Behavioral Sciences & the Law
Behav. Sci. Law **25**: 281–293 (2007)
Published online 23 February 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/bsl.743

The Concept of Free Will: Philosophy, Neuroscience and the Law

Susan Pockett, M.Sc., Ph.D.*

Various philosophical definitions of free will are first considered. The compatibilist definition, which says simply that acts are freely willed if they are not subject to constraints, is identified as much used in the legal system and essentially impervious to scientific investigation. A middle-ground ''incompatibilist'' definition, which requires that freely willed acts be consciously initiated, is shown to be relevant to the idea of *mens rea* and in the author's view not actually incompatible in principle with a fully scientific worldview. Only the strong libertarian definition, which requires that freely willed acts have no physical antecedents whatsoever, makes the existence of free will very hard to swallow scientifically. However, with regard to the middle-ground ''incompatibilist'' definition, three different lines of scientific experimental evidence are then described, which suggest that, in fact, consciousness is *not* the real cause of much of what is generally considered as voluntary behavior. Many voluntary actions are initiated preconsciously, with consciousness kept informed only after the neural events leading to the act have begun. It is suggested that a reasonable way of integrating these experimental findings with the idea that persons do have a somewhat more than compatibilist version of free will is to acknowledge explicitly that a person is a mixture of conscious and unconscious components. In this scenario, the mind in *mens rea* would have to be judged guilty if it contained either conscious *or unconscious* intentions to perform the guilty act. Copyright © 2007 John Wiley & Sons, Ltd.

THE PHILOSOPHY OF FREE WILL

This article is not primarily about philosophy, but in order to write about the neuroscientific and legal aspects of free will it is necessary to begin by making at least

*Correspondence to: Susan Pockett, M.Sc., Ph.D., Department of Physics, University of Auckland, Private Bag 92019, Auckland 1020 New Zealand. E-mail: s.pockett@gmail.com

some attempt to define what is generally meant by the words ''free will''. Until recently, exclusive rights to this area in the academic sense have been enjoyed by the discipline of philosophy—which means that there are probably as many definitions of free will in the literature as there are philosophers. Broadly speaking, however, philosophical definitions of free will seem to fall into three categories.

(1) Many, or even most, present-day philosophers are so-called *compatibilists*. Compatibilists define free will in a very weak sense, basically claiming that as long as nobody is holding a gun to one's head, one's actions can be said to be freely willed. This position is called compatibilism because it is seen as rendering the existence of free will compatible with the existence of determinism. (Determinism is the view that every event in the physical world is either caused by some preceding physical event, or random—i.e. not caused at all).
(2) One opposing position is called *incompatibilism*. Incompatibilists concentrate on the fact that the physical world (even the world as described by quantum mechanics, although misunderstandings on this point are rife) *does* seem to be deterministic, which is taken as meaning that every non-random physical event in the brain must be the direct result of some preceding physical event in the brain. Incompatibilists perceive this latter point as posing a serious problem for the idea of free will, because
(a) they define free will as requiring the causation of voluntary bodily movements by a conscious self;
(b) they see no room in the causal chain of brain events for consciousness (the reasoning here being that one neural event is the necessary and sufficient cause of the next neural event and it is not possible for there to be two causes of one event);
(c) therefore, they conclude, free will can not exist.

The overall position of the incompatibilist is that determinism, taken to its logical conclusion, implies that everything we do throughout our lives is pretty much predestined at birth. A degree of randomness might alter the course of events somewhat, but ''we'', in the sense of conscious selves who can break into the causal chain of brain events to cause actions, can not in fact change anything. The idea that we can do so is seen as nothing but an illusion—an illusion with some evolutionary and social usefulness perhaps, but an illusion nonetheless. In short, incompatibilists believe that free will is incompatible with determinism and determinism is a fact, so they are stuck with a world in which there is no such thing as free will.

(3) The third and most extreme position on the definition of free will is that of the strong *libertarian*. Libertarians require that a freely willed action involves truly *originated* conscious commands. A truly originated command has no physical antecedents whatsoever. Thus in order to believe in the existence of a libertarian free will, one must necessarily be a full-blown dualist with regard to the nature of consciousness—that is to say, one must regard consciousness as being a non-physical phenomenon.

Of these three positions, libertarianism is probably the least interesting from a scientific point of view. There are two reasons for particularly a biological scientist to feel uncomfortable with the existence of the libertarian version of free will. First, libertarians essentially have to be dualists in respect of the nature of consciousness, and dualists face the perennial problem of how a non-physical consciousness could

DOI: 10.1002/bsl

interact with a physical brain. Second, the continued production of truly originated conscious commands, in the sense of commands that arise without any physical antecedents at all, is likely to be considerably disadvantageous in a biological sense. It would lead to behavior so erratic and unrelated to the current environmental conditions that anyone producing such behavior would fairly quickly land in a psychiatric institute at best, or an early grave at worst. Thus, it is difficult for a biologist to believe in the existence of a libertarian free will.

Of the other two definitions, incompatibilism is by far the more interesting from a scientific point of view. Incompatibilists at least confront head-on the difficult issues of cause and effect, while compatibilists merely side-step the problem. However, from the same scientific point of view, there do exist certain problems with the incompatibilist position. First, the attentive reader may have picked up on the phrase in 2b above ''and it is not possible for there to be two causes of one event''. In fact, of course, it is perfectly possible for there to be multiple contributing causes for one event. Indeed, in neurophysiological terms, a single synaptic input is almost never sufficient to cause a postsynaptic nerve cell to fire—the summation of many synaptic inputs is generally an absolute requirement for the output of a single action potential. Thus there is actually plenty of room for not only a number of preceding neural events, but also for consciousness (whatever that is conceived as being) to contribute to the causation of any particular observable behavior. In this view, consciousness could certainly alter the weighting of different neural events and thereby the probability of a particular behavior's occurrence, and it might conceivably even cause the occurrence of a behavior that would otherwise not have happened at all. So this part of the incompatibilist's chain of reasoning is definitely suspect.

The second problem is that the incompatibilist defines free will as involving causation by consciousness, which is fine—but it then makes the very important covert assumption that *consciousness is in some way distinct from the brain*. If consciousness were actually to turn out to be identical with some aspect of brain structure or function, a large part of the problem seen by incompatibilists would disappear. Consciousness would then simply be a normal part of the deterministic causal chain and no issue of incompatibility would arise.

But should the idea that consciousness is identical with some aspect of brain structure or function be taken seriously? Our everyday intuitions are unabashedly dualist: it simply seems obvious that conscious experiences like ''red'' or ''middle C'' are utterly different in kind from material bodies such as brains. Oddly enough, though, the standard working hypothesis of biological scientists at the beginning of the 21st century actually *is* that consciousness is identical either with certain neural activity in the brain (neurophysiologists) or certain functions in or processes of the brain (psychologists). The former position is called psychoneural identity theory, neural identity theory, or, more spectacularly, ''The Astonishing Hypothesis'' (Crick, 1994), and the latter position is called functionalism. So for the standard early 21st century neurophysiologist or psychologist, there is actually no incompatibility between determinism and the philosophical incompatibilist's idea of free will.

At this point I must admit that personally I find both the neural identity theorist's and the functionalist's ideas on the nature of consciousness unsatisfying to the point of being incomprehensible, but I also dislike the idea that Cartesian dualism could be the answer. I am a scientist, and dualism (the notion that consciousness is

DOI: 10.1002/bsl

non-physical) implies that consciousness must be epiphenomenal (unable to cause any physical events) and thus inaccessible to science.[1] While it is not inconceivable that consciousness will ultimately turn out to be inaccessible to science, I choose (freely or otherwise) to believe that this is not the case. My own solution is to adopt the intermediate position that consciousness is identical not with the brain itself, but with a field generated by the brain. Benjamin Libet, whose name will feature prominently in the next section of this article, proposes that consciousness is a non-physical (dualist) field, which can not cause behavior but can, by some unknown means, veto it (Libet, 1994). Leaving aside the perennial dualist's problem of *how* a non-physical entity could cause any physical effect, if one simply assumes that it can do so, one has no problem accepting that the same non-physical entity can violate determinism. Thus the idea of consciousness as a non-physical field does effectively rescue the notion of an incompatibilist free will. For myself, as already remarked I baulk at dualism, so my own suggestion is that consciousness may be a particular kind of spatiotemporal pattern in the electromagnetic field generated by the active brains (Pockett, 2000, 2002). In this hypothesis, consciousness is still part of the physical world, which means that it still obeys the deterministic laws governing the physical world, which means that, again, there is actually no incompatibility between determinism and ''incompatibilist'' free will. In the electromagnetic field hypothesis of consciousness, consciousness is perfectly capable *in principle* of causing the neural activity that underlies behavior, although, as will be seen, the evidence suggests that it if it actually does so at all, it does so much less frequently than we imagine (Pockett, 2004; Pockett, Banks, & Gallagher, 2006).

The upshot of all this is that for *any* currently active species of biological scientist—be they neural identity theorist, functionalist, or field theorist—there is no theoretical problem with the suggestion that conscious free will might exist. Or at least, there is no problem provided one is not a member of the third and most extreme definitional group, the libertarians.

To summarize the discussion so far, we have the following.

1. Philosophical *compatibilists* define free will in such a way that science is irrelevant. They concentrate purely on whether or not there were constraints on a particular action or whether the actor was ''free'' to choose his own course. Constraints in this sense can be either external or internal. Not only a gun-wielding maniac but also one's own ''madness'' can serve to remove an action from the purview of free will under this definition (which is what makes it reasonable for those jurors who are either consciously or unconsciously operating on a compatibilist definition of free will to find a perpetrator not guilty by reason of insanity).
2. Philosophical *incompatibilists* do take science into account in that they focus on the scientific *sine qua non* of determinism. However, the incompatibility they then perceive between determinism and conscious free will seems, at least to me, to result largely from the adoption of a linear view of biological causation and a dualist view of the nature of consciousness, both of which are quite at odds with

[1]It is important to clarify at this point that the words dualism and epiphenomenalism do not mean the same thing. A dualist or non-physical consciousness would necessarily be completely epiphenomenal (i.e. completely unable to cause events in the physical world), but a physicalist consciousness, while *in principle* capable of causing events in the physical world, could also be at least partly epiphenomenal from a functional point of view.

DOI: 10.1002/bsl

the current scientific *Zeitgeist*. In a fully scientific world view, it seems to me that there is actually no incompatibility between the fact of determinism and the idea that we can consciously initiate and control our own actions.

3. Philosophical *libertarians* sit somewhat uncomfortably at the other extreme of the definitional spectrum from compatibilists. Libertarians define free will as involving only genuinely originated acts, which have no physical antecedents whatsoever. Whether or not a libertarian concludes on the basis of this definition that free will does or does not exist depends on whether he or she can or can not accept dualism as a viable theory of the nature of consciousness. Some can, most can not.

My own position falls somewhere in the middle of this definitional spectrum. I have no problem with the compatibilist definition—except that it does not seem really to address the issue. So I am not a compatibilist. However, in my theory of consciousness, there actually is no incompatibility between determinism and the mid-spectrum ''incompatibilist'' definition of conscious free will—provided, that is, that one does not require free will to involve genuinely originated, *de novo* acts. For myself, I do not define free will as involving acts that are wholly uncaused by preceding physical events. I am a scientist by trade, and science has never coped well with the idea of unmoved movers. Even for Aristotle, the only completely unmoved mover was God. We mortals make decisions about whether or not to act on the basis of preceding events, and in a present-day biologist's world view, I believe there is no in-principle problem with the idea that such decisions can be the result of conscious free will.

THE NEUROSCIENCE OF FREE WILL

So if there is no theoretical problem with the idea of non-originating free will, do scientific experiments actually *support* the idea that non-originated decisions—or indeed any of the three kinds of conscious free will—actually exist? Strangely enough, they do not.

Libet's Experiments

On any scientific definition of causation, causes can not operate backwards in time. Thus one obvious means of determining whether consciousness may have caused the neural events that culminate in a particular voluntary body movement is to ask whether the conscious decision to move occurred before or after the beginning of the neural events. If the conscious decision happens before the neural events begin, then the conscious decision may (or may not) be the cause of the neural events. But if the conscious decision arises only *after* the beginning of the neural events, there is absolutely no way that the conscious decision can be said to have caused the neural events. On the contrary, it is more likely that the neural events caused the conscious ''decision''.

The now famous experimental protocol first published by Libet, Gleason, Wright and Pearl (1983) shows, quite repeatably and unequivocally, that human subjects

DOI: 10.1002/bsl

become aware of what they perceive as their conscious decision to initiate a simple finger movement before the movement actually occurs, but considerably *after* the start of the neural activity leading to the movement. In the Libet protocol, each subject is asked to conduct a number of trials in which they report the position of a revolving spot of light at the instant they subjectively decide to move their finger to press a key. Simultaneously, the subject's brain waves (EEG) are measured by electrodes on the scalp. The EEG is stored in a buffer and each key press is used as a trigger on which to back-average over 40-odd trials the EEG occurring in the two to three second period immediately before the finger movement. This averaging procedure reveals a long, slow (at least slow in neurophysiological terms), negative-going wave-form called a readiness potential or *Bereitschaftspotential* (Kornhuber & Deecke, 1965), which starts at some time between 500 and 1000 ms before the finger movement and culminates roughly around the time of movement. The reported time of deciding, ''wanting'' or ''wishing'' to move (W in Libet's terminology) has been measured at anywhere from 200 ms before the start of a finger movement as measured by electrodes on the arm muscles (Libet *et al.*, 1983) to 120 ms before a keypress signifying the end of the movement (Trevena & Miller, 2002)—or in the particular case illustrated in Figure 1 (Pockett, unpublished data) 89 ms before keypress.

It is clear from Figure 1 that the neural events underlying the readiness potential start anything up to a full second before the subject is aware of consciously ''deciding'' to move. A second may not seem very long, but in neurophysiological

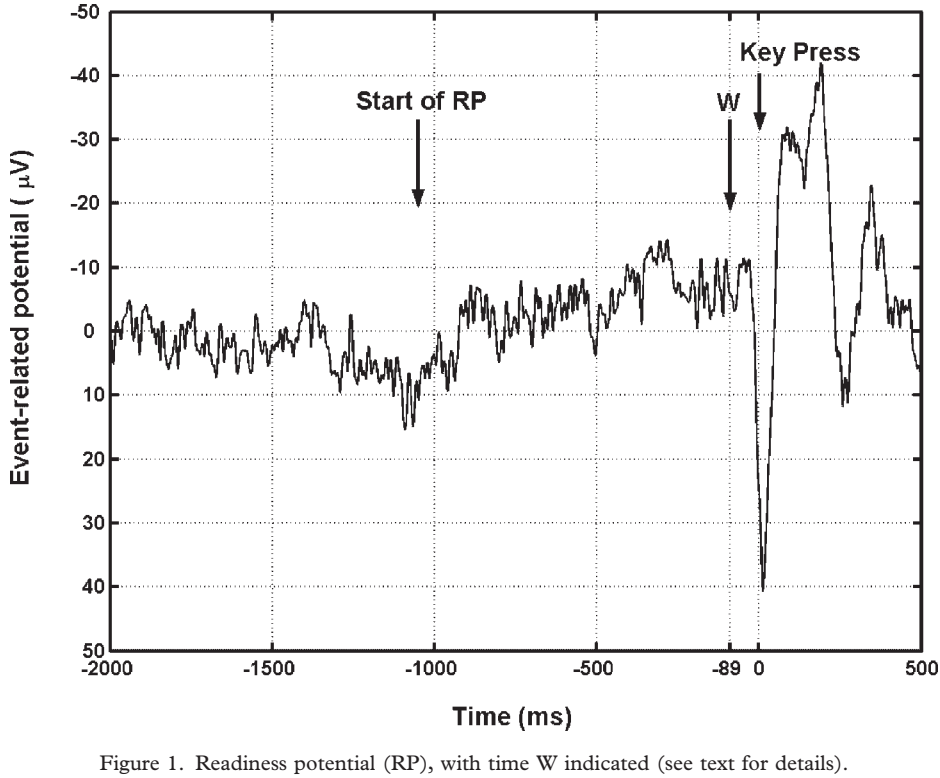

Figure 1. Readiness potential (RP), with time W indicated (see text for details).

DOI: 10.1002/bsl

terms it is an eternity. To put this in perspective, the unit of action in brain cells is an electrical event called an action potential: One action potential lasts 2–3 ms (milliseconds) and a second contains 1000 ms. In any case, the point is that, at least for this very simple kind of voluntary movement, the subjectively perceived decision[2] to move is actually *not* the cause of what the subject thinks of as a consciously caused movement.

Does this result imply that consciousness does not cause behavior in general and therefore that we do not have conscious free will? Libet himself refused to accept such an implication, reasoning that although the conscious decision could not have been the cause of the movement in this simple case, it did arise before the movement, which meant that there was still time for consciousness to *veto* the movement before it took place. Unfortunately it is not possible to back-average off a movement that never takes place, so there is no empirical support for this suggestion. It seems that most professional philosophers (with the notable exceptions of Galen Strawson and Ted Honderich) also find the idea that there is no such thing as free will too unpalatable to contemplate. For example, various philosophers have tried to avoid the conclusion that the Libet experiments imply the illusory nature of free will by arguing that the movements in Libet's paradigm are far simpler than those to which the term free will should reasonably be applied (Gallagher, 2006), that Libet's subjects reported only an ''urge'' to move, when an urge is not a decision (Mele, 2006), that Libet's paradigm depends on introspection, when everyone knows that introspection is fallible (Ross, 2006), that it is not only the initial element in a causal chain that can be considered causal (Pacherie, 2006)—and any number of other excuses. However, as far as I can see, there is no wriggling out of the fact that, at least in the case of the very simple movements studied in the Libet paradigm, an act that the *subject* perceives as being entirely voluntary is probably initiated not by what the subject thinks is a conscious decision to act, but by some other, unknown, preconscious, neural event(s). The awful truth is that, at least in this very simple case, the subject routinely gets it wrong. And if we can get it wrong in the simplest possible case, it seems to me there is more than a passing chance that we also routinely get it wrong in more complex cases.

Wegner's Experiments

Daniel Wegner espouses the model of volition illustrated in Figure 2. According to this model, both the thought preceding a voluntary action and the action itself are actually generated, in parallel, by separate unconscious processes. However, sometimes (if the requirements listed below are adequately fulfilled) we automatically but erroneously *infer* a causal path from thought to action. The requirements for this incorrect inference to be made are the following:

(a) The thought must take place immediately before the action (the priority principle).
(b) The thought must be consistent with the action (the consistency principle).

[2]In the particular case illustrated in Figure 1 the subject actively decided to make each movement, rather than ''allowing the urge to move to arise spontaneously'', as per Libet's original instructions.

DOI: 10.1002/bsl

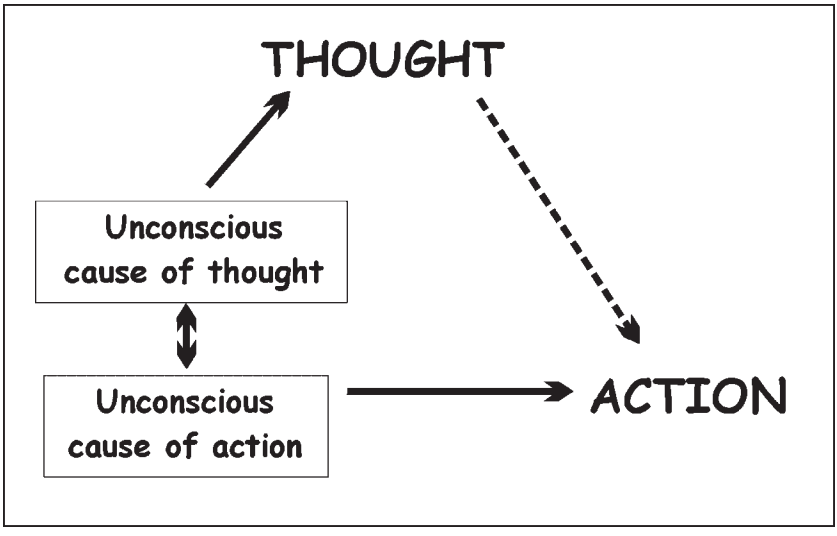

Figure 2. Wegner's model of the relationship between conscious and unconscious processes and the generation of voluntary actions. Solid arrows represent actual causal pathways. Dotted arrow represents erroneously inferred causal pathway.

(c) The thought must be the only apparent cause of the action (the exclusivity principle).

What is the empirical support for this model? Wegner and colleagues conducted two sets of experiments to test whether an observably incorrect inference about the actual cause of an event is made if one of the three listed principles is fulfilled. In the first experiments (Wegner & Wheatley, 1999) subjects were instructed to move a cursor randomly around a computer screen filled with objects and every half minute or so to stop the cursor over an object. It was made clear to the subjects that the experimenter was also capable of stopping the cursor, and after each stop they were to report an ''intentionality'' rating, on a scale from 100% (completely sure they had caused the stop themselves) to 0% (completely sure the experimenter had caused the stop). Surprisingly, when the subjects really had caused the stops, they were not at all sure that they had—the average intentionality rating was only 56%. When stops were actually caused by the experimenter, the rating was exactly the same—56%—if the subject heard the name of the object either five seconds or one second before the stop, and slightly biased towards the experimenter's having causing the stops (52%) if the subject heard the name of the object 30 seconds before or one second after the stop. These results were interpreted as showing that if the subjects were simply led to think about the object over which the cursor stopped at an appropriate time (1–5 s) before the stop, they inferred that their thought had caused the cursor to stop. On the other hand, if they were led to think about the object either after the stop or before the stop but too far away from it in time, they were less likely to infer that they had caused the stop. This was taken as evidence supporting the influence of the priority principle.

The second set of experiments involved subjects viewing other people's gloved hands in a mirror, with the other people's hands in the position where the subjects'

DOI: 10.1002/bsl

own hands would normally be. The subjects were asked to rate on a scale from 1 to 7 whether they thought they had controlled the actions of the hands not at all (1) to very much (7). Again, in the baseline condition the subjects were not completely sure about whether or not they had controlled the movements of what were actually the experimenters' hands, scoring 2.05 ± 1.61 (SD). When the real owner of the hands was moving them according to a set of instructions that were also heard by the subjects, the average control score rose to 3.00 ± 1.09 (SD), a significant increase. This is cited by Wegner (2002) as evidence for the influence of the consistency principle: If one's thoughts are consistent with an event, the tendency is to believe that one caused the event. For example, if one wishes someone dead and then they actually do die (for a completely unrelated reason), one is likely to feel guilt.

Of course, as with Libet, the philosophical literature contains a number of attacks on Wegner. These turn out essentially to be attacks on the possibility that experimental evidence can lead to any firm conclusions about the world at all. For example, Peter Ross (2006) says that Wegner's experimental example of illusory control does not impact on the question of free will because it studies an abnormal situation, ''not relevant to the ordinary example of free will, that is, the case where one's intentions clearly do cause one's actions'' (Talk about begging the question!) In any case, Ross says, all the libertarian claims is that ''*not all* types of control are systematically illusory—that is, that there are also some types of control.'' Wegner's many cited examples of cases where the subjects imagine their will is not the cause of their actions when in fact it *is* the cause are deemed to be no good either, again because they involve unusual or pathological situations. So, according to Ross, Wegner's experiments are of little use because, while they do show that feelings of control are sometimes illusory, (a) they do not show that these feelings are always illusory and (b) they involve abnormal situations. Of course this poses quite a problem for experimental science. On top of the perennial problem of induction,[3] any experimental situation at all could be defined as *abnormal* and therefore inadmissible. In light of these objections, it is perhaps remarkable that scientific investigation in general has been as successful as it clearly has been in understanding the world.

Elisabeth Pacherie (2006) adopts a stance similar to that of Ross. With regard to Wegner's experiments, she concludes that ''As we have independent reasons to think that conscious intentions (in the first-order sense) are causally efficacious in the production of actions and no good reasons to think that our second-order awareness of intentions is always or most of the time the result of a misidentification of mere thoughts with actual intentions, we can, I think, remain confident that the experience of conscious will is a reliable indicator of actual mental causation'' (p. 165). Pacherie does not itemize any of the, ''independent reasons to think that conscious intentions... are causally efficacious in the production of actions''—so one can only assume that she is talking about the standard folk-psychological intuition that because conscious intentions sometimes appear at roughly the same time as actions, the intentions must be causal for the actions. Of course, this ignores the fact that it is exactly the possibility that this intuition is erroneous which is under examination in Wegner's experiments. The ''no good reasons to think that our second-order

[3]The problem of induction is that it is logically impossible to prove by collecting a finite number of examples that anything either always happens or never happens.

DOI: 10.1002/bsl

awareness of intentions is always or most of the time the result of a misidentification of mere thoughts with actual intentions'' comment seems to conflate the ''have not shown it's *always* true'' argument that Ross adduces with the experimental substitution of thoughts for intentions.

Timothy Bayne's complaint about Wegner's work (Bayne, 2006) is slightly different. Bayne's message is that the science is incomplete. He says that both the mental phenomenology and the neuroscience of action are still far from well understood, and until we have a fuller understanding of these matters we should not make pronouncements one way or the other about ''the will''. There is, of course, a good deal to recommend this traditionally scientific, humble-toiler-in-the-vineyard position (although in this case it does seem to fail to do justice to what *has* been shown). Amusingly, though, the moral high ground on which this particular vineyard is planted eventually proves an irresistable pulpit from which to make Bayne's own final pronouncement, that it is ''highly unlikely that the phenomenology of agency is systematically misleading. We experience ourselves as agents who do things for reasons, and there is little serious reason to suppose that we are mistaken on either count.'' No evidence at all is quoted in support of this bold statement. If one refuses someone else permission to make pronouncements on a particular matter because too little evidence is available to support conclusions, surely it is only sensible to remain silent oneself.

My own conclusion is that, while it is clear that none of Wegner's results constitutes completely convincing evidence that conscious thoughts *never* cause actions, his data do tend to support the hypothesis outlined in Figure 2. As Wegner and Wheatley (1999) put it, ''Because we have thoughts of what we will do, we can develop causal theories relating those thoughts to our actions on the basis of priority, consistency and exclusivity. We come to think of these prior thoughts as intentions, and we develop the strong sense that the intentions have causal force even though they are actually just previews of what we may do'' (p. 490). To reiterate, it is not clear that the strong sense that thoughts cause actions is *always* in error: As Wegner and Wheatley say, ''the experience of will *can* [my emphasis] be an indication that mind is causing action... but it is not conclusive'' (p. 490).

Jeannerod's Experiments

Marc Jeannerod (Jeannerod, 2006) summarizes a number of experiments showing that the ongoing control of voluntary actions is often unconscious and that even relatively complicated actions can be initiated, carried through and completed before the fact that they have been done enters consciousness.

In one such experiment (Castiello, Paulignan, & Jeannerod, 1991), subjects had to reach for an object placed in front of them as soon as it became illuminated and also to report verbally as soon as they became aware first of the onset of illumination and then of any changes in the object. The result with regard to the initiation of the reaching movement was that the movement began about 50 ms before the verbal report of illumination. When the illuminated object was moved by the experimenter at the time the reaching movement began, the first sign of correction of the hand trajectory appeared about 100 ms after the shift in target position, and the verbal report on the shift in target position came about 300 ms after the beginning of the change in movement trajectory. In fact, the subjects reported that they only saw the object jump

DOI: 10.1002/bsl

to its new position just as they were about to grasp it (or sometimes even after they grasped it). So it appears that movements like this are both initiated and corrected preconsciously. The conclusion is that fast, accurate movements are executed automatically.

In a second series of experiments, Fourneret and Jeannerod (1998) instructed subjects to draw straight lines between a starting position and a target, using a stylus on a digital tablet. The subjects could see representations of both the target and the output of the stylus on a computer screen, but could not see their hand. When the line on the screen was caused by the experimenter to deviate slightly from the line actually drawn, the subjects were able to compensate for the deviation and reach the target, but reported that their hand had moved in the direction of the target rather than the way it had actually moved. In other words, they paid more attention to the visible aspect of their performance than to proprioceptive (body position) cues. However, in a subsequent experiment (Slachevsky, Pillon, Fourneret, & Pradat-Diehl, 2001), it was found that gradually increasing the degree of experimenter-induced deviation seen on the computer screen eventually resulted in accurate reports that the hand movements had actually been in a direction different from the target. This shows that *conscious awareness* of a discordance between an action and its consequences emerges only when the magnitude of the discordance becomes large enough. The point is confirmed in later experiments by Knoblich and Kircher (2004). Everyday examples of the phenomenon are legion—for example, during a common act such as a more or less automatic reaching for a cup while one is doing something else, one becomes fully aware of the cup-reaching movement only if something goes awry and the action has to be radically modified.

The model Jeannerod proposes to explain this phenomenon is similar to the old "efference copy" model from the cybernetic era (Sperry, 1950; von Holst & Mittelstaedt, 1950). In the efference copy model, for every outgoing motor command, a copy (the efference copy) of the command is also generated. This copy is compared with the visual or proprioceptive feedback signals generated by the results of the action, so that any mismatch between the desired and actual results can be used to generate a correction in the outgoing command. In more recent versions of this model, the concordance between outgoing and incoming signals is postulated to be the factor that identifies self- (as opposed to other-) generated changes and thus allows the recognition of oneself as the agent of a particular action (Frith, Blakemore, & Wolpert, 2000). Jeannerod's suggestion is that a version of the same model could account for mismatch-created consciousness, which is possibly signified by increased activity in a small area of the posterior parietal cortex on the right side of the brain (Farrer, Franck, Georgieff, Frith, Decety, & Jeannerod, 2003).

Jeannerod's overall conclusions about consciousness and action (Jeannerod, 2006) are that

> the role of consciousness [is] to ensure the continuity of subjective experience across actions which are—by necessity—executed automatically. Because it reads behavior rather than starting it, consciousness represents a background mechanism for cognitive rearrangement after the action is completed, e.g. for justifying its result or modifying the factors that have been at the origin of the movement if the latter turned out to be unsuccessful. In line with the idea proposed by Nisbett and Wilson (1977) that we tend to 'tell more than we can know', this mechanism could have the role of establishing a declarative cognitive content about one's own preferences, beliefs or desires (p. 37).

DOI: 10.1002/bsl

Oddly, the philosophical community has not so far expressed any opinion on Jeannerod's conclusion.

NEUROSCIENCE AND THE LAW

As already mentioned, one can never conclude using the method of induction that anything *always* or *never* happens. Even after one has encountered 1,000 white swans there is always the chance that one will one day visit Australia or New Zealand and bump into a black swan. However, it is clear that what scientific evidence is available in the area we are discussing does tend to suggest that, contrary to popular belief, consciousness neither initiates nor modifies so-called voluntary actions, but only carries out a monitoring function. Perhaps this will eventually turn out to be a general finding. Perhaps consciousness never causes behavior. The next interesting question must be, ''Should the legal system care?''.

I have already pointed out that science can have little to say about the truth or otherwise of compatibilism. Compatibilism is a definition. Compatibilists choose to define free will in such a way that neuroscience, or in fact any kind of science, is irrelevant. Compatibilism is not interested in how a behavior is caused—it simply states that, in the absence of external (and arguably also internal) compulsion, acts are said to be freely willed. In intellectual terms this is a relatively weak definition. But it is a definition that is undeniably useful in the day to day conduct of affairs, and it is probably what most judges, lawyers and jurors mean (at least initially) when they use the phrase ''free will''. If this *is* the most relevant definition of free will in a legal sense—if all the law is interested in is the presence or absence of constraints on behavior—then neuroscientific experiments have little relevance in the courtroom.

However, fortunately or unfortunately, this is not all the law is currently interested in. There is also the issue of *mens rea*—the guilty mind. In order for an act to be fully culpable, it has to be initiated deliberately *and consciously*. Otherwise, the act is held to be either an accident or an automatism, and therefore at the very least less culpable. So now the question becomes, ''Was the offending behavior consciously caused?''. More generally, can consciousness ever be the immediate cause of behavior? This latter question is the sort that is accessible to scientific investigation, and as shown above, science has begun to investigate it.

At this stage, the results of the investigation must be said to be inconclusive, but they are only inconclusive in the sense that it is too soon to say that consciousness *never* causes behavior. I believe it is quite reasonable, even at this stage, to conclude that if consciousness ever does cause behavior, it does so far less frequently than has traditionally been supposed. This would seem to make it equally reasonable to suggest that it might be a good idea for the community to start discussing the question of how the law would or should be affected if the ultimate conclusion turns out to be that *all* so-called voluntary behavior is in fact unconsciously initiated. One possible solution would be to acknowledge that mind is a mixture of conscious and unconscious components, and judge any particular mind as guilty if it intended the offending act *either consciously or unconsciously*. Since lawyers have the best handle on where the law currently stands on this issue, it must be lawyers who at least lay the groundwork for discussion of this suggestion.

DOI: 10.1002/bsl

REFERENCES

Bayne, T. (2006). Phenomenology and the feeling of doing: Wegner on the conscious will. In S. Pockett, W. P. Banks, & S. Gallagher (Eds.), Does consciousness cause behavior? (pp. 169–186). Cambridge, MA: MIT Press.

Castiello, U., Paulignan, Y., & Jeannerod, M. (1991). Temporal dissociation of motor responses and subjective awareness. A study in normal subjects. *Brain*, *114*, 2639–2655.

Crick, F. (1994). *The astonishing hypothesis*. New York: Simon and Schuster.

Farrer, C., Franck, N., Georgieff, N., Frith, C. D., Decety, J., & Jeannerod, M. (2003). Modulating the experience of agency: A PET study. *Neuroimage*, *18*, 324–333.

Fourneret, P., & Jeannerod, M. (1998). Limited conscious monitoring of motor performance in normal subjects. *Neuropsychologia*, *36*, 1133–1140.

Frith, C. D., Blakemore, S. J., & Wolpert, D. M. (2000). Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society London B*, *355*, 1771–1788.

Gallagher, S. (2006). Where's the action? Epiphenomenalism and the problem of free will. In S. Pockett, W. P. Banks, & S. Gallagher (Eds.), *Does consciousness cause behavior*? (pp. 109–124). Cambridge, MA: MIT Press.

Jeannerod, M. (2006). Consciousness of action as an embodied consciousness. In S. Pockett, W. P. Banks, & S. Gallagher (Eds.), *Does consciousness cause behavior*? (pp. 25–38). Cambridge, MA: MIT Press.

Knoblich, G., & Kircher, T. T. J. (2004). Deceiving oneself about being in control: Conscious detection of changes in visuomotor coupling. *Journal of Experimental Psychology. Human Perception and Performance*, *30*, 657–666.

Kornhuber, H. H., & Deecke, L. (1965). Hirnpotentialänderungen bei Willkürbewegungen und passiven Bewegungen des Menschen: Bereitschaftspotential und reafferente Potentiale. *Pflugers Archiv für Gesamte Physiologie*, *284*, 1–17.

Libet, B. (1994). A testable field theory of mind–brain interaction. *Journal of Consciousness Studies*, *1*(1), 119–126.

Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness potential): The unconscious initiation of a freely voluntary act. *Brain*, *106*, 623–642.

Mele, A. R. (2006). Free will: Theories, analysis and data. In S. Pockett, W. P. Banks, & S. Gallagher (Eds.), *Does consciousness cause behavior*? (pp. 187–206). Cambridge, MA: MIT Press.

Nisbett, R., & Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231–259.

Pacherie, E. (2006). Towards a dynamic theory of intentions. In S. Pockett, W. P. Banks, & S. Gallagher (Eds.), *Does consciousness cause behavior*? (pp. 145–168). Cambridge, MA: MIT Press.

Pockett, S. (2000). *The nature of consciousness: A hypothesis*. Lincoln, NE: Iuniverse.

Pockett, S. (2002). Difficulties with the electromagnetic field theory of consciousness. *Journal of Consciousness Studies*, *9*(4), 51–56.

Pockett, S. (2004). Does consciousness cause behavior? *Journal of Consciousness Studies*, *11*(2), 23–40.

Pockett, S., Banks, W. P., & Gallagher, S. (2006). *Does consciousness cause behavior*? Cambridge MA: MIT Press.

Ross, P. W. (2006). Empirical constraints on the problem of free will. In S. Pockett, W. P. Banks, & S. Gallagher (Eds.), *Does consciousness cause behavior*? (pp. 125–144). Cambridge, MA: MIT Press.

Slachevsky, A., Pillon, B., Fourneret, P., Pradat-Diehl, P., Jeannerod, M., & Dubois, B. (2001). Preserved adjustment but impaired awareness in a sensory–motor conflict following prefrontal lesions. *Journal of Cognitive Neuroscience*, *13*, 332–340.

Sperry, R. W. (1950). Neural basis of the spontaneous optokinetic response produced by visual inversion. *Journal of Comparative and Physiological Psychology 43*, 482–489.

Trevena, J. A., & Miller, J. (2002). Cortical movement preparation before and after a conscious decision to move. *Consciousness and Cognition 11*, 162–190.

von Holst, E., & Mittelstaedt, H. (1950). Das Reafferenzprinzip. Wechselwirkungen zwischen Zentralnervensystem und Peripherie. *Naturwissenschaften*, *37*, 464–476.

Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge, MA: MIT Press.

Wegner, D. M., & Wheatley, T. (1999). Apparent mental causation: Sources of the experience of will. *American Psychologist 54*, 480–492.

DOI: 10.1002/bsl