones the agent brings about independently of causal influence. They are things she does but, under the exact same causal conditions, could have done otherwise. In contrast, unconscious initiations are subpersonal processes that occur or happen to us because they are instances of exceptionless regularities that are the laws of nature. What is at stake here is whether there are categories of events (miracles?) that violate or are otherwise outside the scope of natural law. While not to digress into the free will–determinism controversy, it is important to identify the common metaphysical space occupied by the concepts of agency, freedom, and conscious mental causation.

Flanagan puts forward three arguments for interpreting Libet's results as endorsing the functional role view of consciousness compatible with physicalism. First, there is the presence of a "veto power" capacity. This supports

---

[22]MacKay (1991) too remarks, "Remember that Libet's (1985) findings are still compatible with the veto capability of volition" (p. 688).

the idea that agents can and do exert causal conscious influence over the course of their actions. "So long as it (mind) can stop a motor movement before it occurs, it does not need to actively trigger it" (1992, p. 137). Even if one does not consciously initiate the actions she can exercise some measure of conscious control over their trajectory and duration. Second, "conscious mental processes emerge out of the neural processes that give rise to them. It would be absurd to expect these emergent conscious neural processes to precede the neural processes they arise from" (p. 137). Third (reminiscent of Dennett's [1978] critique of Skinnerian experimental design), Flanagan argues that the conscious instructions to the subjects already establish the temporal precedence of consciousness to the onset of the readiness potentials. These will be discussed in order.

1. As noted above in passing, Libet seems to be testing the idea of declarative consciousness, the kind that can be reported via introspection. So it is not clear that there is some additional phenomenal property that accompanies the subjects' beliefs about when their urges begin (and end), that is, something subjectively experiential about which they cannot be mistaken. In contrast, there is reason to think that what the subjects report via introspection, are their *beliefs* about the onset of their urges, judgments about when their experience begins, and such beliefs may very well be mistaken.

Flanagan's interpretation of the "veto power" proviso is seductive; it appeals to the folk intuition that the way things feel does make a causal difference to one's subsequent behavior. As Dennett and Kinsbourne (1992) remark, "This picture is compelling but incoherent" (p. 166). To see why consider that, as noted above, it is not obvious that urges feel a certain *urge-like* way or that urges are conscious events or that one can state when an urge is operative. Reporting my own case for example, if I unwittingly trip, I have the urge to resist falling and remain upright. In most case I do not feel some special way other than that I'm falling. I do not consciously feel anything at all. If anything, I consciously find myself in the process of resisting the fall. In this sense, my urge to remain upright is not a conscious effort. Indeed, I do not consciously exercise my capacity to remain upright and even less, in most cases do I find myself falling and with the "speed of consciousness" put my (proprioceptive) capacities into gear, as it were. In contrast to such reflex-like behaviors, Libet's experiments involve deliberative processes. Even with such cases Flanagan's explanation of the experimental results is problematic. The acknowledged veto power capacity fails to explain the *exercise* of the capacity, how it is put into gear. In light of Libet's hypothesis it seems not only plausible but also essential that the onset of the veto is also preceded by and causally explained by readiness potentials.

In reply, Flanagan might argue that although readiness potentials reflect neuromuscular anticipation of movement, a veto would not have to have

this. Rather, the precedence of the neural activity, so the argument might continue, does not undermine the efficacy of consciousness; it merely shows that neural activity precedes conscious awareness. This seems somewhat disingenuous, however, for it still leaves unanswered the question of the etiology of the onset of the veto. This purported reply has it that while there is, consistent with the readiness potential hypothesis, neural activity preceding the consciously initiated veto, this activity is not causally responsible for the exercise of the veto. Alternatively, it would appear that neural activity subsequent to the onset of the initial activation needs either to reach back into the causal chain of production or, at the decisive moment, initiate an alternative chain of neural events to inhibit the action to be vetoed. As Dennett and Kinsbourne remark, "For one thing, such a 'veto' would itself have to be a 'conscious decision,' it seems, and hence ought to require its own 300–5000 ms cerebral preparation — unless one is assuming outright Cartesian dualism" (1992, p. 166).[23] The conclusion here is that neither of these interpretations is consistent with the scientific understanding of the operation of causal laws.

The veto power capacity that purportedly emanates from conscious volition also requires a specific moment of initiation. The timing of phenomenal initiation is a highly vexed question whose coherence has been disputed vigorously (Dennett, 1991; Dennett and Kinsbourne, 1992). The issue here involves the subject's conscious, temporal registration of the onset of the urge to act "by noting the position of a spot on a revolving disk (the 'second hand' on a clock, in effect . . ." (Dennett, 1991, p. 163). The difficulty, for Dennett, lies in the fact that there is no such verifiably determinate moment. To think that there is such a moment of conscious registration is to fall into the Cartesian Theater with all its trappings. The problem is this: the readiness potentials are precisely timed and so is the ensuing movement. The veto's precise timing is questionable because the subject's report or press of a button is not a reliable indicator of the relationship between the instant of a specific conscious occurrence and the underlying neural event in the subject's own brain. Dennett's (1991) distinction between the time of a representation and the representation of time illustrates that the temporal order of brain events (such as readiness potentials) need not correlate with nor mirror the temporal order of "corresponding" conscious events. To preserve representational coherence and to meet task needs "at hand," the brain's control systems can rearrange the temporal order of its own neural activity. The point is, as Dennett and Kinsbourne (1992) emphasize, that the brain's representation of time is of a piece with the way it represents other perceived

[23]Dennett and Kinsbourne add that MacKay (1985) makes a similar point.

objects and events in "ecological" space. To say that the veto simply needs to occur sometime before the behavior even if the experimental subject cannot say exactly when she vetoes, would be to concede that the conscious vetoing need not be the causally efficacious event. By hypothesis, it is the subject's conscious exercise of her veto capacity that causes her ensuing motion. However, if there is no determinably specific moment when she decides to veto (that is, there is no determinable time of the onset of the conscious veto act), then there is no evidence that it was the veto rather than some unconscious (physical) event that caused the motion. Dennett and Kinsbourne suggest the reason for there being no evidence is ". . . because there is no such moment of absolute time" (1992, p. 164). Libet too has acknowledged that the problem of time determination "was not experimentally testable" (1985, p. 560). Recent experimental results by the Damasios and associates suggest a line of interpretation consonant with this view.

The experiments do not establish that exercise of veto power is a case of conscious agent causation in the sense of being explicable without reference to antecedent physical conditions. The phenomenon of veto power then does not seem robust enough to support Flanagan's optimistic, functionalist interpretation of Libet's experimental results. Although testability does not exhaust the empiricist's methodology, when the experimental results are drawn upon to underwrite a hypothesis (the functional role of consciousness) the experimental test in question deserves close scrutiny. This point reappears in the brief discussion of Flanagan's third argument for viewing Libet's results as unfavorable to epiphenomenalism.

2. "A lurking Cartesian intuition" about consciousness, cautions Flanagan (1992), prevents us from recognizing the naturalness of brain processes preceding conscious experience. "It would be completely unexpected if all the causal antecedents of conscious mental processes were themselves conscious. In other words, conscious mental processes emerge out of the neural processes that give rise to them. It would be absurd to expect these emergent conscious neural processes to precede the neural processes they arise from" (p. 137). As conscious cognizers agents are physical systems in which consciousness emerges as the result of complex contingent factors. The complexity of mechanisms in the neural substrate results from selectional pressures that favored organisms whose internal behavioral control systems happened to enable them to adapt to and survive in environments where their competitors failed. Roughly, the idea is that at a certain stage of development, the interactions of certain physical processes became, in Chalmers' phrase, "accompanied by an experienced inner life" (1996, p. xii). The question is not whether there is such a phenomenon as the redundantly described, conscious experience; the issue is rather whether this phenomenon is causally efficacious. In his functionalist view, van Gulick (1990, 1991) also argues

that phenomenal experience does make a difference in the efficiency of the underlying processing mechanisms. So unless it can be established that this accompanying experienced inner life does have causal powers, phenomenal experience might turn out to be a free rider, albeit the most dramatic and striking free rider yet encountered.

The functionalist argument to block this inference to the causal impotence of consciousness is that although phenomenal experience was not selected for, it nevertheless functions in a way that, according to Flanagan (1992, 1996) derivatively enhances survival (and perhaps thus enjoys a vicarious selection). Dreams, for instance, are just such free riders; their presence can be explained without having to claim that they are adaptations selected for naturally because they contribute positively to the dreamer's well being and survival. They do not, but sleep and sleep cycling do.[24] By parity of reasoning, it is argued that Mother Nature selected for nonconscious causal neural mechanisms; consciousness, like dreams, comes as a free rider. The upshot then is that the account of conscious neural processes emerging from underlying non-conscious ones explains the causal antecedents of consciousness but by itself does not address the issue of causal efficacy. The explanation of the emergence of consciousness from non-conscious neural processes is fully compatible with the conscious phenomena being causally inert.

Functionalist theories of consciousness are committed to explaining the causal efficacy of conscious phenomena within the constraints of naturalism. Flanagan's version is a neural correlate theory. Some neural processes make it across the threshold of experiential sensitivity and one becomes conscious of them, not in terms of constituent neural properties, but in terms of their *phenomenological* properties. It is these phenomenological properties that provide the experiencing subjects with a distinct point of view. It seems, however, that there is something missing or suppressed in the explanatory schemes of the functional theories advocated by van Gulick and Flanagan. The tacit assumption is that since unconscious neural processes temporally precede those neural processes that become conscious, their antecedent occurrence explains (or is accompanied by the explanation of) how conscious neural processes are produced from unconscious neural processes. The "production problem" is to explain how physical processes generate phenomenal experience at all (cf. Seager, 1995, on what he calls "the generation problem"). Without such an account one is left with a correlation story that does not enjoy the counterfactual support required of lawful explanation.

[24]"Some think that dreaming is pure noise. But this is wrong. I'll explain why this is false, even though Mother Nature did not *select* for dreams. She selected for sleep and sleep-cycling. Dreams came along as a free rider" (Flanagan, 1996, p. 32).

To spell this out, consider Flanagan's argument. He suggests the analogy that consciousness supervenes on neural activity as ice supervenes on water molecules whose mean kinetic energy decreases to realize 32° F. The slowing of the molecular action precedes the ice formation. The supervenience of both ice and liquidity on $H_2O$ molecules is transparent in the sense that the antecedent and consequent phenomena are ontologically homogeneous. Opponents of the neural-state/mental-state identity theory will argue that unlike the water and ice case, phenomenal experience has qualitative properties that are not fully explicable in terms of physical properties and that the analogy begs the question. The question at issue is about the opaque connection between ontologically heterogeneous features of the world (cf. McGinn, 1989b and Levine, 1982).

Although van Gulick (1995) endorses functionalism, he also points out that there is no adequate account of the required relation between conscious explanda and physical explanans. He argues that from the point of view of scientific explanation, neither logical sufficiency (deductive entailment) nor nomic sufficiency is a plausible candidate for the appropriate psychophysical relation. He defines a weaker notion of intuitive sufficiency as, "a set of processes (that) can be seen intuitively to produce or realize whatever feature of consciousness we are aiming to understand" (p. 71). This more relaxed condition is also unsatisfactory, however, because it omits, "most importantly what is to count as an intuitive process or explanation." He is especially skeptical about just the sort of analogical accounts that Flanagan adopts.

> One can appeal to examples from other domains that seem to meet the standard, such as explaining the room temperature liquidity of water in terms of its molecular structure or its frozen state below 0° C in terms of the intramolecular hydrogen bonds that produce ice crystals. But citing examples is not the same thing as defining intuitiveness, and it is far from clear how we are to generalize from examples like those of liquid water and ice to physical or functional explanations of one or another feature of consciousness. (p. 71)

Flanagan's argument is all the more difficult to understand since his view is an identity theory. So the underlying neural processes cause other processes, some of which cross the threshold of conscious experience. Which processes do and which do not remain a mystery. From an evolutionary perspective, it might be that some neural processes are too "important" vis-à-vis the organism's survival to be entrusted to the good offices of conscious processing. If so, it would also seem that the causal powers ascribed intuitively to consciousness are, in computer scientist Alan Kay's (1984) catchy term, a *user's illusion*.[25] Perhaps the anecdotal explanations of reacting automatically, non-

[25]This term has been popularized by Dennett and is the title of a book by Nørretranders (1998).

consciously, in critical conditions fall fruitfully into the context of discovery.[26] There are similar questions about the relevant kinds of complexity and about the kinds of creatures in which phenomenal consciousness may reasonably be expected to emerge. And last but not least, why, it may be asked, does crossing the threshold of experience produce the qualitative character of phenomenal consciousness at all.

3. The third reason Flanagan offers for his interpretation of Libet's results involves experimental design. Consciousness inevitably precedes the readiness potentials because the subjects receive instructions and are thus consciously aware and consciously comply with the experimental rules. As noted above, there is no airtight way of eliminating the presence of consciousness from experimental procedures designed to test the existence and hypothesized function of conscious experience. The crucial point is, however, that within the scope of the experimental situation there is no relevant conscious data antecedent to the experimental task. Flanagan therefore infers incorrectly when he conclude(s) that Libet's results ". . . are precisely the sort of results one would expect if one believes that conscious processes are subserved by nonconscious neural activity, and that conscious processes play variable but significant causal roles at various points in different cognitive domains" (1992, p. 139). Libet himself concludes, "There is no experimental evidence that would deny a causal role for conscious control function here, although admittedly there is none to demonstrate such a role either" (1985, p. 685). So in this case the results cut neither in favor of nor against epiphenomenalism.

## Causation, Agency, and Functionalism

Teleological functionalist theorists of consciousness (Flanagan 1992; Lycan 1981, 1987; van Gulick 1988, 1990) argue that conscious experience enhances cognitive performance by contributing to the efficiency of the underlying mechanisms implicated in the production of behavior. "There is" Flanagan claims, "ample evidence that consciousness facilitates performance on many activities, despite being not absolutely necessary for these activities" (1992, p. 139). Here the issue is not experiential awareness, but the acquisition of knowledge without phenomenal consciousness, and thus about informational sensitivity. To block what he thinks is a mistaken inferential slide from X *is automatic* to X *is unconscious*, Flanagan distinguishes between "the role of

[26]For another example, at the close of the 2000 U.S. Open Tennis Tournament, when Director Tony Trabert asked Marat Safin how he was able to produce the tennis performance that demolished defending champion Pete Sampras, Safin immediately replied, "I don't know."

phenomenal awareness in acquiring knowledge and in deploying that knowledge" (p. 140). The idea is that once learned, one performs activities, for example, properly executing tennis strokes, without the phenomenal awareness that accompanied the lessons. "The true idea that one eventually goes on 'automatic pilot' can give rise to the mistaken thought that one consciously disassociates oneself from the play" (p. 140). The player makes the strokes automatically under the causal control of the consciously acquired technique she was taught.

Part of the problem, Flanagan thinks, is that typical cases of this kind involve sensory and tactile–kinaesthetic experiences rather than verbal or thoughtful experiences. This would account for the player's shoulder shrug and bemused facial expression when asked to explain her outstanding performance. It is a commonplace in world-class sports to refer to athletes as being in "a zone" to *explain* a streak of extraordinary high-quality performance. Obviously the athletes intend their excellent performance even when they fail to deliver as planned. This is not to insist that all conscious experiences are (must be) "effable." The problem is rather to make sense of the claim that the consciously acquired knowledge of how to make a certain tennis stroke at some time t in the past, and is no longer mentally occurrent, can at time $t_1$, cause the production of an instance of the stroke at $t_1$. To causally explain the production of the motions that constitute the tennis stroke it is sufficient to know the disposition of the physical mechanisms involved and relevant dynamical laws. If, on the other hand, the explanation is in terms of player's memory of how to make the tennis stroke in question, then the account is *intentional* in virtue of its reference to the player's representations. Neither situation is compatible with the naturalistic approach to explanation.

The above critique is based on a (Humean) temporally asymmetrical interpretation of causation. It rejects not only future or teleological causation but more controversially disavows the idea of the cotemporal occurrence of causes and effects. Laws of coexistence express this latter notion; for example, the Boyle–Charles ideal gas law (PV = nRT) expresses the simultaneous functional relationship between pressure, volume, and temperature, "but says nothing whatever about the causal relations among them" (Salmon, 1984, p. 136). So a causal–functionalist theory like Flanagan's could claim that neural activities that cross an agent's experience threshold are not thereby *caused* to become conscious or that the agent is caused to be aware of them (albeit it terms of their content). Instead the idea is that, as with the relations expressed by the Boyle–Charles law, crossing the threshold is cotemporal with consciousness. Such regularities, according to Salmon,

> can be *explained* causally, but they do not *express* causal relations. Moreover, they do not afford causal explanations of the events subsumed under them. For this reason, it seems

> to me, their value in providing scientific explanations of particular events is, at best, severely limited. These are regularities that need to be explained, but that do not, by themselves, do much in the way of explaining other phenomena. (1984, p. 186)

The upshot then seems to be that invoking laws of coexistence does little to clarify the idea of the causal potency of consciousness.

The issue is not how to explain the normative quality of one's performance. Rather, the question is about the causal efficacy of mental states implicated in explanations of how actions are guided. This applies, it seems, to "consciously dissociating oneself from the play." There is no plausible doubt about the irresistible force that has us ascribe causal efficacy to ourselves (and by natural extension, to others). The conjecture is that such force derives from an equally irresistible constraint to ascribe moral responsibility, as captured in Flanagan's poignant plea for meaning in our lives. The depth and significance of this concern can hardly be overstated, but the point here is with the putative causal powers of conscious states implicated in typically human action.

It is implicit in Flanagan's point that in implementing a tennis stroke, the player is the agent of the action, with all the rights and privileges to praise and reproach for its execution. Indeed, his functionalist view has him argue that consciousness enhances knowledge acquisition and task performance. "Let me and twin me receive instruction from the same coach, twin me subliminally, me in the normal way, and let us both practice the same amount. Bet on me in the match" (p. 140). The argument here does not contest the view that the intelligibility of one's *practices* presupposes the applicability of these socially grounded concepts. To the contrary, there is an important inference about agency that can be drawn from acknowledging the relevance of such concepts. The emphasis here is that the very concept of agency harbors the mystery of how the agent's consciousness figures in the causal production of her actions.

Consider that at least part of the motivation for Flanagan wanting to ascribe causal efficacy to his conscious states in defeating his twin at tennis is that he, Flanagan, is the tennis-playing agent. If his movements were to have been controlled by some tennis-omnipotent Oscar on the side-lines manipulating a remote control device, the victory over his twin would not only be less sweet, Flanagan might well wonder if the triumph was *his*. Had *he* really vanquished his twin? Can he justly claim the victor's spoils? These are deeply disturbing questions. If he cannot attribute the victory to his own talents, that is, to himself, then he cannot justly enjoy the praise and other fruits of what, in the proprietary sense, now seem not to be his successful efforts at all.

These musings are provoked by a conflation of two concepts of self. One is the conscious self identified with the continuity through one's projects and

memories, the psychological self. This is the conscious self to whom causal powers are ascribed: How could it be Flanagan's tennis triumph unless he, by his conscious effort, brings it about? A satisfying answer to this question requires a satisfactory explanation of the causal efficacy of conscious mental states. Accordingly, this is also the idea of self that proves difficult to pry free from the deeply entrenched Cartesian intuition against which Flanagan cautions. It is the intuition that the relevant causal antecedents to one's actions are, by hypothesis, one's conscious states for which there are no causally relevant physical states other than those under one's conscious control.

The other concept of self is what may be called the neurobiological self. Roughly, this is the self constituted (circumscribed) by the system's biological integrity, its biophysical continuity through time and the overlapping continuities in its social relations with others. This self is identified similarly with the psychological self's projects and memories except that its very same activities are not a source of causal conundrums. They are activities that, in accordance with causal generalizations, flow from the neurobiological processes that underlie one's conscious experiences. This is not a zombie-like self because the idea is that there is an, albeit causally inert inner, experiential life which might well figure in theories of personal identity, self-expression, and rationalizing explanations. Flanagan's causal–functionalism wants to have it both ways. On the one hand, he wants to be able to lay coherent claim to being causally responsible for his tennis results in virtue of his conscious efforts so that such results confer meaning on his activities. On the other hand, the explanation of his bodily behavior involved during the match must be compatible with the naturalistic constraints on causal laws. Cartesian intuitions lurk within the former; *I play, therefore I am.* (cf. Huizinga's *Homo Ludens*, 1938). The dreaded specter of epiphenomenalism skulks about the latter.

The source of the conflation lies with the intentionality of one's mental states and, in particular, the self-representational capacity of mental states. Conscious self-representation is something of an analogue of nonconscious self-monitoring. Over a period of more than ten years, Flanagan (1991, 1992, 1996, 2000) has developed a systematically refined, comprehensive theory of the self. While the details of this sophisticated and insightful theory are beyond the scope of the discussion here, there is one feature that is directly relevant to it. The self in Flanagan's view is not the traditional mysterious or otherwise illusory phenomenological presence that subserves a person's continuous identity through time and space. The self emerges from brain activity in what Dennett (1991, p. 418) has characterized as a "center of narrative gravity." But whereas Dennett argues that this idea is a theoretical *fiction*, Flanagan's concept has the reality of an organizing principle.

> First, self-representing can and often does correctly represent the person. . . . Second, although the narrative self is "an organizing principle," it does have causal efficacy in the way any complex model or set of representations does. Since it has causal efficacy, there is no escaping the conclusion that it is supervenient on multifarious brain processes. Self-representation, even massively deceived self-representation, is causally efficacious — it causes the person to say wildly false things about himself. This is inexplicable unless the self-deceived thoughts that (sometimes) precede and prompt false self-descriptions are realized in the brain. (Flanagan, 1992, pp. 209–210)

Flanagan's conclusion appears inescapable, but the claim that the organizing principle that emerges (or as Dennett puts it, "bubbles up") depends on the *explanatory force* of the supervenience relation. One of the burdens of the argument here has been that the prospect (or promise) of the supervenience concept has weakened from when it was thought to be the "silver bullet" that finally dispatches the Cartesian ghost, to now, when it is said to redescribe rather than resolve the relationship of the mental and physical (Kim, 1998). The appeal to supervenience then does not ground (or articulate) the mentalistic explanation of one's actions for it does not sustain a naturalistic account of the causal efficacy of conscious mental states. Instead, the language of supervenience reformulates the vexed relationship between physicalistic explanans and mentalistic explanandum. For example, the functionalist's explanation of the causal efficacy of the self (as an organizing principle) is in terms of its supervenience on underlying neural activity. This is the commonsense view that Libet's experimental results seem to challenge. While there are genuine limitations on experimental methods that can and in some cases *should* generate skepticism about experimental results, the project of naturalizing consciousness shares naturalism's adherence to the best available empirical data. If the results are taken seriously that neural activity in terms of readiness potentials precede the onset of conscious activity, the case for the causal efficacy of consciousness must explain the negative evidence, consonant with physical causation.

It would seem then that Libet's evidence suggests either that the apparent causal efficacy of conscious states is explained by the real neural activity of the readiness potentials or that consciousness has causal potency but is incompatible with the understanding of causal regularity as a physically explicable phenomenon. This is very bad news on both sides of the dilemma. If conscious causation is apparent then one can empathize with Flanagan's anxiety about making sense of whatever meaning our lives have. Alternatively, if there are (as Chalmers, 1996, argues) genuinely lawful, non-physical, conscious causal powers, then naturalism's commitment to physical causation would seem parochial. Even more puzzling is the idea that exercising conscious control imparts the kind of meaning to one's life upon which to establish moral and epistemic responsibility. For it would seem that our (current?)

ignorance of such purported principles of conscious causation would correspondingly deprive us of the ability to exercise the appropriate and causally reliable conscious control over our behavior. This is the unenviable situation as van Inwagen (1983) portrays the free-will theorist's explanatory dilemma: one is damned by determinism as unable to do otherwise because one lacks causal control of behavior and damned similarly by indeterminism because there are no identifiable causes of it.

## The Functional Role of Consciousness: Understudy

Another set of experiments designed to test for the causal efficacy of conscious reasoning bears directly upon the cogency of the functionalist view. The findings serve to highlight the frailty of functionalist explanations of conscious experience. The February 27, 1997 issue of *Science* carried a brief report on results of recent experiments into epiphenomenalism that, quite remarkably, caught the attention of National Public Radio news in the United States. The research team of Bechara, Damasio, Damasio, and Tranel, (1997) investigated the theory that the ventromedial frontal cortices hold dispositional knowledge and explored the claim of causal efficacy for conscious, occurrent beliefs vis-à-vis corresponding nonconscious, standing beliefs. Although the conscious phenomena under investigation are subjective, there is no further presumption that they have or are accompanied by some sort of phenomenal feel. The Damasio team tested two hypotheses. The first was that "deciding advantageously in a complex situation . . . require(s) overt reasoning on declarative knowledge, namely, on facts pertaining to premises, options for action, and outcomes of actions that embody the pertinent previous experience" (p. 1293). The second was "that overt reasoning is preceded by a nonconscious biasing step that uses neural systems other than those that support declarative knowledge" (p. 1293). The reference to "neural systems other than those that support declarative knowledge" suggests that distinct neural systems subserve the processing of nonconscious, dispositional and conscious, occurrent information quite differently.

In contrast to the non-verbal, tactile–kinesthetic experience in the tennis case, the Damasio team study deals with cases of verbal reporting of thought and thus goes directly to Flanagan's claim of "ample evidence that consciousness facilitates performance." In particular, the experimental evidence appears to support the second hypothesis, "that overt reasoning is preceded by a nonconscious biasing step . . . " (p. 1293). If so, this result would challenge the functionalist's intuition that consciousness makes a causal difference.

The two subject-groups, normal persons and patients with bilateral damage to the ventromedial prefrontal cortex, were tested for decision strategies in complex gambling situations involving a loan of $2000 facsimile United

States bills. Subjects were to draw from four decks of cards (ABCD) with the goal of losing the least and winning the most amount of money. Decks A and B carried rewards of $100 but unpredictably large losses. Decks B and C had rewards of $50 but smaller penalties. Choosing decks A and B leads to overall loss, choosing C and D leads to overall gain. The researchers recorded the subjects' behavioral performance, on "the number of cards selected from the good decks versus the bad decks," their anticipatory skin conductance responses (SCR) before each card selection, and at ten-card intervals, and "the subjects' account of how they conceptualized the game and the strategy they were using" (p. 1293). Bechara et al. found no significant anticipatory skin conductance responses when the four decks were sampled without penalties. "After encountering a few losses in decks A or B (usually by card 10), normal participants began to generate anticipatory SCRs to deck A and B. Yet by card 20, all indicated that they did not have a clue about what was going on." After about fifty cards the normal subjects "began to express the hunch that decks A and B were riskier and all generated anticipatory SCRs whenever they pondered a choice from deck A or B" (p. 1293). Seven of the ten normal subjects reached conceptualization after about eighty cards. "None of the patients generated anticipatory SCRs" (p. 1293) Significantly however, three patients reached conceptualization and correctly distinguished the bad and good decks; nevertheless they chose disadvantageously.

With respect to the question of the causal efficacy of conscious, Bechara et al.'s results are, to be sure, not decisive. Nevertheless, like Libet's findings, they are striking and cry out for an explanation that reconciles the causal efficacy of nonconscious neural events with our irrepressible intuitions of mental causation. The problems of design, even in ingeniously sophisticated experiments on consciousness, are formidable, and as discussed above, the results admit ambiguous interpretations. On this issue, the point is that the functionalist interpretations of the experimental results typically reflect a default attitude toward conscious causation as "guilty (consciousness is causally efficacious) until proven innocent (i.e., causally impotent)." The findings are notable because they seem to sustain rather than allay epiphenomenalist suspicions (skepticism) about conscious causation. These suspicions are reinforced insofar as "nonconscious signals . . . act as covert biases on the circuits that support processes of cognitive evaluation and reasoning" (p. 1294). The evidence suggests that a nonconscious biasing using "neural systems other than those that support declarative knowledge" precedes overt reasoning. The functionalist view has conscious occurrent states enhancing nonconscious cognitive performance. These striking experimental results suggest quite the reverse. And in some perverse and comforting way they help to explain the hapless tennis hacker's frustration in her failure to impart topspin to the ball despite her conscious efforts.

## References


Bechara, A., Damasio, A., Damasio, H., and Tranel, D. (1997, February 28). Deciding advantageously before knowing the advantageous strategy. *Science, 275*, 1293–1294.

Block, N. (1986). Advertisement for a semantics for psychology. In P. French, T. Uehling Jr., and H. Wettstein (Eds.). *Studies in the philosophy of mind, 10, Midwest Studies in Philosophy* (pp. 615–678) Minneapolis: University of Minnesota Press.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18*, 227–287.

Botterill, G., and Carruthers, P. (1999). *The philosophy of psychology*. Cambridge: Cambridge University Press.

Burge, T. (1979). Individualism and the mental. In P. French, T. Euhling, and H. Wettstein (Eds.), *Studies in epistemology, 4, Midwest Studies in Philosophy* (pp. 73–121). Minneapolis: University of Minnesota Press.

Burge, T. (1986). Individual and psychology. *Philosophical Review, 95*, 3–46.

Chalmers, D. (1996). *The conscious mind*. Oxford: Oxford University Press.

Chisholm, R.M. (1976). *Person and object*. London: George Allen and Unwin.

Churchland, P.M. (1995). *The engine of reason and the seat of the soul*. Cambridge, Massachusetts: MIT Press.

Dennett, D.C. (1971). Intentional systems. *Journal of Philosophy, 68*, 87–106.

Dennett, D.C. (1978). Skinner skinned. In D.C. Dennett (Ed.), *Brainstorms* (pp. 53–70). Montgomery, Vermont: Bradford.

Dennett, D.C. (1991). *Consciousness explained*. New York: Little Brown.

Dennett, D.C., and Kinsbourne, M. (1992). Time and the observer: The where and when of consciousness in the brain. *Behavioral and Brain Sciences, 15*, 183–247.

Flanagan, O. (1991). *Varieties of moral personality: Ethics and psychological realism*. Cambridge, Massachusetts: Harvard University Press.

Flanagan, O. (1992). *Consciousness reconsidered*. Cambridge, Massachusetts: MIT Press.

Flanagan, O. (1996). *Self-expressions*. Oxford: Oxford University Press.

Flanagan, O. (2000). *Dreaming souls: Sleep, dreams, and the evolution of the conscious mind*. New York: Oxford University Press.

Fodor, J. (1974). Special sciences (or the disunity of science as a working hypothesis). *Synthèse, 28*, 97–115.

Fodor, J. (1987). *Psychosemantics*. Cambridge, Massachusetts: MIT Press.

Goldman, A. (1993). The psychology of folk psychology. In A. Goldman (Ed.), *Readings in philosophy and cognitive science* (pp. 347–380). Cambridge, Massachusetts: MIT Press.

Huizinga, J. (1938). *Homo ludens*. Haarlem, The Netherlands: Tjeenk Willink.

Jackendoff, R. (1987). *Consciousness and the computational mind*. Cambridge, Massachusetts: MIT Press.

Jackson, F. (1982). Epiphenomenal qualia. *Philosophical Quarterly, 32*, 127–136.

James, W. (1890). *Principles of psychology* (three volumes). Cambridge, Massachusetts: Harvard University Press.

Kay, A. (1984). Computer software. *Scientific American, 251*, 52–59.

Kim, J. (1993). *Supervenience and mind*. Cambridge: Cambridge University Press.

Kim, J. (1998). *Mind in a physical world*. Cambridge, Massachusetts: MIT Press.

Lackner, J, and Garrett, M. (1973). Resolving ambiguity: Effects of biasing context in the unattended ear. *Cognition, 1*, 359–372.

Levine, J. (1982). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly, 64*, 27–40.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action *Behavioral and Brain Sciences*, 8, 529–566.

Libet, B. (1991). Conscious functions and brain processes. *Behavioral and Brain Sciences, 14*, 685–686.

Loftus, E.F. (1979). *Eyewitness testimony*. Cambridge, Massachusetts: Harvard University Press.

Lycan, W. (1981). Form, function, and feel. *Journal of Philosophy, 78*, 24–50.

Lycan, W. (1987). *Consciousness*. Cambridge, Massachusetts: MIT Press.
Lycan, W. (1997). *Consciousness and experience*. Cambridge, Massachusetts: MIT Press.
MacKay, D.M. (1985). Do we "control" our brains? *Behavioral and Brain Sciences, 8*, 546–547.
MacKay, W. (1991). Consciousness is king of the neuronal processors. *Behavioral and Brain Sciences, 14*, 687–688.
McGinn, C. (1989a). *Mental content*. New York: Basil Blackwell.
McGinn, C. (1989b). Can we solve the mind–body problem? *Mind, 98*, 349–366.
McGinn, C. (1996, April 5). Wise incomprehension, Review of Chalmers' *The Conscious Mind*. In *The Times Higher Education Supplement*, pp. vii and ix.
McGinn, C. (1999). *The mysterious flame: Conscious minds in a material world*. New York: Basic Books.
Nagel, T. (1974). What it is like to be a bat. *Philosophical Review, LXXXIII, 4*, 435–450.
Nørretranders, T (1998). *The user illusion*. New York: Penguin Putnam. (Originally published in Danish as *Maerk verden*, 1991.)
Salmon, W. (1984). *Scientific explanation and the causal structure of the world*. Princeton: Princeton University Press.
Seager, W. (1995). Consciousness, information, and panpsychism. *Journal of Consciousness Studies, 2*, 272–288.
Silvers, S. (1997). Nonreductive naturalism. *Theoria, 12*, 163–184.
Sternberg, S. (1996). High-speed scanning in human memory. *Science, 153*, 652–654.
van Gulick, R. (1988). A functionalist plea for self-consciousness. *Philosophical Review 97*, 149–188.
van Gulick, R. (1990). What difference does consciousness make? *Philosophical Topics, 17*, 211–230.
van Gulick, R. (1991). Consciousness may still have a processing role to play. *Behavioral and Brain Sciences, 14*, 699–670.
van Gulick, R. (1995). What would count as explaining consciousness? In T. Metzinger (Ed.), *Conscious experience* (pp. 61–79). Thorverton, United Kingdom: Imprint Academic.
van Inwagen, P. (1983). *An essay on free will*. Oxford: Clarendon.
Velmans, M. (1991). Is human information processing conscious? *Behavioral and Brain Sciences, 14*, 651–726.
Wegner, D.M. (2002). *The illusion of conscious will*. Cambridge, Massachusetts: The MIT Press.
Wells, G.L., and Murray, D.M. (1984). Eyewitness confidence. In G.L. Wells and E.F. Loftus (Eds.), *Eyewitness testimony: Psychological perspectives* (pp. 155–170). Cambridge: Cambridge University Press.
Wilson, T.D. (2002). *Strangers to ourselves*. Cambridge, Massachusetts: Harvard University Press.