ON JAMES'S ARGUMENT AGAINST EPIPHENOMENALISM

JOHN WRIGHT

ABSTRACT

Epiphenomenalism is the doctrine that mental states lack causal efficacy. A common objection against epiphenomenalism is that this makes it mysterious how or why mental states ever evolved. One particularly powerful form of this objection was developed by William James. James argued that epiphenomenalism cannot account for the familiar fact that what we find pleasurable is typically good for us, while what we find painful is typically bad for us. In this paper it is argued that James's objection to epiphenomenalism is sound. But it is further argued that James's argument constitutes a difficulty, not just for epiphenomenalism, but also for the thesis that mental states do possess causal powers. The paper concludes with some suggestions about how the problem to which James has drawn our attention might be solved.

Epiphenomenalism is the doctrine that, although mental events and properties exist, they lack causal efficacy. One objection to this view is that, if the mental is indeed wholly causally inefficacious, there is no apparent reason why creatures with a mental life should ever have been naturally selected. A particularly telling form of this objection to epiphenomenalism was raised by William James. In this paper James's argument against epiphenomenalism is defended. More specifically, in this paper it is argued that James is correct in saying it is *necessary* to attribute causal powers to the mental.

But this paper has an additional aim. It will be argued that although it is *necessary* to say the mental has causal powers, this is not *sufficient* to satisfactorily explain the points James raises. The main aim of the paper is to argue that James's considerations give rise to a puzzle that has hitherto received insufficient attention in the philosophy of mind. It is a puzzle that is not completely removed even if we *do* attribute causal powers to the mental. Attributing causal powers to the mental is, then, necessary to accommodate James's points, but not sufficient. In the final sections of this paper a possible solution to James's problem is briefly sketched.

1. JAMES'S ARGUMENT AGAINST EPIPHENOMENALISM

In his *The Principles of Psychology* and "Are We Automata?", William James argued against the view that mental states and events, specifically: feelings of pleasure and pain , lacked causal efficacy. He called the theory that they do lack such efficacy "The Automaton Theory"; now we more usually refer to it as "epiphenomenalism". James was directing his argument against T. H. Huxley, Herbert Spencer, C. K. Clifford and Shadworth Hodson.[1] The core of James's argument is very simple: the things we like are generally good for us, while the things we dislike are generally bad for us. Here is what James said:

> It is a well-known fact that pleasures are generally associated

> with beneficial, pains with detrimental, experiences. …Mr Spencer and others have argued that these co-incidences are due, not to any pre-established harmony, but to the mere action of natural selection which would certainly kill off in the long run any breed of creatures to whom the fundamentally noxious experience seemed enjoyable… But if pleasures and pains have no efficacy, one does not see (without some such *a priori* rational harmony as would be scouted by the "scientific" champions of the automaton-theory) why the most noxious acts, such as burning, might not give thrills of delight, and the most necessary ones, such as breathing, cause agony.[2]

We will refer to the fact that what we enjoy is generally good for us, and what we find painful generally bad, as the "felicitous alignment".[3]

Although James does not fully spell out his argument against epiphenomenalism, it is presumably along the following lines: Why should it be the case that the things that give us pleasure tend to be good for us? Why should this "felicitous alignment" exist? Surely the reason is just this: in the past, organisms that got pleasure from health-producing activities *performed* those health giving activities and so were naturally selected, while those that got pleasure from health-damaging activities performed those health damaging activities, and therefore died out. As James remarked:

> An animal that should take pleasure from a feeling of suffocation would, if that pleasure were efficacious enough to make him immerse his head in water, enjoy a longevity of four or five minutes.[4]

Conversely, since we do not find in nature animals that experience agony from breathing, it is natural to suppose that any such animals would have been removed by natural selection. But – James is evidently arguing – natural selection would only remove such organisms if the agony they experience *caused* them to refrain from breathing. Thus, natural selection has brought it about that the only organisms surviving are those that get pleasure from health producing activities, and pain from health damaging activities.

It is an essential feature of this explanation that it apparently attributes *causal powers* to pleasure and pain. It says, for example, that organisms that derive pleasure from health producing activities will perform those activities. And presumably it is *because* they derive pleasure from those activities that they engage in them: the pleasurable nature of the activities is a *cause* of organisms partaking in those activities. But if this is the case, then epiphenomenalism (James's "Automaton Theory") is false.

It is worth noting that, as a challenge to epiphenomenalism, James's argument is powerful. In particular, it raises difficulties not raised by a related, and perhaps better known, argument against epiphenomenalism from evolution. Popper and Eccles, for example, argue against epiphenomenalism on the grounds that it makes it a mystery as to why the mental should ever have evolved.[5] Briefly, their argument is that if the mental were to be naturally selected it would have to increase an organism's fitness, and to do this it would have to make a difference to the organism's behaviour. But, if mental events are to make a difference to an organism's behaviour, they must surely (and contrary to epiphenomenalism) have causal powers. Thus, Popper and Eccles argue, considerations from the theory of evolution lead us to reject epiphenomenalism.

The epiphenomenalist can, however, reply to the argument of Popper and

Eccles. Their argument assumes that if the mental is to arise as a result of natural selection, it must *contribute* to fitness. But this is simply false. It *may* be naturally selected as a result of contributing to fitness, but it may also become prevalent within a species (without contributing to fitness) by being a causal consequence of some other feature F that does contribute to the fitness. This approach is developed and defended by, for example, Frank Jackson.[6] Jackson argues that it is presumably the case that the insulating properties of fur contribute to an animal's fitness. And if an animal is covered in fur, that animal will also be soft to the touch: its softness to the touch is a causal *consequence* of it being covered in fur. But the property of being soft to the touch need not itself increase the animal's fitness. Similarly, it is at least possible that the mental life of an organism might, without itself increasing fitness, be a causal consequence of some other feature of the organism that does. And such a possibility is clearly compatible with epiphenomenalism.

Whatever strength Jackson's suggestion may have against the argument of Popper and Eccles, it has rather less strength as a reply to James. Suppose that the feelings of pleasure we usually get from things that are good for us (and pain from things that are bad) were merely a by-product of something else that did contribute to fitness. Then: there would be no evident reason why the things that are good for us should produce feelings of pleasure rather than pain. If the feelings lack all causal efficacy, there would, as James says, be no reason why burning should not give rise to thrills of delight and breathing to agony. But, as matter of fact, the beneficial activities *do* give rise to pleasure and the harmful ones to pain. The fact that it is this felicitous state of affairs that *actually* holds, rather than the infelicitous one in which breathing causes agony etc., is not satisfactorily accounted for merely by saying the mental is a (causally inefficacious) by-product of something else that does increase fitness. It does not explain why we are in the *felicitous* situation we are in. James's argument has a strength not possessed by the argument of Popper and Eccles.

James's argument raises a difficulty that has not been resolved in still more recent discussions. Jack C. Lyons defends a form of epiphenomenalism that is naturally termed "property epiphenomenalism"[7]. On Lyons' view, although pains may be causally efficacious, the property of painfulness is not. Lyons gives a useful parallel: mountains may be causally efficacious while the property of "mountain-ness" might not be. If a plane flies in to a mountain it certainly causes the metal of the plane to crumple and bend, but the concept of "mountain-ness" may play no causal-explanatory role here. All the explaining might be done by properties such as the hardness of the rock, its mass, and so on. Similarly, a given event that is a pain might, on Lyon's view cause Smith to withdraw his hand, but the property of *painfulness* may play no explanatory role in accounting for Smith's behaviour. All the explaining might be done by, for example, properties of neurophysiology.

But Lyon's "property epiphenomenalism" still leaves it unexplained why James's felicitous alignment should hold. The things that are good for us tend to be correlated with the *property of pleasantness*, the things that are bad for us with the *property of painfulness*. And, to foreshadow issues raised in the next section, we seem to be drawn towards certain things *because* they are pleasurable and repelled by others *because* they are painful. We would appear to be left without an explanation of these correlations if the properties of painfulness and pleasurableness are held to be causally inefficacious.

One author who has paid particular attention to the challenge James's argument presents for epiphenomenalism is William Robinson[8]. Robinson allows that, initially at least, James's argument might seem powerful, but argues that in fact it fails to refute epiphenomenalism. Robinson's essential point is that

if epiphenomenalism fails to explain the felicitous alignment, then so must the view favoured by James. If this is correct, concludes Robinson, James's argument gives us *no reason to prefer* James's own view to epiphenomenalism.

Let us now consider Robinson's argument. Suppose it to be the case that epiphenomenalism provides us with no explanation of the felicitous alignment, more specifically, suppose epiphenomenalism provides us with no explanation of why those things we find pleasurable tend to be good for us. How might we obtain such an explanation by attributing causal powers to, for example, pleasure "itself"? Presumably, any causal powers we could attribute to the psychological state of pleasure "itself" could also be attributed to a neural correlate of pleasure. For example, perhaps the pleasurableness of eating chocolate increases the chances it will be eaten in the future, but so, surely, could some neural state increase the probability chocolate will be eaten in the future. Conversely, if there are no causal powers we could possibly attribute to any neural state that would explain the alignment, what sort of causal powers might we attribute to pleasure itself that would enable us to explain it? Robinson asserts there are none. We gain no explanatory advantage by rejecting epiphenomenalism, according to Robinson. He concludes that James has not, therefore, given us any reason to reject epiphenomenalism.

It is worth noting that Robinson himself has not given us an explanation of the felicitous alignment. All he has done is argue that the prospects for explaining it seem, on the face of it, to be just as good with epiphenomenalism as without it. It seems to the present author that on this issue Robinson is partially right. It will be argued later in this paper that *merely* attributing causal powers to the mental is not enough to explain the felicitous alignment. More specifically, it will later be argued that while it *is* possible to explain the felicitous alignment, to do so it is necessary, but not sufficient, to attribute causal powers to the mental. On the view to be advocated, epiphenomenalism must be rejected if the felicitous alignment is to be explained, and so James's argument against epiphenomenalism remains good. But in order to account for the felicitous alignment, more work needs to be done than simply rejecting epiphenomenalism. To repeat a claim already made, here it will be argued that to explain the felicitous alignment it is necessary, but not sufficient, to reject epiphenomenalism.

2. THE FELICITOUS ALIGNMENT AND THE FORTUNATE CORRELATION

As we have reconstructed it, James explains the "felicitous alignment" between what is pleasurable and what is good for us by saying that if (for example) a human being finds some activity pleasurable they will tend to engage in that activity, and if they find some activity painful, they will tend to avoid it. There is, we may say, a *fortunate correlation* between the stimuli we subjectively experience as pleasurable, and our *behavioural* tendencies to seek out those stimuli, and between what we find painful, and our behavioural tendency to avoid the painful stimuli. This fortunate correlation between pleasure and seeking out, and between pain and avoidance, is used by James to explain the "felicitous alignment" between pleasure and beneficial qualities, and pain and harmful qualities.

Note that what we are calling the "fortunate correlation" is distinct from James's "felicitous alignment". The "fortunate correlation" is a correlation between our feelings of pleasure and our *behaviour* of seeking out things that produce those feelings (and between pain and our behaviour of avoiding the causes of pain.) It is a correlation between feelings and behaviour. James's "felicitous alignment" is a correlation between feelings of pleasure and that

which is *beneficial for our health or survival* (or between pain and that which is detrimental to our survival).

It is clear that James relies on what we are here calling the "fortunate correlation" (between pleasure and seeking out behaviour) to help *explain* the felicitous alignment (between pleasure and healthfulness). As we have noted, he evidently assumes that if an animal were to find some activity (such as immersing its head in water, or breathing air) pleasurable, *they would engage in that activity*. So natural selection ensures the only animals left surviving will be those that engage in, because they take pleasure in, the healthful activities. *Given* that the fortunate correlation exists, we can explain the felicitous alignment. But, of course, we are now confronted with the question: "Why should the "fortunate correlation" between pleasure and seeking out behaviour (and between pain and avoidance behaviour) exist?" James himself does not address this question, and it will be argued that addressing it leads to puzzles both for epiphenomenalism and for the doctrine that the mental does have causal powers.

3. DOES THE "FORTUNATE CORRELATION" ACTUALLY REQUIRE EXPLANATION?

We tend to move away from things, such as fire, dangerous animals and so on that would cause us subjectively unpleasant pain or distress. Our bodily movements tend to decrease the chances of us experiencing pain and distress. Our movements also tend to *increase* the chances we will experience subjectively pleasant experiences. Generally, there is a much higher chance my hand will move towards the chocolate cake, but not towards the pain-causing flame. *But*: intuitively, things did not have to be that way. We seem to be able to imagine what I will call a "nightmare world". In this nightmare world, we find ourselves trapped in physical bodies that tend to seek out things that cause us to experience subjectively unpleasant feelings. For example, we seem to be able to imagine a world in which we *behave* exactly as we do in this word – perhaps a world in which we walk in to a shop, buy a bar of chocolate and eat it – but we subjectively experience terrible agonies as we do so. On the outside, we are eating the chocolate and saying: "This is delicious!", but on the inside we are thinking: "Oh no, not chocolate again!", and we subjectively experience excruciating pain each time some of the chocolate goes in to our mouth. This nightmare world would be a world in which our physical bodies did not obey our conscious will. We could not even communicate our situation to the outside world, because we could not control the movements of our lips.

Fortunately, however, we do not live in this nightmare world. In the actual world, what we are here calling the fortunate correlation does in fact hold: our bodily movements do in fact tend to increase the chances we will experience subjectively pleasant feelings, and decrease the chances we will experience subjectively unpleasant feelings. But it seems things could have been otherwise. We *might have* lived in the nightmare world. And so, it seems, an explanation is required of why, in the actual world, the fortunate correlation holds.

It should be observed that, on some positions, the fortunate correlation is *not* the type of thing that actually requires an explanation. For example, advocates of analytic functionalism, behaviourism or some verificationist theories of meaning might hold that the nightmare world sketched above is not a genuine conceptual possibility at all. On such views, it is (perhaps) simply not the case that we *might have lived* in the nightmare world. And if it is not the case that we might have lived in the nightmare world then, it is natural to claim, we do not need an explanation of why the fortunate correlation holds.

In this paper it will be *assumed* that the nightmare world is indeed a genuine conceptual possibility. Although it will be assumed to be possible, there are three things that can be said that might at least partly soften the resistance of those who are inclined to deny its possibility.

First, although analytic functionalism, behaviourism and so on are *perhaps* incompatible with the possibility of the nightmare world, the fact that they do appear to exclude even the conceptual possibility of such scenarios is surely one reason why these positions have been found to be less than entirely convincing. I think it is fair to say that a substantial number of philosophers have felt it is at least a conceptual possibility that our private, subjective mental states could have been very different from those *actually* correlated with certain behaviour, or functional organisation. And this is one reason why behaviourism and analytic functionalism have been thought to not tell us the whole of the story about our mental states.

Second, I think it is clear that we can imagine certain situations in which "the nightmare world" holds for at least a limited period of time. Suppose, for example, that a person is hooked up to a device that is capable of "scanning" their preferences and desires. Let us further suppose that the person hates, for example, eating olives. When they eat olives their subjective, conscious gustatory sensations are extremely unpleasant. Under normal circumstances, they would avoid eating olives. The fiendish device, however, is able to detect their aversion to olives, and then sends electrical signals to the muscles, forcing the person to go through the motions of gulping down piles of olives. Further, the device also forces their throat and mouth muscles to say: "These olives are delicious; more please." But, at the same time, what is going on "inside" is that the subjective experiences of the person are intensely unpleasant. On the inside, the person is perhaps screaming for the thing to stop, even though on the outside they are commenting on the deliciousness of the olives.

I think it is clear we are able to imagine a person being placed in such a device for, say, one hour. This seems to be clearly a conceptual possibility. But if a person could occupy such a device for one hour, it also seems plausible they could occupy it for two hours, or three hours, or indeed their entire lives. And if it is a conceptual possibility that *one* person could spend their entire life imprisoned in such a device, it is not clear why there could not be a community of such people. But if we admit this, then it appears we have, in effect, admitted the conceptual possibility of the "nightmare world".

Finally, it is worth noting that the nightmare world perhaps is *not necessarily* ruled out by behaviourism or analytic functionalism. Consider a world in which a person is hooked up to the fiendish device described above. A behaviourist, for example need not say that what it means for such a person to hate olives is for them to avoid eating olives. Rather, a behaviourist may say that all it means is that the person has a *disposition* to avoid eating olives. And it may well be maintained that in the above scenario, the person still has within them a state that disposes them to avoid olives, but other factors are preventing that disposition from displaying its normal effects. The same might be said about some part of that person's functional organisation. So, it seems, both a behaviourist and an analytic functionalist might allow that the nightmare world is a conceptual possibility.

It might be protested that the writings of James himself contain an argument against the possibility of the nightmare world. In his "What is an Emotion?" James argues that the physical accompaniment of an emotion is necessary for us to experience that emotion in all its "colour and warmth"; if, for example, we do not behave in a fearful way in the presence of the spider then our fear of the spider reduces in intensity. However, as Capek notes[9], even if this

is granted, it does not show there is nothing to be explained. We still need, for example, an explanation of why *initially*, at least, we felt fear at the sight of the spider or pain at the touch of the hot plate. These facts would still require explanation even if the intensity of the emotions were to fade if we refused to act on them.

Of course, the assumption that the nightmare world is a conceptual possibility is controversial. But I hope the considerations of this section have gone at least some way to softening the opposition of those who are inclined to say the nightmare world is uncontroversially or obviously not possible.

4. *HOW MIGHT THE FORTUNATE CORRELATION BE EXPLAINED*?

In the actual world the "fortunate correlation" holds. But, we are here assuming, it might not have held: we could instead have lived in a "nightmare world". Since the fortunate correlation *might not have* held, we are confronted with the question: Why does it *actually* hold?

Presumably, subjective, conscious experiences of pleasure and pain only exist in humans and (some) animals. This strongly suggests that any explanation of the fortunate correlation will be at least broadly "evolutionary" in character. However, it will be argued that giving an evolutionary explanation of the fortunate correlation is not as easy as it might at first be thought. Most evolutionary explanations either presuppose the fortunate correlation (or something very similar to it), or else explain something other than the fortunate correlation.

Here is a natural first attempt to explain the fortunate correlation. Suppose a particular species eats berries. The red berries are good for the species, the green berries bad. Can we explain the existence of the fortunate correlation in the species by saying that those members of the species that liked the red berries ate them and so survived, while those that liked the green berries ate them and so died out? The trouble with this suggestion is that it does not explain *the fortunate correlation*. What it explains is why organisms that preferred the red berries survived. It explains the prevalence of the preference for red berries. But in saying "those members of the species that liked the red berries ate them", the proposed explanation presupposes the fortunate correlation: it assumes that *liking the red berries* will be correlated with the *behaviour* of eating them. What needs to be explained is why preferring the red berries, in the sense of having a pleasant, subjective conscious sensation when eating them, should be associated with the physical behaviour of eating them.

It might perhaps be suggested that those animals that tended to seek out things that gave them subjective, conscious pleasure also tended to survive, while those that did not had a lower rate of survival. Whether or not this is true as an assertion of fact, it is clearly unsatisfactory as an explanation of the fortunate correlation. We are still left with the question: Why did the organisms that sought out things that gave them pleasure have a better chance of survival? It might perhaps be suggested: If organisms liked the things *that were good for them*, and also sought them out, then those organisms would have a better chance of survival. But this suggestion also has problems. One main difficulty is the organisms *liking* of the things that are good for it seems quite otiose in the explanation. All an organism has to do is in fact seek out things that are good for it. Whether this seeking out behaviour is associated with subjectively pleasant conscious sensations, subjectively unpleasant conscious sensations, or no conscious sensations at all, would appear to be superfluous to the explanation. Since the behaviour of seeking out things that are good for the organism is all that is required for survival, the question arises: Why should this behaviour be

associated with pleasant, conscious experiences? The explanation given simply *assumes* that the behaviour of seeking out things that are good for the organism is associated with subjectively pleasant conscious experiences. But this is precisely what we want explained. The proposed explanation therefore leaves us back at square one. Another approach is needed.

What we want is to explain is the existence of a correlation between subjectively pleasant conscious experiences and seeking out behaviour. Frequently, correlations are seen as evidence for, and are typically taken to be explained by, causal laws. So, it might be suggested we simply take it to be a causal law that subjectively pleasant conscious sensations tend to produce seeking out behaviour in organisms. It is this causal law that explains the fortunate correlation.[10]

However, it will be argued that this suggestion does not give us a satisfactory explanation of the fortunate correlation. Briefly, the difficulty is that while it perhaps explains the *correlation*, it does not explain its *fortunateness*. We have already noted that it seems to be possible that subjectively *unpleasant* conscious experiences could have been correlated with seeking out behaviour. We described such a possibility as the "nightmare world". But if the nightmare world is a possibility, it seems there also could have been a world in which there was a lawlike causal link between subjectively *unpleasant* conscious experiences and seeking out behaviour. Our behaviour in such a world could have been exactly the same as in this world. And so the question arises: Why were we lucky enough to live in a world where the lawlike connection that actually obtains is between pleasant conscious experiences and seeking out behaviour, rather than between unpleasant conscious experiences and seeking out behaviour? Postulating a lawlike connection fails to explain why the *fortunate* correlation is the one that obtains in the actual world. In this respect, it reduces us to saying: "Well, we were just lucky."

The conclusions of the last paragraph have an important consequence: the problems associated with explaining the fortunate correlation persist *even if* we attribute causal powers to mental events: saying the mental has causal powers, by itself, fails to explain why the *fortunate* correlation holds.

It might perhaps be suggested that the whole problem arises from assuming "pan-selectionism", or the doctrine that if a feature of some class of organisms exists, it must somehow contribute to fitness. But amongst biologists this doctrine is at least to some extent controversial. We have so far simply been assuming that the fortunate correlation has to increase fitness. But perhaps it doesn't: perhaps it is merely a by-product of other processes, and does not itself help to make organisms fitter.

However, as we have already argued in connection with the Popper-Eccles objection to epiphenomenalism, it is clear this is not satisfactory. It is one thing to say that consciousness itself is, from the point of view of increasing fitness, a kind of epiphenomenal by-product, but it is not at all plausible to claim this about the fortunate correlation. The problem is that the fortunate correlation has certain features which it is very implausible indeed to say are merely an epiphenomenal by-products of other processes.

There are some possible "nightmare worlds" that would be truly horrific. For example, there are nightmare worlds in which everything we did caused us unspeakable agony. But in the actual world there is at least a pretty high correlation between what we do and what gives us pleasure, and what we avoid and what causes us pain. If this is just good luck, then we have been very fortunate indeed.

There is another respect in which we have been very lucky. It is *a priori* highly unlikely that there should be the *degree* of correlation there actually is

between our bodily movements and what gives us pleasure. The number of bodily movements over an individual's life runs in to the hundreds of thousands, if not millions. The *a priori* probability that such a high proportion of these should be linked (directly or indirectly) to the obtaining of what gives us pleasure and the avoidance of what causes pain is surely very low. It does not seem to be plausible to say this is merely due to good luck. The idea that the fortunate correlation is just an accidental or epiphenomenal consequence of other processes must therefore be rejected as unsatisfactory.

In summary, in this section we have considered a number of possible ways in which the fortunate correlation might be given an evolutionary explanation. It has been argued that none of them are satisfactory. Explaining the fortunate correlation is not as easy as it might at first seem.

5. THE SIGNIFICANCE OF THE FORTUNATE CORRELATION

In the previous section a range of possible explanations of the fortunate correlation were considered. Problems were found for all of them. In particular, it was argued that aspects of the fortunate correlation remain unaccounted-for even if we *do* attribute causal powers to the mental. The fortunate correlation is, therefore, not merely a problem for epiphenomenalism: it is a problem for *all* the views of the relation between the mental and the physical.[11]

On the face of it, the issues raised by the fortunate correlation (and the possible existence of a "nightmare world") resemble those raised by the possibility of "inverted spectra". It is a familiar observation that the outward behaviour of persons is compatible with very different hypotheses about their inner, private experiences. Perhaps the most frequently discussed case of this sort is the inverted visual spectrum: It seems to be possible for a speaker's publically observable use of terms such as "red" and "green" to be exactly the same as that of normal speakers, and yet for the private sensations associated with those terms to be exchanged. And, of course, we can also imagine parallel possibilities for other sensory modes: inverted aural sensations, permuted olfactory and gustatory sensations and so on. On the face of it, the possibility of our "nightmare world" might seem to be just another hypothesis of this sort – in the nightmare world a speaker's publically observable use of the terms "pleasure" and "pain" would be the same as that of speakers in the actual world, but the private sensations associated with those terms would be exchanged or "inverted". So: it might be tempting to assume that the issues raised by the possibility of a nightmare world are no different from those raised by the possibility of inverted spectra and similar hypotheses. This, however, is not the case. There is an additional feature raised by the possibility of the nightmare world that is not raised by the other variants on our inner or private experience.

We are, quite plainly, *fortunate* that we do not live in the nightmare world. But there is no sense in which we are fortunate to have a normal rather than inverted spectrum. Having an inverted spectrum would merely be *different* from having a normal spectrum: it would not be horrible or horrific or extremely unfortunate. Similarly, having inverted or permuted aural or olfactory or gustatory sensations need not be obviously horrific or bad. But this means that the fact that the fortunate correlation holds gives rise to the need for *more explanations* than does the fact that we have the spectrum we actually have. We are lucky to live in a world in which the fortunate correlation holds, and ought to be glad we do. A satisfactory explanation of the fortunate correlation would, as we noted in the previous section, therefore need to explain our *fortunateness*. But no such requirement lies on a satisfactory explanation of why we have the visual (or olfactory or gustatory etc.) spectra we actually have.

6. *A STRATEGY FOR EXPLAINING THE FORTUNATE CORRELATION*

The aim of the following sections is to tentatively suggest a possible explanation of why the fortunate correlation actually exists in humans and (presumably) in other animals. It should be noted that the explanation to be advanced is intended as no more than possible, speculative hypothesis.

It is suggested that one strategy for finding an explanation of the fortunate correlation is to ask the question: How must organisms have been *prior to* the fortunate correlation if organisms in which the fortunate correlation held were to be naturally selected? In this section it will be argued that the fortunate correlation would have arisen if organisms *already possessed the power to freely choose*. The power of free choice was, on this view, evolutionarily prior to the fortunate correlation.

First, let us note that the following conditional seems *prima facie* plausible:

> If an organism has the power of free choice, then, *ceteris paribus*, the organism will tend to choose those options that it believes will give it pleasure and avoid those it believes will give it pain____(P).

Of course, there are many circumstances in which an organism (in particular, a human being) with the power of free choice will not choose to seek out pleasure and avoid pain. They may do this because of their ethical beliefs, or a sense of duty, or because they believe greater long term pleasure will come from some short term pain, and so on.[12] However, we may include qualifications such as "unless duty, ethics etc. incline them to act otherwise" in the *ceteris paribus* clause.

A perhaps more puzzling case for P is the phenomenon of masochism, or the deriving of pleasure from pain. One possible way of dealing with masochism might be to replace the word "pain" as it appears in P with, perhaps, "displeasure" or "feelings of unpleasantness". However, this complication will be ignored in what follows.

Subject to the qualifications noted above, there seems to be a sense in which P is not something that requires *further explanation*. Suppose a person freely chose the painful option over the pleasant one. In such a case we would surely look for an explanation of the person's behaviour. And the explanation might be, for example: a sense of duty, or a belief that the course of action will result in greater good in the future, or masochistic tendencies, etc. But now, let us suppose instead that a person chose the pleasurable option and avoided the painful one. We also confirmed that in this case the *ceteris paribus* clause held: there were no ethical or other factors present that would have constituted a reason for choosing the painful option. In such a case it would be clearly odd to ask: "But *why* did the person freely choose the pleasurable option?" That the person freely chose in this manner does not appear to be something that can be given, or requires, further explanation. There seems to be a sense in which the need for explanation comes to an end at that point.

It should, however, be noted that to say that P does not require explanation does not mean that it is *necessarily* true that, *ceteris paribus*, people will choose the pleasurable option and reject the painful one. If the choice is made freely, then presumably there is a sense in which they could have chosen otherwise, and this is surely incompatible with it being *necessarily* true that they choose the pleasurable option. But if a person chooses the pleasurable option, and the

ceteris *paribus conditions* hold, it does seem to be unnecessary to seek a *further* explanation of why they chose in the way they did.

7. *WHY PAIN*?

Suppose an organism lives in an environment in which there is something harmful to it. More specifically, assume the organism is a species of animal that lives on grassland, and there are occasional small fires that would harm it if it went too close. Clearly, it will be to the species' advantage to avoid the fires. There are many possible ways in which a disposition to avoid the fires might develop in the species, and be naturally selected for. *One* possible mechanism might be for the organism to experience pain on getting too close to the fire, and so be compelled to withdraw. But, of course, this is only *one* possible mechanism. We can imagine many other possible mechanisms, not involving conscious experience at all, by which an organism (or a robot) might be caused to avoid fires. For example, an organism (or robot) might be equipped with a structure capable of detecting heat. If the heat rises above a certain level, the organism is caused to withdraw. This could, quite easily, occur without any subjective, conscious sensation of heat or pain. Why would *pain*, as a subjectively felt, *nasty feeling* be selected for as the mechanism for fire-avoidance, rather than one of the many other possible mechanisms?

It is suggested that the considerations of the previous section provide a possible answer. If organisms *already* have a power of free choice, then they will tend to avoid that which causes subjectively unpleasant sensations. *What would be the point in pain having its subjective feeling of nastiness or unpleasantness unless the experiencer had the power to freely choose the direction in which it will move*? In organisms or mechanisms *without* a power of free choice, *any* device or internal mechanism that causes the organism to move away from the harmful stimulus will do. But in organisms that have a power of free choice, the mechanisms responsible for getting it to behave in a particular way will be subject to a particular constraint: they seem likely to work more effectively if they get the animal to move in a way that is in accordance with the freely made choices it is disposed to make. Imagine, for example, an organism that had within it a mechanism that gave it a propensity to *avoid* fires; but the organism *also* had the power of free choice and found being dangerously close to fires pleasurable. Given the organism had the power to freely choose its movements, it might find itself "torn in two directions": to move both towards and away from the fire. In an organism with the power of free choice, it is *not* the case that any mechanism that causes it to move away from the fire is as good as any other: mechanisms seem likely to be more effective if they cause the organism to behave in a way consistent with the free choices it would make. But if the organism already has the power of free choice then principle P tells us the peculiar mechanism of a nasty or unpleasant sensation will in itself have the effect of getting the organism to avoid the fire. Only if organisms have the power of free choice would subjectively pleasant and unpleasant feelings appear to "have a point".

We can perhaps make this point more vividly by considering, not the evolutionary development of an organism, but the ways in which it might be rational for the designer of a robot to include in it devices for controlling its behaviour. Suppose a designer is constructing a robot which, it is hoped, will be able to successfully get around in the world. It will, plausibly, need to contain within it mechanisms that cause it to move away from things that are harmful to it, and other mechanisms that cause it to seek out things that are beneficial to it. But the robot does not, we may suppose, have anything that could be called the

power of free choice: all its actions are determined by mechanisms in which nothing that could plausibly be called a power of free choice have any role. Now, let us suppose that the designer discovered that certain of the mechanisms within the robot generated subjective, conscious sensations. The designer discovered that they could produce, say, pleasant sensations in the robot by building certain types of mechanisms in to it, and unpleasant sensations by including others. There is no *a priori* reason why any unpleasant conscious sensations that arose within the robot should cause it to move away from the source of those sensations, and neither is there any *a priori* reason why it would exhibit behaviour that would increase the likelihood of pleasant sensations. The conscious sensations within the robot might be like "a wheel that turned even though nothing else turned with it." But the designer wants to ensure the robot moves away from things that are harmful to it, and towards the things that are good for it. Under the circumstances imagined, it seems there would be no point in the designer ensuring that that the robot felt unpleasant sensations in the presence of harmful things, and pleasure in the presence of beneficial things. There would be no point in this if these pleasant/unpleasant sensations did not *also* give rise to avoidance and pursuit respectively.

We can contrast this hypothetical robot with an organism or mechanism that *does* have the power of free choice. One way in which the constructer of a mechanism that did already have the power of free choice might get it to avoid harmful stimuli is by designing it in such a way that harmful stimuli produce in the mechanism subjectively unpleasant sensations. If the mechanism has a power of free choice then, by principle P above, the organism *would* tend to avoid harmful stimuli. Consequently, if the mechanism does have the power of free choice, then there would be point to the designer using the peculiar mechanism of subjectively pleasant and unpleasant conscious sensations to control the organism's behaviour.

As noted above, without the power of free choice, subjective pleasantness and unpleasantness may be a "like a wheel that turned even though nothing turned with it, that is, not a part of the mechanism". To continue this metaphor, the power of free choice is the cog that links up the wheel of subjective pleasant and unpleasant conscious experience with the wheel of physical behaviour. It is the component that makes subjective experiences a part of the mechanism.

Of course, we can imagine a mechanism being constructed that lacks the power of free choice, but which experiences, say, unpleasant sensations. We can also imagine the designer of the mechanism going on to construct it in such a way that these unpleasant sensations then gave rise to avoidance behaviour. But if the mechanism lacks the power of free choice, then there would seem to be no reason to use the *particular device* of unpleasant sensations, rather than one of many other possible mechanisms, to produce avoidance behaviour. Only if organisms, or mechanisms, *already* have a power of free choice does the subjectively unpleasant or nasty aspect of pain seem to "have a point". *Only if an organism already has the power of free choice does it become explicable why the peculiar mechanism of a nasty, unpleasant conscious feeling should have developed as the means by which avoidance behaviour is assured.*

Postulating a prior power of free choice in organisms enables us to give a possible explanation of the fortunate correlation. If organisms already possess a power of free choice, then they will *ceteris paribus* tend to seek out those stimuli that bring them pleasure and avoid those that bring them pain. So, provided that the *ceteris paribus* clause holds, if organisms already have a power of free choice, the fortunate correlation will tend to hold. Of course, this at most only provides an explanation of why seeking out/avoidance behaviour should be correlated with pleasure/pain. It does not explain why pleasure should

be correlated with seeking out things that are good for the organism, and pain with avoiding things that are harmful. To explain that correlation, the theory of evolution is presumably also required. And it is easy to see the general form such an explanation might take: Organisms that, by random variation, got pleasure from things that were beneficial to them tended to seek out those things and so had an increased chance of survival. Similarly those that got pain from things that were bad for them avoided them, and so tended to survive. In this way, natural selection ensured that those that survived got pleasure from things that were good for them and pain from things that were bad from them. But note: *this* is *not* the fortunate correlation. The fortunate correlation links feelings of pleasure with seeking out behaviour, and feelings of pain with avoidance behaviour. *That* correlation is not to be (directly) explained by evolution, but by hypothesising that those organisms in which it holds already had, prior to the correlation, a power of free choice.

8. THE NOTION OF FREE CHOICE USED HERE

On the view suggested in this paper, a "power of free choice" is used in explaining the fortunate correlation. But, it is natural to ask: "How is this notion of free choice to be defined?" Here, no definition of this notion will be offered. This does not mean, however, that the meaning of the term cannot be explained. Like many terms that cannot be given an analytic definition, it is possible to ostensively explain the meaning of "free choice" to which we have here appealed. We are all familiar with instances of the exercise of free choice in this sense: we all experience it when we, for example, voluntarily choose to withdraw our hand from a painful stimulus *because* it is painful. Experiences of this sort give us "knowledge by acquaintance" with free choice in this sense. We can *recognise* that we are exercising this faculty of free choice when we assert that we have voluntarily removed our hand because of the pain. So: the reference of the notion of free choice used here can plausibly be fixed by ostension, even if it cannot be analytically defined. Moreover, if we are able to recognise as such the conditions under which it is appropriate to assert sentences such as "I chose to move my hand away because it was painful" then it seems we have some sort of implicit knowledge of the meaning of the notion. And, of course, in these respects the notion of free choice used here is perhaps no different from many other significant philosophical notions: the notions of truth, knowledge, causation and consciousness, for example, have all been claimed to be primitive and not definable in more fundamental terms. The lack of a definition does not show the term to lack meaning, and neither ought it to prevent us from employing it.

It is worth observing that the conception of free choice used here can be incorporated within the model of free will developed by James[13]. On James's model, the exercise of free will is to be seen as a two-stage process.[14] In the first stage, a number of options become available to an agent. These options might be: to walk home one way rather than another, to put their hand near the fire to retrieve an object or to refrain from doing so, and so on. Both courses of action are possible, given the agent's present circumstances. In the second stage the agent freely chooses one of these options, where the choice made causally follows from the agent's psychological make-up, particularly their emotions and desires. Clearly, this second stage in James's model of free will is compatible with the view advocated here. On this view, an agent might, for example, freely choose to not retrieve the object from the fire because of the pain of doing so, and on the view advocated here, the avoidance of the pain involves the exercise of free will. So: on this view, the act is both caused by attributes of the agent's

psychological make-up, but is also the exercise of free-will: it therefore fits James's account "like a glove".

9. *A CONSIDERATION OF SOME OBJECTIONS*

On the view advocated here, there is a type of "free choice" or "free will" that is evolutionarily prior to the ability to feel pleasure and pain. So, it seems, this power of free choice must be present in a species *before* it develops the capacity to feel pain. But this, it might be objected, is not very plausible. A worm squirming on a hot rock, for example, presumably feels pain, but we would surely feel wary about attributing to the worm a power of free choice. The attribute of having a power of free choice, is, it might be protested, a higher or more developed cognitive capacity than the mere capacity to feel pain. And, if this is granted, we would surely expect it to come after, not before, the capacity to feel pain.

One reply to this objection is that the notion of free choice used here does not require any highly developed cognitive capacity. The only *explanatory* role played by the notion of free choice used here is that it gives an agent the *tendency* to move away from a painful stimulus because it is painful, and towards a pleasant stimulus because it is pleasant. It does not, for example, require an organism to make a conscious choice between, to deliberate between, or even to have an understanding of, a range of possible behavioural options. It is at least not obvious that such a minimal capacity must be evolutionarily posterior to the feeling of pleasure and pain.

It might be objected that the notion of free choice used here is no more genuinely explanatory than Moliere's "dormative virtue". What we want to explain is why organisms seek out pleasurable situations and avoid painful ones. The "explanation" given is that this is because organisms have a "power of free choice". But – so the objection may be made – the power of free choice used here is nothing more than a power that enables organisms to seek out the pleasurable and avoid the painful. And so the proposed explanation would seem to come to no more than: organisms seek out the pleasurable and avoid the painful because they have a power to do so. But that would hardly seem to be any "explanation" at all!

There are two things that can be said in reply to this objection. First, it is false that on the view offered here the *sole content* of the notion of a "power of free choice" used here is "a power to seek out the pleasurable and avoid the painful". Part of its content is also given ostensively: we refer to this power when we report that we have exercised it in, for example, choosing to move our hand away from a flame or towards a chocolate cake. We are, plausibly, *acquainted* with this type of free choice. Secondly, and perhaps more importantly, the view offered here also advances a thesis about evolutionary priority. It says that the type of free choice we exercise when moving our hand from the flame is evolutionarily prior to the fortunate correlation.

It might perhaps be objected that the explanation of the fortunate correlation offered here merely replaces one puzzling phenomenon without another: Our initial problem was to explain why the fortunate correlation exists, and, on the view offered here, we do so by postulating a "power of free choice". But this surely gives rise to the question: "Why does this power of free choice exist?" One puzzling phenomena is explained merely by postulating another.

In this paper an explanation of the existence of the power of free choice shall not be offered. However, three points may be noted. First, and most obviously, all explanation must stop somewhere: any explanation must, for example, leave certain laws as (at least for the time being) unexplained "brute

facts". Secondly, the problem of explaining the fortunate correlation has been shifted to the problem of explaining why organisms should have a power of free choice. And if it is accepted that organisms do indeed have this power (and introspection seems to reveal we do have it) its existence is something that we would have needed to explain quite independently of the claims of this paper. It is something we would need to have explained "anyhow". And so, in this respect, shifting the problem to explaining free will reduces the total number of things that require explanation. The third point to be noted is that the prospects for being able to find an evolutionary explanation of a power of free choice seem, at least on the face of it, to be rather better than those for finding an explanation of the fortunate correlation. One reason why the problem of explaining the latter perhaps seemed so intractable was that organisms for which the fortunate correlation held were behaviourally indistinguishable from those that found themselves trapped in a "nightmare world". The fortunate correlation would not seem to confer upon an organism any behavioural tendencies that would give it an increased chance of survival. But an organism's having a power of free choice *would* seem to make a difference to its behaviour. The prospects would seem rather better for finding an evolutionary explanation for organisms having a power of free choice.

Perhaps the objection most likely to be made to the view offered here is that, in asserting organisms have a power of free choice, it is asserting the existence of something that empirical research has revealed to be dubious at best. More specifically, it may be objected, empirical findings due to Benjamin Libet show the view we have free will to be highly questionable.

Libet's findings can be briefly summarised as follows.[15] Libet asked his subjects to do two things: (i) press a button at a time (supposedly freely) chosen by the subject and (ii) record the moment in time when they made the decision to press the button. Libet found that there was in the brain of each subject an electrical event of a particular type that always preceded the *formation of the decision* to press the button. On the face of it, such a discovery suggests both that the decision to press the button was caused by this earlier electrical event, and that the decision was therefore not freely made.

There are, however, two reasons for thinking that Libet's findings need not be decisive against the view adopted in this paper. First, and perhaps most obviously, the correct interpretation of Libet's findings is a matter of some controversy. While some workers in the field do see these empirical findings as showing the notion of free will to be illusory, not all do.[16] Some prominent commentators have argued that Libet's results are just what we ought to expect to be the case if we have free will.[17]

Secondly, it is not even clear that Libet's findings cast in to doubt the *specific kind* of "free will" or "free choice", appeal to which has been made in this paper. In Libet's experiments, subjects were asked to freely choose the moment at which they pressed a button. This action was not in response to either pleasure or pain; it was rather more like a wholly arbitrary choice. But the type of free choice that is the concern of this paper *is* exercised in response to pleasure and pain. It manifests itself when organisms move towards something because it is pleasurable or away from it because it is painful. It is free will in this sense that plays the causal role identified in James's argument against epiphenomenalism, and which is also the object of his theory of free will. And so the type of free will which Libet's findings (perhaps) show to be illusory would seem not to be the type of free choice we have utilised here.

10. EPIPHENOMENALISM AGAIN

On the view advocated here, in moving our hand away from the flame, we are exercising our capacity for free choice. We are choosing to avoid the *painful* flame in exercising this capacity. And we make this choice *because* the flame is painful. But in saying that we choose to move away *because* it is painful, we are clearly attributing causal powers to the subjective property of painfulness. And so this position is incompatible with epiphenomenalism.

But although the position advocated here is incompatible with epiphenomenalism, a main theme of the paper has been that to explain the fortunate correlation it is not sufficient to reject epiphenomenalism. Merely attributing causal powers to the mental is not enough: something more is need. In this paper it has been suggested that this "something more" is to say that a power of free choice is evolutionarily prior to the capacity to feel pleasure and pain.

CONCLUDING REMARKS

The aim of this paper has been to argue that the "fortunate correlation" is a puzzle in the philosophy of mind, the significance of which has perhaps not yet fully recognised. In particular, it raises problems not raised by apparently similar puzzles such as the inverted spectrum.

In the final section of the paper an explanation of the fortunate correlation was offered. The explanation says that a "power of free choice" is present on organisms that feel pleasure and pain, and this power must be evolutionary prior to the development of pleasure and pain. Only if a power of free choice is already present in organisms would the peculiar mechanism of a subjectively nasty feeling as a means of securing avoidance seem to "have a point". The suggestion was defended against some likely objections. But, whether or not the suggested explanation is accepted, the fortunate correlation perhaps deserves more attention than it has so far received.

University of Newcastle
John.wright@newcastle.edu.au

NOTES

[1] James argument was directed against Shadworth Hodgson, T. H. Huxley, Herbert Spencer and W. K. Clifford.

According to Adela Pinch (Thinking About Other People in the Nineteenth Century, p.68) , the thinker who initiated discussion of epiphenomenalism in this period was Shadworth Hodgson. Hodgson's views were developed in his *Time and Space: A Metaphysical Essay* (London, Longman, 1865). Hodgson saw mental properties as being analogous to the colours on the tiles of a mosaic. It is the shape of the pieces in the mosaic that are causally relevant in holding the mosaic in place: the colours of the pieces are causally irrelevant in doing this. In an analogous way, Hodgson saw mental properties as causally irrelevant.

T. H. Huxley embraced a view according to which animals and humans were "automata". Huxley famously compared mental phenomena to the steam given off by a steam train. His views were developed in his "On the Hypothesis that Animals are Automata, and Its History" in *Methods and Results: Essays by Thomas H. Huxley*, (New York: D. Appleton and Company, 1898.)

W. K. Clifford's *Body and Mind* (*The Fortnightly Review*, 16, December 1874, pp.714-736) contains a number of passages in which Clifford certainly appears to explicitly endorse mind-body parallelism. (Although in the opinion of the present author it is not entirely clear that Clifford actually does subscribe to what we usually regard as parallelism, or to a form of "double-aspect" theory.)

[2] William James *The Principles of Psychology* (H. Holt, 1890)

[3] The expression "felicitous alignment" seems to be originally due to W. S. Robinson in his article "Epiphenomenalism" in the *Stanford Encyclopaedia of Philosophy*.

[4] James, *The Principles of Psychology*, Chapter V The Automaton Theory, Section: "Reasons against the Theory".

[5] See Popper, K. and Eccles, J. *The Self and Its Brain* (New York, Springer-Verlag, 1977) pp.72-75.

[6]See Frank Jackson "Epiphenomenal Qualia" in *Philosophical Quarterly* vol. 32, 1982, pp,127-136, especially p.134.

[7] See Jack C. Lyons "In Defence of Epiphenomenalism" in *Philosophical Psychology* vol. 19, no. 6, 2006, pp.767-794.

[8] See William S. Robinson http://www.public.iastate.edu/~wsrob/EvoEpi.pdf

[9] See Milic Capek "James' Early Criticism of the Automaton Theory" in *Journal of the History of Ideas*, XV, (April, 1954), pp.260-279, especially p.267.

[10] This is the type of explanation favoured by David Chalmers in his *The Conscious Mind*, esp. p. 158.

[11] It is worth noting that in his paper "Evolution and Epiphenomenalism" William Robinson also comes to the conclusion that James' felicitous alignment presents a problem for all views of the mind-body relationship. However, the reasons Robinson gives for this are different from those used here.

[12] Capek (*op cit*) notes that there are numerous counter-examples to the suggestion that what we find pleasurable tends to be good for us, but notes that statistically the tendency is for the pleasurable activities to be healthful. The same can surely be said for the "fortunate correlation" with which we are presently concerned: statistically the tendency is surely for us to tend to seek out the pleasurable and avoid the painful.

[13] James explains his model of free will in his "The Dilemma of Determinism", *Unitarian Review*, vol XXII (1884), p.193.

[14] A discussion James' two-stage conception of free-will, and of the way it pre-dates similar conceptions, is given in Bob Doyle "Jamesian Free-will, the Two-Stage Model of William James" in *William James Studies*, 2010, vol 5, pp.1-28.

[15] See, for example, Benjamin Libet et al "Subjective referral of the timing for a conscious sensory experience: a functional role for the somatosensory specific projection system in man." *Brain*, **102** (1979) (1): pp.193-224. See also "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action" in *The Behavioural and Brain Sciences*, vol 8, pp.529-566.

[16] For a defence of the view that Libet's findings do not cast free will in to doubt, see, for example Owen Flanagan "Conscious Inessentialism and the Epiphenomenalist Suspicion" in *The Nature of Consciousness* edited by Ned Block, Owen Flanagan and Guven Guzeldere, MIT Press, A Bradford Book (1997)

[17] See Flanagan *loc cit*.