TRAMES, 2011, **15**(**65/60**), 2, 147–155

NEUROSCIENCE, FREE WILL AND MORAL RESPONSIBILITY

Gardar Árnason

Leibniz Universität Hannover

Abstract. Neuroscientific challenges to free will work on at least three levels: there is a metaphysical level, an epistemological level, and an empirical level. In this paper I discuss the main neuroscientific challenges on each of these three levels. Three fundamental conditions for free will can also be placed on these levels, and I briefly discuss how these conditions can be met in the context of the neuroscientific challenges. In conclusion I strongly doubt that neuroscientific evidence can show free will not to exist at all.

Keywords: free will, neuroscience, responsibility, determinism, reductionism, Libet experiments

DOI: 10.3176/tr.2011.2.03

1. Introduction

There are scientists and philosophers who have claimed that neuroscience is undermining the concept of free will, or even revealing free will to be an illusion (Spence 1966, Pereboom 2001, Roskies 2006, Sie and Wouters 2008). In this paper I will look at the nature of the main neuroscientific challenges to free will and then sketch how we might be able to answer them, or at least pull out their sharpest teeth.

First, let me say a few things about what is at stake. Without free will, there can be neither moral responsibility nor legal culpability. Without free will, no one deserves punishment for breaking the law and no one deserves blame for immoral behavior. No one would deserve praise for good work either. In other words, on a conceptual level, free will is a precondition for moral responsibility.

If the view that free will is an illusion becomes widely accepted, it will have various implications for society. One of the implications is that the legal system would have to be drastically revised. Justice and desert cannot play any part in punishment, punishment could only be determined with regard to its beneficial effects for society (deterrence, satisfy needs for justice or revenge) and for the

person punished (rehabilitation). Some think this would be a good development towards a more humane punishment (Green and Cohen 2004, Burns and Bechara 2007).

Another implication is that people may behave less morally. According to recent psychological research, those who do not believe in free will are more likely to behave immorally than those who do think that they have free will (Vohs and Schooler 2008, Shariff, Schooler, and Vohs 2008).

A third possible implication may be more surprising. The spread of the belief that we do not have free will may have a negative economic impact. A recent study found that believing in free will predicts better work attitudes and better work performance than not believing in free will (Stillman et al. 2010). Believing in free will has stronger effects on job performance than well established predictors such as conscientiousness, belief in being in control of one's life, and Protestant work ethic.

Even if neuroscientific challenges to free will turn out to be poorly justified or even wrong, scientists' claims about free will being an illusion are quickly and easily circulated in the media, and can hence have negative impact on society. When neuroscientific claims are made against free will, we should take them seriously, but we need also to view them critically.

What do the neurosciences, then, tell us about free will? The neuroscientific challenges to free will work on at least three levels:

1. On a metaphysical level there is a determinist challenge: The mind is nothing more than what the brain does, and the brain is a physical, deterministic system.
2. On an epistemological level there is a reductionist challenge: The mind, mental phenomena, can be fully explained in terms of neural states, structures and functioning.
3. On an empirical level, there is a cognitive challenge: Decision-making is fundamentally unconscious and therefore not free.

The first challenge is primarily metaphysical in the sense that it has to do with the ultimate nature of the mind and the brain. The second challenge is primarily epistemological in the sense that it has to do with what we can know about the mind and the brain. Both challenges are *assumptions* of science, they cannot be supported by empirical evidence, at least not directly. These assumptions might survive if the science based on them is successful in producing knowledge, and as long as the knowledge produced does not conflict too badly with these assumptions. In that sense, the first two challenges could be said to be *self-vindicating* assumptions. Only the third challenge can rely directly on empirical evidence, but even here there are significant conceptual issues at stake.

In what follows I will discuss these challenges in more detail. I will also introduce three conditions or requirements for free will, one at each level, and attempt to make sense of them in the context of the corresponding neuroscientific challenge.

2. The metaphysical-determinist challenge

The deterministic challenge is that the mind is nothing more than what the brain does, and the brain is a physical, deterministic system. If this is the case, it seems that for any decision and action we make, we could not have done otherwise. So, we have no true choice, and therefore no free will.

This challenge is parallel, if not identical, to the deterministic challenge to free will in the philosophical debate about free will. My suggestion how to answer this challenge has also a parallel in the philosophical debate. It is a position called the compatibilist view of free will. The compatibilist view of free will holds that determinism (whether it is a determined universe or a determined brain) does not exclude the possibility of free will. So, how can that be?

First let me point out that physically determined causation is not coercion (see Levy 2007:223). We are coerced when we are forced to do something against our will. If a strong wind throws us against a window so that it breaks, we were physically forced to break the window. It was not a freely willed act. But the thief who breaks a window to steal jewelry is not being causally forced to break the window against his will. He is not being coerced or manipulated by external physical forces. So, how is determination any better than coercion? For the compatibilist, what matters is that our decisions and actions are determined in the right way, that is, by our beliefs, desires, values and intentions, and without external coercion or manipulation, and without internal coercion, such as a dysfunction of the brain.

But, one may ask, is it not possible that quantum mechanics is at work in the brain, with the result that the brain is not a deterministic system? And if so, is it not possible that we have a truly free will, which is not determined by physical causation at all? It may be tempting to introduce quantum mechanical indeterminacy in order to rescue free will, that is, to show that our decisions and actions are not physically determined and therefore we could have done other than we did. The problem with this is that our decisions and actions would be entirely random and arbitrary if they were not determined at all. There may be a role for quantum mechanics, which I will discuss in the next section, but it does not help at this stage at all. On the contrary, quantum mechanical indeterminacy may be a worse alternative to physical determinacy.

I said that if the brain is a deterministic system, then *we could not have done otherwise* than we did, and it may still seem the case that this is so and that this excludes free will. Although some have doubted that this is a necessary condition for free will (Frankfurt 1969), I think we need at least to make sense of what this requirement means.[1] The compatibilist claims that we could have done otherwise in the sense that *if* our beliefs, desires, values and intentions, or external circumstances, had been different, *then* we could have acted differently. There are

---

[1] For a recent criticism of Frankfurt-style counter-examples to the principle of alternative possibilities, see Robinson (2010).

at least two serious problems with insisting that having free will must mean that we could have done otherwise with everything relevant to our action being exactly the same.

The first problem is the familiar indeterminacy problem: if we could have acted differently, in the case when everything else was exactly the same up to the millisecond when we made our decision, then our decision would again be entirely random and arbitrary and no freer than if it was determined. In other words, if exactly the same deliberation, with exactly the same thoughts, intentions, beliefs and values, in exactly the same external circumstances, could result in two or more different decisions and actions, then we are not truly in control of our decisions and actions.

The second problem is that when we reflect on a past act and wonder whether we could have done differently, we are not interested in the case where everything leading up to the action is exactly the same. We are interested in things being *almost* the same, but with some minor, relevant changes. I may wonder whether I could have run my last marathon in less than four hours. Everything being exactly the same, of course I could not have. I could not have run *that exact same run* faster. Rather, what interests me, is whether I could have finished the run in a better time if I had put in more effort at the end in spite of the pain and cramps, or if I had started with a slower pace, or if I had eaten power gel during the run, instead of the bananas, and so on. I am not interested in whether I could have run faster if I had spent the previous five years in a Kenyan training camp for elite runners. What is of interest, when I say I could have done otherwise, is what minor but relevant changes would have made the action different, and this is entirely compatible with determinism (see Dennett 2003:75–77).

3. The epistemological-reductionist challenge

The reductionist challenge is that the mind, and mental phenomena, can be fully explained in terms of neural states, structures and functioning. This challenge is epistemological because it has to do with *explanation*, that is, whether and how we can *know* the neurological basis of mental phenomena.

To many neuroscientists, this may seem as an obvious or a necessary assumption for neuropsychology, at least in so far as it is concerned with mental phenomena or cognitive functions. If you want to study the neural structures and functions underlying, for example, perception, emotion, memory, consciousness or decision-making, you would have to assume that the explanations for these things are to be found on the neurological level, otherwise a neurological study of them would be rather pointless.

There is no doubt that mental phenomena and cognitive functions can be explained to a great extent in terms of neural states, structures and functioning. For example, we know a lot about which parts of the brain are involved in many mental phenomena and cognitive functions, such as perception, emotion and

decision-making. We know that lesions in specific parts of the brain cause specific mental or cognitive dysfunctions. What I doubt, however, is whether we can *fully* explain mental phenomena neurologically.

First of all, explanations have a very strong pragmatic element. Any fully reductionist model of conscious thought, decision-making or other complex cognitive functions would likely be of no pragmatic interest, as it would likely be as complex as the brain itself. For example, knowing all the zeros and ones going through a computer at a given moment will not explain to any human being what the computer is doing at the time, let alone does one have a sufficient explanation if one is provided with the location and movement of all the sub-atomic particles which make up the computer. To explain what is going on at a high level in a complex system, the explanation may have to work on the same or similar level of complexity. A more poetic example is that of a very short story by Borges named "On Exactitude in Science". In the story cartographers in an empire come up with better and better maps, until at last they create a map of the empire which is on the scale of 1:1. Of course such a map is useless, the empire itself could just as well serve as its own map.

Another potential obstacle to reductionist explanations of mental phenomena is that it can be argued that mental states, and possibly free will itself, are (non-mysterious) emergent properties of that complex system which is the brain. If this is the case, knowing the micro-physical, or neurological, basis of these emergent properties might not suffice to provide either predictions for what sort of phenomena emerge from any micro-physical state, nor an explanation of them.

It is possible to create very complex systems based on very few simple rules and given a certain initial setup, where the system is completely determined, yet knowing the rules and the initial setup does not help in predicting how the system evolves or what properties emerge in it. These can only be seen by letting the system run its course. In some such systems, for example the so-called Conway's Game of Life, one can observe complex patterns where events take place and events stand in causal relations to each other, without these events and their causal relations having any corresponding explanations at the level of the rules and initial setup (even if that level fully determines the emerging events and causal relations).

The example of complex systems can help us understand how mental properties, and even free will, might emerge in the brain, and be fully determined by the brain's neurological states, structures and functions, but still not have full explanations at that level. It may very well be the case that some aspects of mental phenomena can only be explained in terms of other mental phenomena, and not in terms of their physical basis, even if the mind is nothing more than what the brain does.

The epistemological condition for free will has to do with how we understand and explain our decisions and actions: We must be, and experience ourselves as, the *authors or origin of our decisions and actions*. In normal circumstances, we explain our decisions and actions with reference to our reasons for them (which may also involve our desires, beliefs, intentions etc.). The decisions or actions are ours if they follow from, and make sense in the context of, our reasons (and

desires, beliefs, intentions, etc.). And they must also not be coerced or manipulated.

Some philosophers are perfectly happy to consider this condition then fulfilled. Others, however, are concerned that if our physical and mental states are determined by past physical/mental states and the deterministic laws of nature, then we are really not the authors or origin of our actions (Searle 2001). One solution is to bring in complexity, possibly along the lines of my discussion of complex systems and emergent properties (see Walter 2002). Another solution is to bring in quantum mechanics. The German neurobiologist Martin Heisenberg has proposed in an essay in *Nature* in 2009, that quantum mechanical effects in the brain might provide the basis for free will (Heisenberg 2009). As I discussed above, quantum indeterminacy is not helpful in providing an alternative to having our decisions and actions determined by our reasons. But I think this is a more promising place for quantum indeterminacy: It would make us the origin of our actions and decisions, because quantum indeterminacy cuts us off from the physical determinacy of the external world. We are then, in this sense, uncaused causes. Quantum indeterminacy in the brain might have the form of a random generator of sorts, which feeds into the otherwise deterministic processes of the brain. This sort of random generator might perhaps have the role of opening a non-deterministic space of possibilities for the mind/brain to work with. So one could imagine, perhaps somewhat simplistically, such phenomenological results as getting ideas out of the blue, or an odd unexplained feeling when we are about to make a decision, or we think again about what we are doing, without any reason at all. Still, what we do with any of this is then determined by our reasons, beliefs, etc. However this quantum indeterminacy might actually work in the brain, the point of introducing quantum indeterminacy at this stage of the argument, is to show how the person could be the causal origin of her action, an uncaused cause, and *not* to explain 'how we could have done otherwise'.

The complexity solution, which I mentioned earlier, could also work differently from what I discussed above, that is, without having to rely on emergent properties. The complexity of the brain could create *pseudo-randomness* (without any particular unique properties emerging). Pseudo-random effects would be determined, but wholly unpredictable with regard to our knowledge of any part of the complex system. Since I am at his point considering free will only on the epistemological level, that is in terms of how we explain and understand our actions and decisions, a pseudo-randomness created by the complexity of the brain could do all the work of quantum indeterminacy. We would never be able to *know* the difference, and that is all that counts at the epistemological level.

4. The empirical-cognitive challenge

The cognitive challenge is that decision-making is fundamentally unconscious and therefore not free. This is the only level of the neuroscientific challenge to free will which is directly empirical.

Libet (1983; see also Haggard and Eimer 1999, Sirigu et al. 2004, Lau et al. 2004), Matsuhashi and Hallett (2008), and Soon et al. (2008) have claimed to show that a conscious decision to move, or the intention to move, happens after the brain has started preparing for movement. Soon could predict choice up to 7 seconds before the subject reported a conscious decision. On one interpretation of Libet's study, there is no free will but maybe a 'free won't' or a conscious veto of action initiated by an unconscious part of the brain.

If these studies and their interpretations are correct, then our consciousness is informed of our decisions, but does not have an active role in forming them (except, possibly, by having a power to veto them). Since free will and moral responsibility are often taken to suppose that we consciously will our actions, this seems to exclude the possibility of free will.

These studies have been much criticized in terms of methodology and design. In the studies by Benjamin Libet's group and John-Dylan Haynes group, the experimental subjects had to report the timing of their intention to move. For one thing, this is a rather subjective feeling. Also, the time which may pass from the subject noting an intention to move to noting the position of the 'clock' may skew the result. Finally, the subjects did not need to deliberate before making their movement, or have any reason for it, the intention to move was trivial. This may mean that the experimental setup only detected the subjective feeling of having an impulse to move, rather than a deliberate decision or a conscious intention to move. In simplified terms, since the decision to move was of no consequence and did not require any higher cognitive functions, the subject may have delegated the decision to the motor cortex or other unconscious parts of the brain, which then originated the action as it was supposed to, only briefly notifying consciousness in case there might be a veto.

Another objection concerns the meaning of consciousness. There is a difference between being (minimally) conscious on the one hand, and being directly conscious of one's decisions and the reasons behind them on the other hand. The first is necessary for free will, but the second is not. Most of the time we are not directly conscious of our decisions and actions, or the reasons for them. It is typically only if we need to particularly focus on or deliberate a decision, or if something out of the ordinary is going on, that we become directly aware of our decision and action. Most of the time our decisions and actions just barely enter consciousness, and that is a good thing. We know what we are doing, when everything is working right, in the sense that if we are asked what we are doing and why, we can give answers, even if we are not fully conscious of our decisions and actions at the time. We are not like automata in these cases, since automatism means that we are *not* even minimally conscious, for example in the case of sleepwalking.

Now I come to the cognitive condition for free will: We must be (at least minimally) conscious and rational, and able to act on reasons.

A person who is not conscious, but acting, is not acting freely, for example if she or he is sleepwalking. A person who is significantly irrational due to

mental illness has a correspondingly diminished free will and reduced moral responsibility.

There are reports of patients with damage to the ventromedial prefrontal cortex, which causes them to lose the ability to act on reasons, while retaining rationality. The patient can tell you what is rational to do, but then fails entirely to act on that knowledge. A patient with this sort of dysfunction has very limited moral responsibility and lacks in some significant sense free will.

There is much positive work for neuroscience to do on how rational decision-making and deliberation takes place in a fully functioning brain. In specific cases, failures of the functions involved can lead to diminished responsibility, even if free will is not altogether lacking.

5. Conclusion

What I wanted to do in this discussion of neuroscientific challenges to free will was not to 'prove' the existence of free will in the face of contrary evidence from neuroscience. My main point is that the issues are a lot more difficult than (at least some) interpretations of the neuroscientific case against free will might suggest. Still, I think the philosophical considerations of the neuroscientific challenges to free will quite strongly suggest that a *universal* challenge to free will based on neuroscientific evidence is unlikely to be successful. In other words, I think that neuroscience has not revealed free will to be an illusion and that it is not likely ever to do so.

Neuroscience may, however, affect our views of moral responsibility, not by showing that we do not have free will at all, but by showing that in many specific cases where we now consider people responsible, they were actually not responsible, because of lack of rationality or lack of relevant control over their decisions and actions.

Address:
Gardar Árnason
Leibniz Universität Hannover
Institut für Philosophie
Im Moore 21
D-30167 Hannover
Germany
E-mail: gardar.arnason@philos.uni-hannover.de

References

Burns, Kelly and Antoine Bechara (2007) "Decision making and free will: a neuroscience perspective". *Behavioral Sciences and the Law* 25, 263–280.

Dennett, Daniel C. (2003) *Freedom evolves*. London: Penguin Books.

Frankfurt, Harry G. (1969) "Alternate possibilities and moral responsibility". *The Journal of Philosophy* 66, 23, 829–839.

Green, Joshua and Jonathan Cohen (2004) "For the law, neuroscience changes nothing and everything". *Philosophical Transactions of the Royal Society* (London) B 359, 1775–1785.

Haggard, Patrick and Martin Eimer (1999) "On the relation between brain potentials and the awareness of voluntary movements". *Experimental Brain Research* 126, 128–133.

Heisenberg, Martin (2009) "Is free will an illusion?". *Nature* 459,14, 164–165.

Lau, Hakwan C. et al. (2004) "Attention to intention". *Science* 303, 1208–1210.

Levy, Neil (2007) *Neuroethics: challenges for the 21st century*. Cambridge: Cambridge University Press.

Libet, B., C. A. Gleason, E. W. Wright, and D.K. Pearl (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.

Matsuhashi, Masao and Mark Hallett (2008) "The timing of the conscious intention to move". *European Journal of Neuroscience*, 28, 2344–2351.

Pereboom, Derk (2001) *Living without free will*. Cambridge: Cambridge University Press.

Robinson, Michael (2010) "Modified Frankfurt-type counterexamples and flickers of freedom". *Philosophical Studies*, published online 28 September 2010, at http://www.springerlink.com/content/l263t84798574186/, DOI: 10.1007/s11098-010-9631-z.

Roskies, Adina (2006) "Neuroscientific challenges to free will and responsibility". *Trends in Cognitive Sciences* 10, 9, 419–423.

Searle, John (2001) "Free will as a problem in neurobiology". *Philosophy* 76, 491–514.

Shariff, Azim F., Jonathan Schooler, and Kathleen D. Vohs (2008) "The hazards of claiming to have solved the hard problem of free will". In *Are we free? Psychology and free will.* J. Baer, J. C. Kaufman, and R. F. Baumeister, eds. New York: Oxford University Press.

Sie, Maureen and Arno Wouters (2008) "The real challenge to free will and responsibility", *Trends in Cognitive Sciences* 12, 1, 3–4.

Sirigu, Angela et al. (2004) "Altered awareness of voluntary action after damage to the parietal cortex". *Nature Neuroscience* 1, 80–84

Soon, Chun Siong et al. (2008) "Unconscious determinants of free decisions in the human brain". *Nature Neuroscience* 11, 5, 543–545.

Spence, Sean A. (1996) "Free will in the light of neuropsychiatry". *Philosophy, Psychiatry, and Psychology* 3, 2, 75–90.

Stillman, Tyler F., Roy F. Baumeister, Kathleen D. Vohs, Nathaniel M. Lambert, Frank D. Fincham, and Lauren E. Brewer (2010) "Personal philosophy and personnel achievement: belief in free will predicts better job performance". *Social Psychological and Personality Science* 1, 1, 43–50.

Vohs, Kathleen D. and Jonathan W. Schooler (2008) "The value of believing in free will: encouraging a belief in determinism increases cheating". *Psychological Science* 19, 49–54;

Walter, Henrik (2002) "Neurophilosophy of free will". In *The Oxford handbook on free will*, 565–576. Robert H. Kane, ed. Oxford and New York: Oxford University Press.