IN DEFENCE OF INTERACTIONISM

IN DEFENCE OF INTERACTIONISM

OLE ANDREAS KLÆBOE KOKSVIK
B.A.

Submitted for the degree of Master of Arts
Department of Philosophy
Monash University, September 2006

The dualism of the common man holds that experiences are nonphysical phenomena which are the causes of a familiar syndrome of physical as well as nonphysical effects. This dualism is a worthy opponent, daring to face empirical refutation, and in due time it will be rendered incredible by the continuing advance of physicalistic explanation.

DAVID LEWIS
'An Argument for the Identity Theory'

Abstract

Mind–body dualism is intuitively a plausible position, the important contemporary varieties of which are epiphenomenalism and interactionism. The former acknowledges causal relations only *from* the body *to* the mind; the latter insists that the causal relations go both ways. This thesis defends interactionist dualism against a set of commonly raised objections.

Given dualism about phenomenal experience I argue that interactionism has a decisive advantage over epiphenomenalism, first because epiphenomenalism clashes thoroughly with widespread and entrenched intuitions about motivation, secondly because epiphenomenalism leads to the incredible conclusion that whether or not judgements about phenomenal experience are justified fails to depend on the experiences they are about.

The advantage I argue that interactionism holds relies on the causal efficacy of phenomenal experience. I consider and reject a conceptual argument which charges interactionism with being unable to deliver this advantage.

Interactionism is often thought to be unsustainable on the grounds of conflict with results from science. One of these is the law of conservation of energy. I argue, however, that there are well supported and useful formulations of that law that do not exclude interactionism.

The thesis that the physical world is causally closed is a result with which interactionism is indeed incompatible. I argue that we lack credible evidence for the thesis, and that the arguments that are supposed to show that the thesis is true rely on a premise the dualist has no reason to accept.

I conclude that interactionism is an attractive and viable position in contemporary philosophy of mind.

Contents

Statement

This thesis contains no work which has been submitted for the award of any other degree or diploma in any other university or other institution.

To the best of my knowledge, this thesis contains no material previously published or written by any other person, except where due reference is made in the text.

OLE KOKSVIK
12th of February 2007

Note on Language

Australia is, along with the other countries in the 'first world', a society in which half the population is effectively discriminated against. Anyone who can read statistics can easily identify numerous areas in which women occupy unequal positions in relation to men; being paid less and being underrepresented in nearly all the places of power are just two examples.

Biased language encourages biased thought. It is therefore the clear responsibility of all those whose work is the written word to ensure that bias is not perpetuated by our hand.

This thesis is prepared with the goal of not perpetuating sexist bias through the use of language. I have found the American Philosophical Association's 'Guidelines on the Nonsexist Use of Language' (Warren 1986) helpful in this regard. For any sexist use of language that remains, I apologise.

Acknowledgements

I have been fortunate enough to be supervised by three great philosophers: Ian Gold, Graham Oppy and John Bigelow. I thank them for sharing their time, knowledge and advice so generously, and for the many careful and valuable comments which each, at various stages, provided to numerous earlier versions of the manuscript. They all made many helpful suggestions, many of which I have gratefully adopted. A special thanks to John for his tireless and patient efforts in the crucial final stages.

The philosophy department at Monash is a friendly and encouraging place, where students are treated more as peers than as underlings; I am very grateful to those who constitute it for making it so. Throughout the candidature I have benefited from conversations with many of the faculty, *inter alia* with Dirk Baltzly, Sam Butchart and Monima Chadha. Lynda Burns, Toby Handfield and Laura Schroeter read a chapter each, and provided useful comments.

I am grateful to my fellow postgraduates at Monash. In addition to good discussions, helpful comments and questions at work–in–progress presentations, they also provided a supportive environment — at times suffused with a great deal of excitement about research and teaching in philosophy — without which my candidature would have been much less enjoyable.

Erik Brown at the University of Bergen was my first philosophical hero. His clarity of thought and presentation much impressed me.

My parents, Merete and Arne, and my sisters, Kristine and Kari, are open–minded people who have been unfailingly encouraging of me. My home was one in which the serious exchange of opinion was appreciated and consistently practised. Only later have I come to realise how valuable that is.

My mother–, father– and sister–in–law, Lorraine, Chubby and Sue, have never let me doubt that I am part of their family. They did much to make Australia a home for me in ways that only family can, as did also the large extended family here.

I owe my greatest debt of gratitude to Jo-Anne Weinman, my wife. For your unfailing support, encouragement, care and patience, for small and large kindnesses, numerous and generously given: my thanks.

Chapter 1: Interactionism

1.1 Introduction

Mind-body dualism is, on the face of things, an attractive position. Before philosophical inquiry begins, we have on the one hand interacted with a multitude of different physical objects, and on the other enjoyed a rich 'inner life', abounding with thoughts, emotions, experiences and so forth. That the two *seem very different* is rarely disputed.[1] What is disputed is how much weight this appearance should be given when we try to find out what the world is like.

Appearance of difference does not always track real difference in the world. *Wind* appears differently to us than do most physical objects of our experience. We cannot see wind, but we can see the effects it has, for sometimes leaves, branches, sand or fog move in the wind. We are used to seeing the causes of physical changes in our environments, so in this respect (and others) wind is unusual. We do not, however, take this to indicate that wind is deeply different from physical objects of our everyday experience. Wind is air that is moving. We know how air is composed, and although we cannot make accurate small-scale predictions, we can with reasonable accuracy predict what the general direction and strength of wind will be on most days. Moreover, even in those few instances — such as tornadoes — where even large-scale predictions elude us, there is widespread consensus that it is only the large number of variables and the complexity of their interactions that stop us. Wind, although invisible, is no longer deeply mysterious to us.

Mind-body dualism therefore requires considerations to motivate the belief that the appearance of difference between mind and body runs deeper

[1] Rarely, but sometimes. In 'The Headless Woman Illusion and the Defence of Materialism' Armstrong (1968) can be understood as suggesting that rather than being aware of apparent difference, we are *un*aware of apparent similarity, between the two. "It can now be suggested by the Materialist that we tend to pass from something that is true: 'I am not introspectively aware that mental images are brain-processes' to something that is false: 'I am introspectively aware that mental images are not brain-processes'." (pp. 48-49). I have added single quotation marks for clarity.

than the appearance of difference between wind and everyday physical objects such as rocks and books. There is, in other words, a need to show that we are now *not merely* in a position *vis-à-vis* the mind similar to that which people *were* in *vis-à-vis* wind, when wind was still mysterious to them.

Mind-body dualists are usually taken to believe that the difference between mind and body is deep, or even fundamental. One consideration that may underpin such a belief is the following. A pre-scientific encounter with wind might have been characterised by a sense of a mysterious and unseen force that moves branches and leaves. That force has now been explained by, as it were, constitution; we have an account of what 'constitutes' the 'force' that moves the branches, and of why it is that we cannot see it. With the mind, however, many dualists hold that no analogous explanation *could* be forthcoming, because any attempt at giving an explanation in terms of (say) atoms and how they move would amount to changing the subject, in the strong sense of changing what we are talking about, not just how we talk about it.

We know a fair bit about the *physical requirements* for our inner life. The indications that our inner life is strongly connected with the brain are very clear; we can for example 'turn off' the conscious part of a person's inner life by giving them a solid knock to the head. Evidence to show that damage to a specific part of the brain is likely to cause specific changes in the mental life also abounds. Yet, mind-body dualists think that the appearance of difference is not compromised by this knowledge of the physical requirements for our inner life. Even though we know a lot about the working of the brain, we would be wrong, they claim, to say that *that* is what we are talking about when we report on our inner life. It is not as if we are inaccurate when we say e.g. 'I feel sad', and really should have stated something about the neural activity in the brain. What is referred to by expressions such as 'I feel sad' *cannot* turn out to refer to brain states, or so the dualist is said to think.

This report, 'I feel sad', is carefully chosen, for it is most common today to be a dualist about only a limited section of our inner life. 'I feel sad' is taken to report on a state within that section. What most dualists believe today is that there are some properties of experience that are not physical. There are many names for these properties, *phenomenal experience* is one, *qualia* is another, *what it is like to be* (someone who has a certain experience) is yet another. The thought is that there is something it is like to have an experience, there is a something it *feels* like to see a colour, hear a sound, smell something. This is something it is very hard to argue for. If you were inclined to deny that there really is something it is like to have an experience, there is not much I could do to convince you to change your mind. For the purposes of this thesis it is assumed that phenomenal experience really is a property of experience. That is not too controversial.

It will furthermore be assumed that this property of experience, the '*what it is like*' to have that experience, is *not physical*. It will be assumed, in other words, that any attempt to replace reports like "I feel sad" with a report like "My brain is in such-and-such a state" *has to* fail; the two reports would not be co-referential. Someone would have changed the subject.[2] That is *very* controversial. In fact, a significant proportion of recent philosophy of mind has been occupied by argument over this point. Nevertheless, I think it is worthwhile to make that assumption, and I will try to explain why.

In philosophy of mind *monism* — the view that mind and body are fundamentally similar — is the received view. Jackson and Braddon-Mitchell exaggerate when they claim that "dualism is almost as unpopular as idealism" (1996, p. 4); a publication such as Chalmers' *The Conscious Mind* is either a symptom or a cause of a resurgence in the popularity of dualism; most likely a

[2] Again, this is meant in a strong sense. On some accounts the use of a different predicate guarantees in itself that a different property is being predicated. On such accounts it comes out as trivially true that changing the 'mode of referring' from "I feel sad" to "My brain is in such-and-such a state" changes the predicate. I assume that changing the sense does not necessarily suffice to change the reference, and thus that the claim of having changed the subject in this example is non-trivial.

bit of both. Furthermore, even among monists it is surely true that dualism, in one form or another, has *some* pull on them *some* of the time. The intuition that mind and body are very different from one another, perhaps fundamentally different, is ubiquitous and compelling. So dualism is not nearly as unpopular as idealism. However, though it does not rule the scene as unopposed as Jackson and Braddon-Mitchell claim, they are right in saying that monism is orthodoxy.

Against this backdrop the decision to *assume* that dualism is true is perhaps surprising. Less so, hopefully, once it is realised that the dominance of the monist position does not rest solely on the merits of that position itself. Rather, it depends, to a not inconsiderable extent, on the elimination of dualist positions as unviable. Many philosophers believe that no credible dualist position is available. If none were, it would be unreasonable to blame someone for retreating to a (presumably) less problematic monist position instead. The purpose of this thesis is not to attack the merits of monist standpoints in the philosophy of mind, but rather to question whether the dismissal of a particular dualist position — interactionism — can be justified in the way that it is widely supposed that it can.

Assuming the truth of dualism for the purposes of this thesis is justified, because the widespread belief that interactionism is unviable rests to a very large extent on arguments that are inapplicable to other dualist positions. The assumption does not, therefore, lend interactionism any unfair dialectical advantage. It serves merely to focus the discussion in the following way: it opens for an evaluation of whether interactionism can be defended from the attacks that it is *actually* subjected to. It remains the case, of course, that any general or principled criticism that can be levelled against *any* dualist position applies just as forcefully to interactionism. This is, however, as a matter of fact not how the criticism of interactionism usually proceeds. Interactionism is criticised for claims that are specific to *that* position, and crossed off the list over

available dualist alternatives. Only when other dualist alternatives are crossed off as well is the conclusion reached that monism must be true. Interactionism is, in other words, to a very large extent criticised *qua* interactionism, and not *qua* dualism. Assuming that dualism is true for the purposes of evaluating the viability of interactionism takes this fact seriously.

In addition to leaving any general or principled criticism of interactionism *qua* dualist position intact, defending interactionism from an argument widely used to discredit it does not, in itself, contribute to *motivate* belief in the position. A comprehensive argument for interactionism would contain *either* arguments that lead directly to interactionism (if such could be devised) *or* arguments that lead first to dualism and then, in the second step, to interactionism as the best dualist position. The present thesis comes closer to employing the second of these strategies, since epiphenomenalism — an alternative dualist position — is criticised in some detail herein. But the other half of the strategy is conspicuous in its absence; I make no independent attempt to motivate dualism here.

Interesting questions can be raised about the viability of the strategy of motivating interactionist dualism in two steps. It is interesting, in particular, to ask whether the arguments that these days are used to motivate dualism could constitute the first step of such a two step strategy.[3] The question is whether these arguments would survive the move to interactionism in the second step; there may be some reason to doubt that they would. For example, one might doubt if the knowledge argument could go through exactly as before.

That argument — the important features of which were present already in Dunne's *An Experiment with Time* (1927/1958), but which was since put into its oft-cited form by Jackson (1982) — describes a brilliant scientist who is confined to a monochrome environment, but who nevertheless studies, among

[3] I am grateful to Leon Leontyev for relentlessly driving this point home to me.

other things, colour perception. The intuition is elicited that when she escapes her confinement she *learns something new*, namely *what it is like* to have a colour experience, and that this is something she could not have known beforehand. However, if phenomenal experience is causally efficacious, as interactionism holds that it is, then Mary *would* know about phenomenal experiences while she was still confined, at least in the sense that she could 'pick them out'. She would be able to refer to them using unique identity conditions, for she would know their place in the causal web of things. She could not pick them out in the same way that most others pick out colour experiences, of course, and there is still a powerful intuition that there would be *something* she would not know. How important that something would be, and whether it would be enough for the argument to go through, is uncertain. Perhaps similar worries could be raised about other arguments for dualism.

I set this question to one side, however, for whether a two-step argument for interactionist dualism is possible is tangential to the current project. The aim of this thesis is not to complete the ambitious project of building a comprehensive case in favour of interactionist dualism. For the purposes of this thesis it is more than sufficient to note that a two-step argument for interactionist dualism does not *obviously* fail, and that interactionism in any case has considerable intuitive attractiveness.

The aims of this thesis are quite modest. The overall aim of the project is just to show that interactionist dualism is a 'live option' in the philosophy of mind, one that deserves more attention and investigation than it is, at present, getting. First it is argued that epiphenomenalism — the currently dominating dualist position — lacks important and attractive features that interactionism has. This motivates the claim that if both positions are viable we have good reason to accept interactionism over epiphenomenalism. Secondly I attempt to show that interactionism is a viable position by showing that the objections most commonly raised against the position hold little force.

In conjunction, these two steps go some way toward showing that interactionist dualism is the most attractive dualist position. Since the starting point is the assumption that dualism is true they can obviously at most contribute toward half the job of showing that interactionist dualism is the most attractive position in the philosophy of mind overall. However, if the claim that monism wins a not inconsiderable proportion of its support by means of elimination of dualist alternatives should be accepted, it is hoped that the merits of showing that one such alternative is not unviable (or at least not unviable for the reasons most commonly given) will be seen as well. If interactionist dualism comes to be seen as a real alternative position, it seems that this may have the potential to impact the debate in philosophy of mind quite significantly.

1.2 How Monism Wins Popularity

It is not easy to argue directly for the claim that the popularity of monism is owed largely to a lack of alternatives, without taking an opinion poll. A good case can be made for the claim in indirect ways, however, for the process that leads to the 'crossing out' of one dualist position after another — and thus eventually to monism — is not too hard to reconstruct.

1.2.1 FOUR DUALIST POSITIONS

In respect to causality, the logical space governing the relation between the mental and the physical contains but four types of dualist positions:[4]

i. There are no causal relations between the mental and the physical.

ii. There are causal relations between the mental and the physical such that mental states of affairs bring about physical states of affairs.

[4] This thesis does not discuss what Chalmers (2002) calls 'Type-F monism'. Aside from the standard excuse that one cannot do everything at once, the most important reason for this stems from the objection that the position seems to fail to cater for the core intuition that the mind *makes a difference* to the physical in a *direct* sense. It seems that the 'subtle' causal relevance that Type-F monism (1996, pp. 153-56) allots to phenomenal properties is *too* subtle to account for our intuitions. Chalmers notes the problem: "It remains the case that natural supervenience *feels* epiphenomenalistic" (p. 156).

iii. There are causal relations between the mental and the physical such that physical states of affairs bring about mental states of affairs.

iv. There are causal relations between the mental and the physical going both ways. That is, there are causal relations between the mental and the physical such that mental states of affairs bring about physical states of affairs, and such that physical states of affairs bring about mental states of affairs.

The first of these positions shoulders a very significant explanatory burden, namely that of explaining why it constantly *seems* to us that the there are causal relations between the mental and physical. This position was famously defended by Leibniz. He argued that God had established a harmony between the mental and the physical from the outset, so that the two domains run in parallel. Whenever there is a wish to, say, raise one's arm, there is also, according to parallelism, operations of the mechanisms in the physical world such that the arm is indeed raised. All of this takes place without any interaction between the two domains. It is safe to say that pre-established harmony strikes most contemporary philosophers as a very implausible doctrine, and I know of no remaining defenders of that or any other form of parallelism.

The second position is not to be confused with idealism. Idealism is a form of monism, but unlike the materialist or physicalist monism now popular, idealism holds that everything is mental. In contrast, the position under consideration is a *dualist* position, one that holds that physical states of (say) human bodies are brought about by mental states, but never the other way around. The position should further be differentiated from occasionalism, one form of which holds that all states of affairs are brought about by God through continuing creation. Occasionalism is similar to the position presently under consideration in that causation between mind and matter is limited to the one

direction *from* mind *to* matter. However the view under consideration is not a view about God, but about finite creatures that have both bodies and minds, and the causal relations between those.

This position has to my knowledge never been defended, although it appears to be a possible stance. One might hold that the (immaterial) soul causes the body to grow and develop, and that the soul controls the body throughout the life of a person.

Considerations mentioned above can be brought to bear against this position. There is abundant *prima facie* evidence for the existence of causal relations from the physical to the mental in our everyday life, so it is very hard to believe that causal relations should be limited to the direction from the mental to the physical. It seems safe to conclude that neither parallelism nor the position just discussed are viable dualist positions. They can safely be set to one side.

1.2.2 CAUSAL CLOSURE

After this first brief round of exclusions we are left with only the last two dualist alternatives. The fourth position is interactionist dualism. The minimal interactionist position is that the mind either sometimes brings about a physical event that would not otherwise have occurred or brings it about that an event occurs *differently* than it otherwise would have. A single, very influential objection is generally taken to put this alternative out of the running for good. The objection relies on what is known as the principle of *causal closure* of the physical world. Even though we are far from completing the scientific project, we know, it is claimed, enough about the world to know that all physical phenomena are fully explainable in terms of other physical phenomena; they have only physical causes. That the principle of causal closure does not rely on

determinism is widely recognised;[5] allowing for indeterminacy you might say that *inasmuch* as a physical event can be said to be an *effect*, inasmuch as the event has causes, all those causes are physical.

It is not hard to see how the principle of causal closure leads to the denial of interactionism. There are two domains, one is physical and the other is not. The principle says that the physical domain is causally closed; that nothing outside of that domain has any causal influence on anything within that domain. Then it follows straightforwardly that the *other* domain cannot have any causal efficacy on the physical domain. That some aspect of the mental has such causal efficacy is precisely the minimal claim an interactionist must make, so the principle of causal closure leads to the rejection of interactionism.

The belief that the objection from causal closure is fatal to interactionism has been extremely influential. Thus, for example, Papineau argues that "[a]ll physical effects are fully caused by purely *physical* prior histories" (2002, p. 17), Jackson and Braddon-Mitchell state that "[a]ll the evidence suggests that there is nothing non-physical ... that steps in and somehow regulates the physical world's goings-on" (1996, p. 9), Chalmers writes that "[t]he best evidence of contemporary science tells us that the physical world is more or less causally closed: for every physical event, there is a physical sufficient cause" (1996, p. 125) and Lewis endorses what he calls the "materialistic working hypothesis", namely that "physical phenomena have none but purely physical explanations" (1966, p. 17). With the exclusion of Chalmers, who has since changed his mind (2002, n. 26), these theorists believe that interactionism is excluded by causal closure.

It is worth noting that there is a significant difference between the formulations just mentioned. Lewis' and Jackson and Braddon-Mitchell's formulations expressly *exclude* non-physical causes for physical phenomena,

[5] E.g. by Papineau (2002, p. 17, n. 2), Jackson and Braddon-Mitchell (1996, p. 13) and Chalmers (1996, p. 150).

there *are no* entities such that they are both causes of physical phenomena and themselves non-physical. Chalmers' formulation here does not; it states merely that for all physical effects there *are* sufficient causes that themselves are physical, and Papineau's formulation is also restricted to an existential claim.

Formulations that assert the existence of sufficient physical causes for all physical effects, rather than deny the existence of non-physical causes for physical effects, open the door to the thesis of *causal overdetermination*. This is the thesis that some events, and in particular those which we tend to believe have mental causes, may have *two* sets of causes, each in itself sufficient to bring about the effect. If that were so it looks like interactionism might not be incompatible with the causal closure thesis defined as an existential claim, for the assertion that *there are* sufficient physical causes for all physical effects is compatible with there being sufficient non-physical causes for some physical events, as well.

It will be argued in this thesis that it is important to direct close attention to the investigation of interactionism as an alternative position on the mind-body issue, and in particular to the scrutiny of the causal closure thesis. It may seem to follow that further research into causal overdetermination should also be recommended. That is not the recommendation, for two reasons. First, while the investigation of causal overdetermination is not a large research project it is fair to say that it has received much more attention (recently) than the project of investigating the viability of interactionism. But even more importantly, causal overdetermination does not seem to deliver enough 'oomph' to the mental for it to account for our intuitions about the causal efficacy of phenomenal experience. It does not offer a significant improvement over epiphenomenalism. As Robinson points out, the central intuition is "that the mental *makes a difference* to the physical, i.e. that it leads to behaviour that would not have happened in absence of the mental" (2003), and that intuition is not appeased by causal overdetermination. Consequently, in what follows, all mention of the causal

closure thesis is to be understood as making reference to a thesis that is incompatible with causal overdetermination, to a thesis, that is, that expressly denies the existence of non-physical causes for physical events.

1.2.3 EPIPHENOMENALISM

The exclusion of interactionism on the grounds that it conflicts with the principle of causal closure thus leaves only one dualist position: epiphenomenalism. According to epiphenomenalism the phenomenal properties of experience — what it feels like to have an experience — have no causal efficacy upon our actions (or upon anything else that is physical). On this view, our non-physical mental life arises from the physical material in the brain, and the changes and events that take place there, so there are causal relations going *from* the physical *to* the mental. The epiphenomenalist holds, however, that it is those physical changes and events themselves that account for all our bodily behaviour; they do not countenance causal relations going from the mental to the physical. Epiphenomenalism has sometimes been marketed as a 'safe' way to be a dualist. An epiphenomenalist do not have to engage in bothersome speculations about how something non-physical can cause a change in something physical — speculations that lead Descartes to the much ridiculed conclusion that the 'cross over' point is in the pituitary gland.[6] One can be a dualist without having to sound like someone who believes in fairies, as Jackson once put it (1982, p. 128).

In this thesis it will be argued that epiphenomenalism is an extremely unattractive position. If that is right it appears that the dualist has no escape; there is no attractive and tenable position left. In reference to the point of how monism wins popularity it may be noted that even if it is not *true* that there is no tenable and attractive dualist position left for a dualist to hold, the *widespread*

[6] In *The Passions of the Soul* he wrote about the pituitary gland: "[T]he activity of the soul consists entirely in the fact that simply by willing something it brings it about that the little gland to which it is closely joined moves in the manner required to produce the effect corresponding to this volition" (1649/1985, p. 343).

belief that this is so goes a very long way toward explaining why monism dominates to the extent that it does. Or so I claim.

Given first and foremost that dualism enjoys intuitive plausibility, and secondly that it has enjoyed renewed popularity of late, considerations that leave the dualist with no tenable position serve to highlight the importance of two projects of inquiry. First, it naturally becomes very important to investigate whether or not epiphenomenalism is as untenable as it may seem. Perhaps the intuitions against epiphenomenalism cannot be translated into rigorous arguments against that position. If so, perhaps the *mere* fact that epiphenomenalism is a counterintuitive position should not lead us to abandon it. This project has already received significant attention. When there appears to be no tenable dualist alternatives left, that situation should *also*, however, alert us to a *second* important project for further research, one that has received far less attention than it deserves.

That project is, of course, the project of determining whether *interactionism* really is an untenable position for the dualist to hold. Do we really have good reasons to conclude that interactionism is incompatible with current scientific knowledge? Are there other solid arguments against the position? Above it was claimed that, as a matter of fact, the most influential objection against interactionism has been that it is incompatible with the principle of causal closure. If that is true it follows that an important project is to investigate whether or not we have good reason to believe that the principle of causal closure is *true*. Or, perhaps even more to the point, it should be investigated whether we have *as good* reason to believe in the principle as the majority of contemporary philosophers of mind seem to believe or assume that we have.

1.3 Epiphenomenalism Is an Unattractive Position

The importance of investigating the viability of interactionism — and thus also the importance of investigating the credibility of the principle of causal closure — depends on the claim made above, namely that epiphenomenalism is an unattractive position. The most accessible reason to reject epiphenomenalism is also the most compelling. It is simply that epiphenomenalism is completely at odds with what most people are convinced is going on all the time in their everyday lives; it clashes thoroughly with our commonsense view of the world.

How so? If you were to ask someone (perhaps, to be safe, someone unpolluted by philosophy) *why* they did something — why they did a bungee jump or jumped out of a plane with a parachute or had a piece of cake or worked out or got their hair done or whatever — then their eventual answer would, in a very large proportion of the cases, be that they did it because it, or something it facilitates, *feels good* somehow. It feels good to have a big adrenalin rush, chocolate tastes good, it feels good to do physical training (or at least to have done it), it feels good to be satisfied with your appearance. Similarly with avoidance behaviour; to ask for an explanation of why we avoid humiliating experiences amounts to admitting that you do not understand what 'humiliating' means. It feels patently *bad* to be humiliated, so we try to avoid it.[7]

This is not to say, of course, that the motivation for all our behaviour can be explained this simply. One of the marks of maturity is the ability to forego an immediate or very proximate desire-satisfaction in favour of one that is a more distant but greater. This ability obviously affects our behaviour in profound ways. We keep working, for example, even when our immediate preference is to go to the beach, and we sometimes refrain from eating tasty foods in order to maintain our health. Nearly as obviously, this ability affects the way we *speak*. Envisage the following (rather caricatured) conversation with a young person:

[7] There are exceptions to this, stemming *inter alia* from certain sexual fetishes. It is no less likely in those cases, however, that phenomenal experience — a certain very arousing phenomenal experience, probably — plays a crucial role in motivating the (unusual) behaviour of seeking out humiliating experiences.

You:	I am sorry, we cannot start that game now; I have to leave very soon.
Child:	Why?
You:	Because I have to be at the university by one o'clock.
Child:	Why?
You:	Because I want to go to a lecture then.
Child:	Why?
You:	Because it is important to be there.
Child:	Why?
You:	Because it will help me get a good grade in this subject.
Child:	So?
You:	I want a good grade in this subject.
Child:	Why?
You:	Because it will help me get a good job.
Child:	So?
You:	Getting a good job is important.
Child:	Why?
You:	A good job will help me care for my family.
Child:	So?
You:	Well, if I cannot care for my family, I will be very sad, and they will be very sad too, because they will not get the things they need.
Child:	Oh, OK.

Here the account of the motivation is gradually being 'pushed backwards', until finally a point is reached where the interlocutor *understands* why; a motivational factor the child shares has been reached. Speaking with young people is a good thing, and speaking with people who hold very different opinions is not bad either. When we speak with people who share our ambitions and values we are not usually challenged to justify our actions, and this can lead to complacency about the value of what we are doing. Sometimes the 'unpacking' of motivations is a complicated and time consuming process. That, however, does

not change the fact that at the bottom of very nearly all motivation is phenomenal experience.[8]

Thus interactionist dualism is arguably very well aligned with common sense; both hold that phenomenal experience plays a causal role upon our behaviour. When it is taken into account how successful a theory commonsense is, this weighs very heavily in favour of interactionist dualism.[9]

That phenomenal experience has causal efficacy on our behaviour is precisely what epiphenomenalism denies. According to epiphenomenalism, the fact that the chocolate tastes *good* and the fact that it feels *bad* to be scared or humiliated has *no* causal influence on your behaviour. In particular, it does not explain why you buy chocolate or avoid heights and belittling people. That is incredible. Inasmuch as epiphenomenalist dualism is committed to this claim, it is an incredible position, too.[10]

One standard reply from the epiphenomenalist to this challenge needs to be mentioned briefly. According to this reply certain brain states have dual effect; they cause *both* a phenomenal experience *and* an action (the latter of which may be stipulated to occur further downstream, after the intermediate occurrence of further brain states, neural impulses, muscular contractions and

---

[8] Of course, we sometimes do things for others. In those cases perhaps we are motivated by the phenomenal experience we intend to elicit in our beneficiaries, or perhaps we are still motivated by a phenomenal experience in ourselves (the latter claim tempts some to deny the existence of true altruism).

[9] For eloquent praise of the "extraordinary predictive power" of commonsense psychology, see *Psychosemantics* (Fodor 1987, pp. 1-10).

[10] Similar objections to epiphenomenalism are widespread in the literature. Kneale argues that "the great paradox of epiphenomenalism" is "the suggestion that we are necessarily mistaken in all our ordinary thought about human action" (1959, p. 453); and that "the proposition ... that mental events are sometimes causes of physical events, is one which belongs to the hard core of common sense" (p. 454); Taylor characterises the thesis that "all bodily behavior is caused by bodily processes alone" as "quite impossible to believe" (1963, p. 23); Shaffer argues that "[i]nsofar as Epiphenomenalism asserts that mental events are never causes, it does seem to be in flat contradiction to our ordinary descriptions of familiar experiences" (1965, p. 101); Lewis argues that if the phenomenal difference between two tastes is epiphenomenal, that "makes it very queer, and repugnant to good sense" (1988/1999, p. 285); and Jackson and Braddon-Mitchell argue that, not only do we have good reason to believe that phenomenal experience, as a matter of fact, causally influences our behaviour, but a propensity to cause certain actions is part and parcel of many of our concepts for phenomenal experience (1996, p. 6), a thought which Lewis also entertains (1995, p. 141). (Kneale, Taylor and Shaffer are expressing opposition to epiphenomenalism as a thesis about a broader range of mental states than the thesis I criticise in the text advocates.)

so on). Because the phenomenal experience is *reliably correlated* with the kind of brain state that causes physical action downstream, it is not surprising that we should come to *believe* that phenomenal experience is causally efficacious on our actions. Nevertheless, so the argument goes, it is fully plausible that the causal relation is an illusion.[11]

The persuasive force of this argument varies very significantly. Some philosophers seem willing to accept the conclusion that phenomenal experience only *seems* to matter for action, and that this appearance is illusion. It does not, however, seem at all plausible to *me* that the causal connection I perceive between the certain experiences and my subsequent behaviour is illusory. Furthermore, it does not seem any more reasonable to suppose that the actions of the people I relate to in my daily life are not similarly influenced by *their* phenomenal experience than it does to suppose that they have no phenomenal experience at all. My belief that there is such a connection is, one might say, as entrenched as my belief in other minds.

For the present purposes, however, whether a solid argument can be mounted to replace these vague musings is not quite to the point. What matters more is that the vast majority of the 'folk' — as well as a fair few philosophers — are entirely convinced that the causal connection is real and not illusory. That does not mean that that belief cannot be mistaken. It *does* mean, however, that denying that belief is a significant theoretical burden, and one that certainly counts heavily against the epiphenomenalist position.

1.4 Burden of Proof

This leads to the question of what the burden of proof for a theoretical position about the mind-body issue should be. In the current dialectic climate one might easily get the impression that interactionist dualism is so outlandish and crazy

[11] See e.g. 'Epiphenomenal Qualia' (Jackson 1982, p. 133), *The Conscious Mind* (Chalmers 1996, pp. 181-82), 'Epiphenomenalism' (Robinson 2003).

that someone wishing to raise the mere possibility that it might be true must present arguments truly out of the ordinary. One cannot help but feeling, sometimes, that this impression is deliberately cultivated. For example, the title of Jackson's review article of *The Self and Its Brain* (Popper and Eccles 1977) was 'Interactionism Revived?' (1980), perhaps a somewhat tendentious choice of words. Similarly, to say that "dualism is almost as unpopular as idealism" and that "[t]o many, dualism is as discredited as vialism" pulls in the same direction, this time for dualism as a whole (Jackson and Braddon-Mitchell 1996, p. 4).

It would of course be silly to think that anything but an honest appraisal of the standing of the debate is being expressed here. Nevertheless, expressions such as these *do* make it appear that a Herculean argumentative effort is needed to raise the possibility that dualism is a live option for contemporary philosophers, and all the more so for interactionist dualism. To borrow a now common legalese expression, one gets the impression that dualism, to be considered, would have to be supported by arguments that establish the case 'beyond reasonable doubt', and further still in the case of interactionist dualism.

That, however, is an unreasonable standard, one we should not impose. Sticking to legalese we might say that evidence that proves the case *on the balance of probabilities* is all we can reasonably demand. We should expect a case for the claim that dualism in general — or interactionist dualism in particular — is *more likely* to be true than are its rival theories, and nothing more than that. That this is the only reasonable standard to impose is a truth that has been somewhat obscured by aspects of the current climate in philosophy of mind, but it is a truth, nevertheless.

Against the backdrop assumption that dualism is true, the considerations given above quickly show that the main rival theory for interactionism is epiphenomenalist dualism. One argument against that position has already

been offered. Another will discussed in the next chapter: the argument from the paradox of justified phenomenal judgement.

It is not here claimed that these two arguments constitute a knock down case against epiphenomenalism. However, if, as I have argued, a proof on the balance of probabilities is what we should be looking for, the two arguments against interactionism jointly go a long way toward making the case that the viability of interactionist dualism needs to be more carefully investigated. To make interactionism appear more likely than epiphenomenalism, it is not necessary to show that epiphenomenalism cannot be true. It may be sufficient to show that it shoulders very significant theoretical burdens. It has been argued that epiphenomenalism runs contrary to widely held and deeply entrenched intuitions about motivation, and that is a significant theoretical burden. Added to that is the slightly more technical argument in the next chapter to the effect that epiphenomenalism cannot satisfactorily account for the justification of beliefs and judgements *about* phenomenal experience. The hope is that together they will at least pique an interest in interactionism.

Piquing an interest in interactionism is necessary in the current climate, but it is not sufficient. If it were true, as is widely held, that interactionist dualism is flatly disproved by current scientific knowledge then the position would hardly merit a second look. The latter two chapters of this thesis argue that this is not the case. In chapter four it is argued that interactionist dualism is compatible with the conservation of energy law in physics and in chapter five interactionism is defended against some common, less convincing, arguments.

1.5 Proportionality

In the prevailing climate in recent analytic philosophy of mind, interactionism has been considered dead and gone; in need of 'resuscitation' by arguments beyond reasonable doubt. Investigation of which position is more likely on the

balance of probabilities is, in contrast, a context conducive to the close appraisal of some arguments that have thus far largely escaped detailed scrutiny. This thesis contends that the argument against interactionism from the causal closure of the physical, in particular, stands in need of much closer scrutiny than it has thus far been subjected to. If causal overdetermination is excluded, the *validity* of the argument is rarely doubted: if the physical is causally closed, then the mental is either itself physical, contrary to the dualist part of interactionist dualism, or it is causally inefficacious, contrary to the interactionist part. What is required is further investigation into the *truth-value* of the crucial premise: that the physical is causally closed.

Consider a simplistic analysis of the circumstances under which one might recommend that a thesis be accepted for the sake of argument. Intuitively, there are two conditions under which that can do little harm, namely when either (i) not much hinges on the acceptance of the thesis, or (ii) we are superbly confident that the thesis is true. If either of these conditions is satisfied to a very high degree it is reasonable to ask that a thesis be accepted for the sake of argument.

It is rarely the case that we know with near-certainty that a thesis we are being asked to accept is true, and it is, of course, also rare that we are asked to accept a thesis on whose acceptance nothing of importance hinges. That does not mean that we are almost never warranted in accepting theses. What we require is a *suitable mixture* of the two conditions, and we have that if the degree of confidence in the thesis is *proportionate* to how much hinges on its being true or false. Call this simplistic analysis *the proportionality constraint* on accepting a thesis.

This work contends that acceptance the thesis of causal closure of the physical *violates* the proportionality constraint, because the importance of what hinges on its acceptance is disproportionate to the reasons we have for accepting it. We should *not* yet accept that thesis of causal closure is true.

Above it was argued that monism wins a not inconsiderable proportion of its support through elimination of the available dualist alternatives. Then an objection to epiphenomenalism was presented, alleging that epiphenomenalism is wildly at odds with widespread and firmly entrenched intuitions about the causal efficacy of phenomenal experiences. Conjoining that argument with the argument in the next chapter yields a solid case for the claim that epiphenomenalism is an implausible position. If that is right — and if the brief considerations above against positions of type i. and ii. are effective — interactionism emerges as the *only* plausible dualist position. Since the thesis of causal closure leads to what is widely taken to be a decisive argument against interactionist dualism, accepting the thesis of causal closure will force a decision on the question of whether to accept dualism or monism, for once interactionism is excluded, no plausible dualist position is left. In this context it is safe to say that *a lot* hinges on the acceptance of the thesis of causal closure.

We might still respect the proportionality constraint if we have reason to be supremely confident that the thesis of causal closure is true. That, no doubt, is what a number of contemporary philosophers believe to be the case. However, as the last two chapters of this thesis show, the arguments in favour of the thesis of causal closure fail to adequately support belief in that thesis.

All of this is not to say, of course, that we should stop considering what *would* be the case if the thesis *were* true; working out the consequences of physicalism is as important as before. It would be well to remember, however, that the results are conditional. It would be well, too, to consider the consequences of the thesis being false. The arguments in favour of the thesis of causal closure fall well short of what would be required to justify allowing the acceptance of the thesis of causal closure to, as it were, 'become stable'. That is, to an extent, what has happened; recent analytic philosophy of mind in the English-speaking world has been dominated by the view that interactionism is dead and gone. This thesis aims to show that that is a conclusion which, by that

tradition's own standards, lacks adequate support. Analytic philosophy of mind would profit from accepting interactionism as what it is: a live option, a reasonable position.

Before seeing interactionism as such, there is no reason to demand a demonstration that the thesis of causal closure is *not at all* supported by the arguments. Once it is accepted that much hinges on the acceptance of the causal closure thesis it becomes clear that, to justify discounting interactionism to the degree which that has been done, what is required is not just *some* support for the causal closure thesis, but *very strong* support. This thesis attempts to show that that kind of support is not forthcoming.

The causal closure thesis may still play legitimate roles in philosophical arguments, for example in arguments for the truth of physicalism. It cannot, however, legitimately play such roles without arguments justifying belief in the thesis itself, or with only brief and superficial arguments. What is suggested here is not that the thesis of causal closure should not be accepted if a solid argument is put forth in its favour, but that it should not be accepted in the absence of such an argument.

Regardless of the merits of the simplistic analysis of when one should accept a thesis, it is certainly a striking fact that arguments for belief in the causal closure thesis are very rarely presented. Of the philosophers mentioned above — Jackson and Braddon-Mitchell, Chalmers, Lewis and Papineau — only the latter presents much of an argument for the causal closure thesis, and yet their arguments all explicitly rely on the thesis. Moreover, it is reasonable to think, with Papineau, that many materialist arguments that do not mention the thesis explicitly nevertheless depend implicitly on the truth of the thesis (2002, pp. 233-34). If the arguments in chapter four and five of this thesis are successful, this predicts difficulty for anyone wishing to exclude interactionism on the basis of an argument from the thesis of causal closure. It does not, of

course, close the issue. To close the issue is, however, not the aim of the current thesis. The aim of the current thesis is to open it up.

Chapter 2: The Paradox of Justified Phenomenal Judgement

2.1 The Paradox

The state of affairs a judgement *is about* usually figures in an explanation of what does, or does not, justify that judgement. Were I to make the judgement that the sky is overcast today, a part of the explanation of why this is a justified judgement would be played by the meteorological state of affairs in the region where I am. It seems obvious that this is not an accident, and that the state of affairs the judgement is about is not merely a useful pedagogical tool. Being *appropriately related* to a state of affairs that is the way the judgement claims that it is seems certain to be what *makes* the judgement justified. This does not, I claim, come under threat from any lack of knowledge about the precise nature of the appropriate relation, nor from lack of clarity about what it means for a judgement to claim that the world is a certain way.

Sometimes we make judgements about our phenomenal experience. One might say (or say to oneself, or write, or sing, or ...) that one's visual experience is dominated by a certain colour, or that one is having a particularly pleasant taste-experience at that time. Most of the judgements thus expressed are no doubt justified.[1] It will be argued below, however, that not only might you be *wrong* about your phenomenal experiences and thus make *erroneous* judgements about them but that you are furthermore capable of making *unjustified* judgements about your own occurrent phenomenal experience. This is by no means uncontroversial but the examples of unjustified phenomenal judgements I give below will hopefully go some way toward defending the claim. (Indeed, that we can be aware of properties of experiences as such is also sometimes disputed, but I do not defend that claim here.) Given unjustified phenomenal

[1] There is an implicit restriction to judgements about *intrinsic* properties of experiences. If that restriction is not imposed, there is no difficulty in supposing either that judgements about phenomenal experiences can be erroneous or that they can be unjustified.

judgements, an explanation of what distinguishes justified judgements from judgements that are not justified is required.

According to epiphenomenalist dualism, judgements are physical events.[2] Additionally they often cause other physical events downstream, for example when the judgement is expressed. Epiphenomenalism holds that phenomenal experience *forms no part* of the causal history of *any* physical event. One natural way of cashing out what it is to be 'appropriately related' to a state of affairs — of explaining in virtue of what most of our judgements about phenomenal experience are justified — is therefore unavailable to epiphenomenalists: they cannot give a *causal* theory of justification. On their account, there *is no* causal link between the experiences and the judgements about those experiences.

Whether epiphenomenalism is forced to accept that the phenomenal experience judgements are about has *nothing to do* with the judgements' status as justified is a further question. On the face of things, however, that seems to be a likely outcome, for once a causal link is excluded it is not clear what role the phenomenal experience a judgement is about *could* play in accounting for the justification of a judgement about it. Thus we have, at least *prima facie*, a paradoxical situation, in which the intentional object of a judgement is *unrelated* to the status of that judgement as justified or not. One might call that result *the paradox of justified phenomenal judgement*.[3]

---

[2] See e.g. Chalmers (1996, pp. 173-74).

[3] Chalmers gives vivid expression to a closely related paradox in *The Conscious Mind* (1996, chapter 5). The paradox Chalmers discusses in his book is couched in terms of *explanatory* irrelevance: given that judgements are taken to be reducible events, there should be a purely physical explanation of how they come to occur. But phenomenal experience, which is *not* reducible, will not figure in that explanation, so phenomenal experience will be *explanatorily irrelevant* to judgements about phenomenal experience (p. 177). Thus the discussion here is pitched in terms of a slightly different paradox than that which Chalmers discusses. This is innocent, however, for although Chalmers does not formulate the paradox of justified phenomenal judgements, he well could have, for he explicitly recognises the problem of how judgements about experience can be justified in the absence of a causal theory of justification (p. 193). I take his attempted solution to that problem as proposed solutions to what I have called the paradox of justified phenomenal judgement, and perpetrate no injustice that I can see in so doing. (See also Kneale (1959, p. 454), Shaffer (1965, p. 100 ff.), Elitzur (1989, pp. 8-9) and Penrose (1987, p. 116; 1989, pp. 408-09) for earlier formulations of closely related paradoxical results.)

One way of making the paradox vivid is by imagining a zombie twin world. There, everything is just like it is in our world, except that all the humanoid inhabitants of that world are phenomenal zombies: they lack phenomenal experience altogether. Any inhabitant of the zombie twin world is *ex hypothesi* just as likely as its twin on our world to make judgements about phenomenal experience. The difference is, of course, that in our world the vast majority of our phenomenal judgements are justified, but in the zombie world a justified phenomenal judgement is almost never made.[4] How do we account for the difference?

2.1.1 INTERACTIONIST RESPONSES

The paradox of justified phenomenal judgement does not arise for an *interactionist* dualist account of the mind. This is not surprising, for material in allowing the paradox to arise for the epiphenomenalist is the causal inefficacy of phenomenal experience. The interactionist, in contrast, can subscribe to a widely accepted and intuitively very plausible theory of justification: the *causal* theory.[5] According to the causal theory of justification it is a condition of a belief being justified that the object of the belief figures appropriately in the causal history of that belief. An interactionist can maintain that a causal theory of justification is appropriate in general, and in particular that it applies to judgements about phenomenal experience. Then our judgements about phenomenal experience are justified because they are at least partly *caused by* the experiences they are about, and the paradox fails to arise. In chapter one it

[4] *Almost* never, but perhaps sometimes, as eliminativists about consciousness would have counterparts in the zombie world, too. Their judgements would represent the limiting case. The judgement 'I have no phenomenal experience' might perhaps be said to be a degenerate case of a judgement 'about' phenomenal experience. Since it corresponds to the state of affairs it is 'about', perhaps an account of justification which made that judgement come out justified could be developed.

The zombie case, by the way, is an illustration here and the claim that justified judgements are almost never made in the zombie world is not required for the argument. All the argument requires is that there be some unjustified judgements in *our* world, a claim I defend below. I think that the claim about the zombie world is plausible, but I do not defend it here.

[5] For arguments in favour of the intuitive plausibility of this theory see section IV of Benacerraf's 'Mathematical Truth' (1973, pp. 671-73).

was argued that interactionism is better aligned with common sense because it takes phenomenal experience to be causally efficacious. Given the intuitive plausibility of the causal account of justification, the solution to the paradox outlined here constitutes a natural choice for the interactionist, a suitable continuation of the alignment with intuition and common sense.[6]

The paradox of justified phenomenal judgements is a very memorable problem for epiphenomenalism. Furthermore, the fact that interactionism has a solution to the paradox is made more noteworthy still by the fact that interactionism can solve other problems that face epiphenomenalism as well. Epiphenomenalism was criticised above for clashing radically with ubiquitous beliefs about motivation. Interactionism has no such problem. It is open to interactionism to claim that it is (at least partly) *because* you know that chocolate tastes good that you purchase a chocolate bar, that you exercise (at least partly) *because* you wish to *feel good* about your physical condition (and you believe that exercise will further this goal), that it is (at least partly) *because* being humiliated *feels bad* that most people avoid humiliating situations etc. Interactionism aligns well with common sense here.

Other problems faced by epiphenomenalism on account of the causal inefficacy of phenomenal experiences are avoided by interactionism in similar ways. The *evolutionary* objection complains that epiphenomenalism leaves us without an account of how phenomenal experience has evolved in humans. A causally inefficacious feature of our constitution cannot contribute to our reproductive fitness, so, *a fortiori*, cannot have been among the traits picked out by evolution. This poses a challenge for the dualist to explain how phenomenal experience came to evolve. The early Jackson's solution to this problem — he

[6] What is needed is *differential treatment* of justified and unjustified phenomenal judgements, and that is what it will be argued that epiphenomenalism fails to provide. It is assumed here that the causal theory is an excellent candidate for supplying such differential treatment, and that a systematic difference between the causal histories of justified and unjustified phenomenal judgements is discoverable; more specifically the difference of containing and not containing, in the appropriate place, the right phenomenal experience. This, of course, remains a somewhat substantial assumption until much more is known about this issue.

later characterises it as 'hardly attractive' — is that phenomenal experience is an unavoidable by-product of traits that *are* adaptive (Jackson 1982, pp. 133-34; Jackson and Braddon-Mitchell 1996, p. 7). Interactionism, in contrast, offers a rather more elegant solution; phenomenal experience was picked out by evolution because it *is* adaptive.

Interactionism will admittedly face an isomorphic problem when it comes to accounting for phenomenal experiences that are manifestly maladaptive. Panic angst is an obvious example. Panic angst is all-consuming and debilitating, and many of the situations where it sets in would be far better dealt with by action than by the inaction that it causes, so surely the capacity for panic angst is maladaptive if anything is. Unsurprisingly, the answer to this challenge is also isomorphic; the interactionist should claim that phenomenal experience comes as a 'package deal', and that it is, overall, more adaptive to have the potential for the whole spectrum of phenomenal experiences than to not have any. This may make it seem like the advantage interactionism can claim in favour of epiphenomenalism here is very modest. Modest or not, it is not insignificant. Resorting to the package deal solution for a *small subset* of phenomenal experience is much less problematic than it is to thus explain the *entirety* of a phenomenon as ubiquitous and all-encompassing as phenomenal experience is. The solution, in other words, is acceptable in very limited cases, but its use should be minimised. That is what interactionism does. Similarly, what Jackson and Braddon-Mitchell call 'the epistemological' objection to epiphenomenalism (p. 7), that of how phenomenal experience can be remembered, does not cause problems for interactionism, for causally efficacious phenomenal experiences *can* leave traces in the world and thus can be known and remembered.

These considerations reinforce the view argued above, that interactionism is a position that fits *very well* with our commonsense view of the world. This is not to say that its plausibility cannot be undermined by argument,

levelled both specifically at interactionism and generally at dualism. It does, however, put strain on any argument summoned to perform this duty, and, to reiterate, particularly if it is true, as was argued in the first chapter, that the plausibility of the dominant alternative position — monism — depends to a not inconsiderable extent on the elimination of tenable dualist alternatives.

2.1.2 CHALMERS' RESPONSE

We now turn to the solution Chalmers proposes to the paradox in *The Conscious Mind*. As mentioned, an obvious candidate solution is to adopt a causal theory of justification; this stops the paradox from arising in the first place. Chalmers rejects this solution. The problem with such an account is, he argues, that causal connections are contingent — "a causal connection that holds might not have held" — so any knowledge that relies on such a connection cannot be *certain* either (p. 195). This, he argues, conflicts with the *certainty* of our knowledge of phenomenal experience.

We take ourselves to know much about, for example, medium sized objects that surround us. Nevertheless, it is widely held that *for all we know* we could be brains in vats; all our sensory input could be the result of an elaborate scam and not of the objects we believe occupy our surroundings.[7] In contrast, according to a similarly widely shared intuition, it is *not* possible to construct a sceptical scenario *about being conscious* analogous to the sceptical brain-in-vat scenario about being embedded in a physical world. Sceptical scenarios rely on the possibility that everything might *seem just the same* to the subject while the situation is radically different from what the subject believes it to be. Therefore, given that "[t]here is no situation in which everything seems just the same to us

[7] Chalmers denies elsewhere (2003b) that brains–in–vats would be massively deluded, but the contrast to the sceptical scenario and its uncertainty is really only illustrational here. What matters is that Chalmers takes the causal account to be unable to deliver the required degree of *certainty*, a certainty he thinks that our knowledge that we are conscious demands: "There will always be a skeptical scenario in which everything seems just the same to the subject, but in which the causal connection is absent and in which [the purportedly known object] does not exist; so the subject cannot know for certain about [that object]. But we do know for certain that we are conscious; so a causal account of this knowledge is inappropriate" (1996, p. 195).

but in which we are not conscious, as our conscious experience is (at least partly) constitutive of the way things seem", Chalmers argues that a sceptical scenario about being conscious is impossible to construct (p. 195). This shows, he argues, that our knowledge about conscious experience is certain, and consequently that any account that allows for uncertainty about that knowledge (and the causal theory of justification is among them) is inadequate.

According to Chalmers in *The Conscious Mind*, judgements about phenomenal experience are justified simply by "*having* the experiences" (p. 196):

> [T]here is something intrinsically epistemic about experience. To have an experience is automatically to stand in some sort of intimate epistemic relation to the experience—a relation that we might call "acquaintance." There is not even a conceptual possibility that a subject could have a red experience like this one without having *any* epistemic contact with it: to have the experience is to be related to it in this way (pp. 196-97).

2.1.3 CRITICISM OF CHALMERS' RESPONSE

There is something right about Chalmers' picture here, but, I shall argue, something that is wrong about it as well. It is clearly true that it is impossible to *have* an experience without standing in *any* relation to it. The problem, however, is determining the nature of this relation with as much precision as possible. The claim that the relation is epistemic does not on its own take us any closer to an explanation or understanding of the relation, and neither does the assertion that experience itself has an 'intrinsically epistemic' aspect. Both statements are surely true on *some* reading of them. However, like the claim that simply *having* an experience justifies judgements about the experience, they do not elucidate what requires elucidation; they restate what needs to be explained in a different way.

Is having phenomenal experience a *necessary* condition for being able to make justified judgements about phenomenal experience? Can a judgement about phenomenal experience be justified if there is no such experience? It is

hard to think of a case where this would be so, but perhaps not impossible.[8] Intuitions may vary, however, and not much hinges on this point.

2.1.3.1 HAVING EXPERIENCE IS NOT SUFFICIENT

What is much more important is that just having phenomenal experience is *not a sufficient condition* for making justified judgements about phenomenal experience. We can imagine someone having one set of conscious experiences but making judgements as if they had a different set. It appears that Chalmers would have to agree, for this is just a description of the zombie case. However, contrary to Chalmers, I think the claim holds *not just* in the limiting case when the set of experiences that obtains is the null set, as is the case with phenomenal zombies. We can coherently imagine someone who has the right 'amount' of phenomenal experience, but who nevertheless makes a large number of mistakes in their judgements about phenomenal experience.

That this is possible follows from the possibility of making single erroneous judgements about phenomenal experiences, which Chalmers does not contest. If it is possible to make one erroneous judgement about phenomenal experience, it is also possible to make another mistake

[8] Imagine subjects repeatedly administered a drug with a dual effect: first to cause strong pleasant taste experiences, and secondly to interfere with the mechanism by which gustatory experiences are judged, causing it to 'go haywire'. In the first few trials the subjects might, while still under the influence of the drug, intermittently report very pleasant and very awful taste experiences. After the drug wore off they might be expected to report that, in fact, their taste experiences had been very pleasant throughout.

After repeated trials, and so long as no memory impairment is supposed, it is reasonable to expect the reports during the time under the drug's influence to become gradually more and more positive. What is going on, we might stipulate, is that the subjects remember that the drug has a dual effect, and overrides their initial inclination to judge that they are having dreadful taste experiences, because they know that they are influenced by the very same drug as before, and know that, in the previous instances, when the drug wore off, they confidently asserted that their taste experience had been pleasant all along.

If unbeknown to the patients the drug was now changed to one whose first effect was to *suppress* all gustatory experience while the second was the same as before, it seems likely that the subjects would still judge that they had pleasant gustatory experiences, though perhaps with decreased frequency and somewhat lessened confidence. In such a case I think it would be best to say that the judgements were justified. The causal theory can account for this. The distinctive feature of the situation is that the subjects learns to base their judgements about experiences immediately prior to the judgements on *previous* judgements, judgements that in turn are causally linked to experiences. The effect is that the subject's latest judgements become causally dependent on experiences much further removed from the judgements than what is normally the case. Nevertheless, there is a causal connection between experiences and judgements about those experiences that explains why we should judge them to be justified.

immediately following the first one, adding up to two consecutive mistaken judgements. Then it is also possible to have a long series of nothing but mistakes. The larger the number of consecutive mistakes the more improbable the scenario, but given a large enough sample to draw on even the very improbable will obtain. (Put enough monkeys to work on enough typewriters, and you will — later rather than sooner, probably — end up with the complete works of Shakespeare.) There may be an upper limit to the amount of mistakes we can coherently attribute to another person.[9] But those limits allow enough mistakes to make it reasonable to hold that some proportion of such a person's judgements would have to be classified as *unjustified*; lest the concept of justification lose all content.

Perhaps some would protest, however, and argue that to admit that we might make *mistaken* judgements about phenomenal experiences is not yet to admit that those judgements are unjustified, regardless of the number of mistakes we suppose are made.[10] How should we evaluate such a claim? We might start with the observation that while infallibility may or may not imply justification, the failure of infallibility certainly does not imply failure of justification.[11] The question of whether someone can fail to be justified in beliefs about their phenomenal experience is therefore quite correctly claimed to be a further, distinct question from that of whether one can be mistaken about one's phenomenal experience. Moreover, the two questions appear to be *independent* of one another, in the following sense: it seems possible to make mistaken but justified judgements about phenomenal experience (cf. n. 8 above), and it seems

[9] See e.g. 'On the Very Idea of a Conceptual Scheme' (Davidson 1974, especially p. 19).

[10] I take it that such a person would be defending about phenomenal experience the view that Alston calls 'self-warrant': "For any proposition, *S*, of type *R*, it is logically impossible that *P* should believe that *S* and not be justified in believing that *S*" (1971, p. 235).

[11] Alston claims that infallibility implies justification: "[O]ne could hardly have a stronger (epistemic) justification for holding a certain belief than the logical impossibility of the belief's being mistaken" (1971, p. 229). I think this is less than obvious, for, since they are necessarily true, true mathematical beliefs are logically guaranteed to be correct; nevertheless there certainly seems to be examples of people holding true mathematical beliefs without being justified in so doing; *viz*. the first person to ever entertain the belief that there is no largest prime number. However, I remain agnostic on this issue here.

possible to make unjustified but correct judgements about phenomenal experience (as when a judgement is passed in excessive haste, but happens to be correct; more on this in a moment). Given this, some might claim that all our judgements about phenomenal experience, erroneous or not, are justified. For the paradox to arise in the form discussed here (cf. n. 3 above) it must be possible to make *unjustified* phenomenal judgements, so the question about the possibility of making unjustified phenomenal judgements is relevant here.

2.1.3.2 SOME UNJUSTIFIED PHENOMENAL JUDGEMENTS

Are all phenomenal judgements justified? I think not.[12] Examples of erroneous phenomenal judgements where it seems fully possible for the subject to avoid the erroneous judgement — for instance because the mistake arises from haste or carelessness — seem amply available. Furthermore, it seems reasonable to say that when it was fully within a subject's capabilities to avoid making an erroneous judgement, then she was *unjustified* in passing that judgement. Chalmers mentions as an example of an erroneous phenomenal judgement someone misclassifying the phenomenal experience of cold for heat when passing a quick judgement in the context of expecting to be burnt (2003a, p. 241). Given that it is likely that at least some such judgements could have been avoided with relative ease the example works just as well as an example of *unjustified* judgement.

Here is another example, from an exhibition I once enjoyed at the technical museum in Oslo. In an experiment subjects place their head level with a fixed indicator at some length from the subjects, perhaps three or four metres away. Another indicator is arranged on a sliding rod so that the subjects can move it closer and further away. The task is to attempt to place the two indicators next to one another, first while keeping one eye closed, and then with

[12] I take the view I express here to be the negation of what Alston calls 'self-warrant' (cf. n. 10), this despite the fact that the thesis he discusses is formulated in terms of beliefs and the present discussion centres on judgements.

both eyes open, without lifting one's head to get a vantage point view from above.

It is clear that the point of this setup is to show that experience is less reliably correlated with reality when that experience is the result of a one-eyed input than when it is the result of a two-eyed one. It also seems clear that participants have to pass judgements about their phenomenal experiences to complete the task. Given the knowledge we all share that objects appear smaller when they are further away, one might aptly describe what the participants do by saying that they move the rod until it *appears* that the two markers are of equal size. That participants pass judgement on their phenomenal experiences is a conclusion further strengthened by the fact that, in this setup, the participants *know* all along that the markers *are* of equal size.

The results on this task are not only better when both eyes are used than when one eye is covered, they are *also* better when *careful attention* is paid to the task than when this is not the case. It seems implausible to claim that the phenomenal experience *changes* between a careless first attempt and a second, careful attempt, and no other new information is made available to the participants, either. The only factor that changes is the level of attention paid to the phenomenal experience the participant is having. When the task is performed quickly and carelessly it is evident that it could be performed *better* than it actually is performed, and assuming that this fact is not beyond the ken of the participants (i.e. assuming that the participants could know this), quickly passed judgements seem clearly unjustified.

Although not put forth as such, further examples of unjustified phenomenal judgements are available in the literature. In *The Problem of Knowledge* Ayer (1956/1958, p. 69) discusses two drawn lines of approximately equal length and argues that one can be unsure about which of the lines looks longer than the other. Ayer puts this forth as an example of an phenomenal judgement that can be *erroneous*; I think, as before, that a hastily reached

conclusion here can be unjustified. In 'Vagueness and Coherence', Burns (1986) discusses a series of colour samples, such that each is indistinguishable from the next one in the series, but such that a member of the series *is* distinguishable from a later sample, some number of further steps removed from it. Simplifying this example we can imagine three colour-experiences, A, B and C, such that A is indistinguishable from B and B is indistinguishable from C, but such that A is just distinguishable from C.[13] A person who judged A to be an *identical* colour-experience to B would do so unjustifiably, since the resources necessary to determine that A and B are *not* identical experiences would be readily available to her (although only on some reflection): A and B differ in the property of *being discernible* from C. It seems clear that a judgement about phenomenal experience is being made here, and also that it is not justified. (Of course, in the absence of C, the judgement that A and B are identical may well be justified, though erroneous; sometimes we need 'props' to make judgements, also about experiences.)

Some would resist the conclusion that phenomenal experience can be unjustified. However, when taken together these examples do seem to show that it is possible to make unjustified phenomenal judgements about phenomenal experience one is having. If that is right, it follows that just *having* phenomenal experiences *cannot* be enough to justify our judgements about them. The position in *The Conscious Mind* must be wrong in this respect. Something more is required. What is needed is relation of an appropriate kind between the phenomenal experience and the judgement. What might that relation be?

2.1.4 INTERACTIONISM AND CERTAINTY

We have seen that it is open to interactionism to claim that the judgement is justified just in case the phenomenal experience figures appropriately in the

[13] Whether or not the series would in practice have to be longer is unimportant here.

causal history of the judgement. This is a solid account of justification, and one that we have a reasonably good grip on. But can that account resist the charge of not delivering enough certainty for our phenomenal knowledge?[14] If the mechanism by which phenomenal experience is linked to our judgements is contingent, what accounts for the *certainty* with which we know *that* we are conscious?

An interactionist might be tempted to insist that while it is true that nearly all causal relations are contingent, the causal relation between phenomenal experience and judgements about those experiences is *necessary*. This approach feels more substantial than the mere claim that there is something intrinsically epistemic about experience; it fleshes out this intuition somewhat, and goes some way toward explaining the intuition. It also shares the spirit of Chalmers' insistence that there is not even conceptual room for the having of an experience without having *some* epistemic contact with it. It arguably goes further, for in denying that *this* causal relation might have failed to obtain it claims that the relation that allows us to make *justified* judgements about phenomenal experience obtains necessarily. That approach thus denies the possibility of unjustified phenomenal judgements.

Justifying the denial of contingency for one particular causal relation is sure to carry with it significant difficulties. Furthermore, there is, as we have seen, good reason to think that unjustified phenomenal judgements really do occur, so a necessary connection here would prove too much. Aside from these difficulties, however, the interactionist has additional polemical reasons to avoid the proposal. If the possibility of unjustified phenomenal judgements could be denied, the advantage of the interactionist solution to the paradox of justified phenomenal judgement compared with the epiphenomenalist solution

[14] Cf. *The Conscious Mind* (Chalmers 1996, p. 195).

in *The Conscious Mind* would melt away.[15] It is the possibility of making unjustified phenomenal judgements (while still having phenomenal experience) that makes untenable the claim that simply *having* phenomenal experiences justifies phenomenal judgements, and it is this, in turn, which underpins the insistence that a more 'substantial' relation between the experiences and the judgements must obtain in order for the latter to be justified.

There is a better approach to the problem of certainty. We must distinguish between judgements *that* we are conscious ('*judgements that*'), and judgements about the *particular character* of various phenomenal experiences ('*judgements about*').[16] For creatures with phenomenal experience, *judgements that* seem infallible; that accounts for the apparent impossibility of the sceptical scenario.[17] There does not seem to be a way to construct a scenario where I have the same conscious experience, but in which I could not be certain *that* I am conscious.

The interactionist who espouses a causal account of justification can explain the certainty of *judgements that* in the following way. Suppose that I make an erroneous *judgement about* a phenomenal experience I have. Such a judgement may or may not be unjustified, but since it is a *judgement about* a phenomenal experience, it is still, on the interactionist account, the case that the phenomenal experience figures somewhere in the causal history of that judgement. (The error might e.g. be the result of the phenomenal experience not figuring in the causal history in the appropriate way.) On the other hand, if a phenomenal experience instead causes me to make the *judgement that* I am conscious, it does not matter *how* the experience figures in the causal history of

[15] Of course, the advantage of the interactionist account in securing *explanatory* relevance would still obtain cf. n. 3.

[16] Or, if existence is taken to be one of the characters a phenomenal experience can have, then we must distinguish between judgements about *that* character (and possibly also the essential properties of the experience) and judgements about all the other characters a phenomenal experience can have.

[17] I am following Alston (1971) in using the term 'infallibility' (and its cognates) in the sense in which it is *not* interchangeable with 'incorrigibility', i.e. to designate the impossibility of erring, not the impossibility that someone else should be in a position to show that I had erred.

my judgement, so long as it figures there *somewhere*. Where *judgements that* are concerned, phenomenal experience figuring *anywhere* in the causal history of the judgement *just is* for it to figure appropriately therein. So the certainty with which we know *that* we are conscious is accounted for.

It might be thought that the interactionist faces problem cases, in which subjects *judge that* they are conscious, but where no phenomenal experience figures anywhere in the causal history of those judgements. Not so. First, it is plausible and certainly open to the interactionist to claim that this never occurs. Given that we have phenomenal experiences all the time,[18] and given the interactionist insistence that phenomenal experiences are causally efficacious, it is natural to think that the causal history of *judgements that* always contains phenomenal experience. That is, I think, the response the interactionist should give.

Secondly, if that response should be judged inadequate, another is available. The interactionist should then claim that if *judgements that* are sometimes made without it being the case that the judgement is (partly, somehow) caused by a phenomenal experience, those judgements are *not justified*. The account must then be amended slightly; it will not be the case that just *having* phenomenal experience sufficies to justify *judgements that*. To qualify as justified, each particular *judgement that* must be partly caused by phenomenal experience; a phenomenal experience must figure in the causal history of the judgement somewhere. If that condition sometimes fails, then even *judgements that* can turn out to be unjustified. Strict infallibility about *judgements that* must be jettisoned.

How serious would that admission be? Not very serious. *Judgements about* the particular character of an experience are *also* — albeit indirectly — *judgements that* one is conscious. There are very many such judgements; so there

[18] A largely uncontroversial, though not uncontested claim. Also one I do not argue for in this thesis.

will be very many judgements which count as justified *judgements that*, either directly or indirectly. Compared to this number, the frequency of *judgements that* which *in no way at all* are caused by phenomenal experience — where phenomenal experience figures *nowhere* in the causal history of the judgement — is certain to be vanishingly small.

Then interactionist can retain very–near–certainty for *judgements that*, while falling short of absolute certainty. Any particular *judgement that* would not be *absolutely* guaranteed to be justified. Nevertheless, we *approach* justified certainty for those judgements as closely as is indicated by the ratio of (i) *judgements that* (direct and indirect ones) such that phenomenal experiences *do* figure in the causal history of that judgement, to (ii) *judgements that* where *no* phenomenal experience figure in the causal history. So the admission is a very minor one indeed, and the lessening in the degree of certainty for *judgements that* is, I think, one with which the interactionist could very well live.

Judgements about the particular character of phenomenal experience are, as we have seen, not infallible. That brings out what is wrong with the suggestion that *having* phenomenal experiences suffices to make justified *judgements about* them; the account it gives of the justification of *the particular character* of a given phenomenal experience is wrong. Just having some phenomenal experience or other does not justify a judgement of having a pleasant gustatory experience right, nor a judgement of experiencing intense heat.

A sensible interactionist opts for a combined account. For *judgements that* — expressed by statements such as "Egad! I am conscious!" — *any* causally connected phenomenal experience contributes equally well to the justification. An absurdly wayward causal chain can do all sorts of damage to the justification of *judgements about*, but, I claim, no damage at all to *judgements*

that.[19] On the other hand, about *judgements about* the particular character of some phenomenal experience a sensible interactionist should say that a necessary condition for a judgement to be justified is that the phenomenal experience the judgement is about figures *appropriately* in the causal history of the judgement (where the account of what it is to figure appropriately is still to be given).

If the causal relation between phenomenal experiences and *judgements about* them were to be completely severed in some subjects, those subjects would indeed make systematically unjustified *judgements about* the particular character of their phenomenal experience. What is more, the subjects would *not even* be justified in *judging that* they were conscious. That result is, I claim, a strength of the causal account of justification, not a weakness.[20] Regarding *judgements about*, if the phenomenal experience the judgement is about figures nowhere in the causal history of the judgement, there is not even a *prima facie* case for the claim that the judgement is justified. Equally, if *no* phenomenal experience *at all* forms part of the causal history of a *judgement that*, there is no *prima facie* case for the claim that that judgement is justified either. *A fortiori*, there is no *prima facie* support for the claim that there is *certain knowledge* to that effect.

Where there is causality there is contingency. The causal relation that ensures that our judgements are overwhelmingly justified might not have obtained. One can be unfortunate in other ways than by being a zombie. In particular, one can be unfortunate by having the causal connection between one's phenomenal experience and one's judgements about phenomenal experience severed. None of this, however, constitutes a problem for the causal

[19] I am bracing for counterexamples!

[20] It is only the assumption that *judgements that* are always at least partially caused by phenomenal experience which opens up for the idea that just *having* phenomenal experience may be all that is required for judgements *that*. If *judgements that* sometimes occur without being even partially caused by phenomenal experiences, the causal connection between the phenomenal experiences and the *judgement that* would be severed. Then there could be no sustainable claim of the *judgement that* being justified.

theory. The causal theory can explain all the reasonable *explananda*. The paradox of phenomenal judgement, while truly a headache for epiphenomenalism, is thus well handled by interactionism.

2.2 A Revised Account

I have argued that Chalmers gives an unsatisfactory answer to the paradox of justified phenomenal judgement in *The Conscious Mind* (1996). My complaint has been that the claim that simply *having* phenomenal experience sufficies to justify judgements about phenomenal experience must be false, and that the account given merely restates what needs explaining in a different way, without elucidating it. I also claimed that interactionism has a significant advantage over epiphenomenalism in that the paradox of justified phenomenal judgement fails to arise on that theory when the intuitively plausible causal account of justification is accepted. It is open to an interactionist to say that it is a necessary condition for a phenomenal judgement to be justified that the phenomenal experience the judgement is about figures appropriately in the causal history of the judgement.

However, Chalmers sees the view expressed in the later article 'The Content and Epistemology of Phenomenal Belief' as a (partial) *replacement* for the chapter in *The Conscious Mind* that deals with the paradox (Chalmers 2003a, n. 16). It therefore becomes crucial to answer the following questions: Does the account in the paper offer an elucidating explanation of what justifies phenomenal judgement? In particular, does it offer an explanation that diffuses the advantage alleged above for interactionism over epiphenomenalism? To that end we must now take a careful look at some parts of that paper.

2.2.1 EXPOSITION

In 'The Content and Epistemology of Phenomenal Belief' Chalmers (2003a) discusses an important area of interaction between experiences and beliefs; first

person phenomenal beliefs. The discussion is premised on what he calls 'phenomenal realism', the view that phenomenal properties "type mental states by what it is like to have them" and, importantly, "are not conceptually reducible to physical or functional properties" (2003a, p. 221). According to Chalmers, most dualist views about phenomenal experience are phenomenal realist. In addition, some materialist views are also phenomenal realist views, e.g. views that deny both conceptual reduction and ontological dualism by claiming that the physical necessarily entails the phenomenal only *a posteriori* (2003a, p. 222).

The paper proceeds by first constructing a framework for the analysis of the content of phenomenal beliefs, and secondly by applying this tool to epistemological questions. Many of the concepts employed in the paper are new, and their specific meanings often play crucial roles in the analysis and argument, so they will be discussed in some detail here.

2.2.1.1 PHENOMENAL CONCEPTS

In the first instance a distinction is drawn between phenomenal colour properties on the one hand and external colour properties on the other. Both an experience and an object can have the property of redness, and both are, Chalmers argues, "respectable properties in their own right", distinct from one another (2003a, p. 223). Corresponding to phenomenal properties are phenomenal *concepts*, which refer to or 'pick out' the phenomenal properties. Phenomenal concepts are divided into three sub-categories: relational, demonstrative and pure.

Relational phenomenal concepts have their reference fixed in ways involving paradigmatic objects in the external world. A *community* relational phenomenal concept — designated by a subscript C (e.g. '*redc*') — has its reference fixed via a route that includes normal observers under normal

circumstances; it is the concept of "*the phenomenal quality typically caused in normal subjects within my community by paradigmatic red things"* (2003a, p. 224).[21]

In contrast, an *individual* relational phenomenal concept — designated by a subscript I (e.g. '*red*$_I$') — is the concept of "*the phenomenal quality typically caused in me by paradigmatic red things*" (2003a, p. 224). Chalmers argues that both concepts refer to intrinsic properties rigidly designated; the concept picks out the same property across possible worlds as it picks out in the actual world, even across worlds where experiences with those intrinsic properties are not brought about by red[22] objects; and further that the public language expression 'red', as applied to experiences, can be understood to express either the community concept or the individual concept (*red*$_C$ or *red*$_I$).

Demonstrative phenomenal concepts, designated $\underline{E}$, form the second category. Demonstrative phenomenal concepts are expressed by locutions such as 'this quality', they function indexically, and they designate "whatever quality the subject is currently ostending" (2003a, p. 225). The job of picking out in this way is done by the demonstrative concept's *character*. However, the concept also has *content*, "corresponding to the quality that is actually ostended" (p. 225).[23] As with relational phenomenal concepts, a demonstrative phenomenal concept also designates rigidly.

In addition to these three types of phenomenal concepts, which pick out their reference *relationally*, Chalmers argues that there is another "crucial"

---

[21] Chalmers uses italics to mark concepts and beliefs. I shall follow this convention in the rest of this chapter, but I will additionally underscore phenomenal concepts designated by single letters (e.g. $\underline{E}$, $\underline{R}$) since it is sometimes a little hard to see whether a single letter is italicised or not.

[22] The referent of the concept of *external* redness is of course also rigidified here; the claim is that both relational phenomenal concepts rigidly designate (possibly distinct) phenomenal experiences, also across possible worlds in which those phenomenal experiences are not brought about by objects that *we would call red*.

[23] The text here is ambiguous between two interpretations: first, that the content of a demonstrative concept *is* the ostended quality, and second, that the content of a demonstrative concept merely *corresponds* to the ostended quality. I confess that I cannot make sense of what the content of the demonstrative concept would be if it merely corresponded to the ostended quality, so I am left with the interpretation that the content of demonstrative phenomenal concept *is* the ostended quality. This interpretation clearly has some impact on the criticism I present in section 2.2.2.1 below, but that criticism does not, I think, stand or fall on this point.

phenomenal concept, which picks out its reference "directly, in terms of its intrinsic phenomenal nature" (2003a, p. 225). Referring to Jackson's famous thought-example from 'Epiphenomenal Qualia' (1982) and 'What Mary Didn't Know' (1986), he writes:

> [C]onsider the knowledge that Mary gains when she learns for the first time what it is like to see red. She learns that seeing red has such-and such quality. Mary learns ... that red things will typically cause experiences of such-and-such quality in her, and in other members of her community. She learns ... that the experience she is now having has such-and-such quality, and that the quality she is now ostending is such-and-such. Call Mary's 'such-and-such' concept here $\underline{R}$. (2003a, p. 225).

Chalmers argues that the concept $\underline{R}$, which he calls a *pure phenomenal concept*, can be shown to be distinct from the other concepts (in the instances where they also pick out phenomenal redness) by means of tests for cognitive significance of the beliefs that we can expect Mary to gain when leaving her room. The first such belief is the belief that $red_C = \underline{R}$, the belief that "the quality typically caused in her community by red things is such-and-such" (p. 225). Both that belief and the corresponding belief for the *individual* phenomenal concept are, he argues, cognitively significant.

Importantly, Chalmers also argues that the belief that the quality currently being ostended is such-and-such, or $\underline{E} = \underline{R}$, is also cognitively significant:

> Mary's belief $\underline{E} = \underline{R}$ is as cognitively significant as any other belief in which the object of a demonstrative is independently characterized: e.g. my belief *I am David Chalmers*, or my belief *that object is tall*. For Mary, $\underline{E} = \underline{R}$ is not *a priori*. No *a priori* reasoning can rule out the hypothesis that she is now ostending some other quality entirely .... Indeed, nothing known *a priori* entails that the phenomenal quality R is ever instantiated in the actual world (pp. 225-26).

The demonstrative concept is intended to capture the conception of the object of the demonstration "*as* the object of a demonstration"; whereas the pure phenomenal concept is claimed to characterise "the phenomenal quality *as* the phenomenal quality that it is" (p. 226). The thought or belief $\underline{E}$ = $\underline{R}$ is thus intimately dependent on the ability to conceptualise one and the same experience simultaneously in two different ways or modes; one *as demonstrated*, and the other *as quality*: "one conceives the object of a demonstration *as* the object of a demonstration ('this [quality], whatever it happens to be'), and at the same time attributes it substantive qualitative properties, conceived non-demonstratively" (p. 226).

Characterised in terms of the two-dimensional framework, we can bring out the difference between $\underline{E}$ and $\underline{R}$ on Chalmers' account in the following way: the secondary or subjunctive intension of all the four concepts under consideration are the same, since the four concepts are all taken to be rigid designators, and rigid designators pick out the same referent across all possible worlds considered as counterfactual. However, $\underline{E}$ and $\underline{R}$ differ in the primary or epistemic intension. $\underline{E}$ picks out whatever quality the subject at the centre of the world ostends.[24] Consequently, the primary intension of the demonstrative concept $\underline{E}$ varies with the variance of the quality ostended at the centre. In contrast, Chalmers argues that the primary or epistemic intension of the pure phenomenal concept $\underline{R}$ "picks out phenomenal redness in all worlds" (p. 228).

Chalmers argues separately for the claim that the content of phenomenal concepts, and so also of phenomenal beliefs, fails to supervene on the physical; the details of that argument are passed over here. He furthermore argues that there is a supervenience base on which the content of phenomenal concepts *does* supervene, and that is the conjunction of the physical and phenomenal

[24] Chalmers discusses a complication to this on p. 228, but the complication is irrelevant for the present purposes.

properties of the world (p. 234). On Chalmers' account, the content of a phenomenal concept is partly *constituted* by a phenomenal quality,

> "in that the content will mirror the quality (picking out instances of the quality in all epistemic possibilities) and in that ... [when] the underlying quality is varied while background properties are held constant, the content will co-vary to mirror the quality" (p. 235).

Chalmers calls such a concept a *direct phenomenal concept*, and says that the clearest example occurs "when a subject attends to the quality of an experience, and forms a concept wholly based on the attention to the quality, 'taking up' the quality into the concept" (p. 235).

Direct phenomenal concepts are pure phenomenal concepts, but they do not, on Chalmers' view, exhaust that class. In addition, *standing* phenomenal concepts are also pure phenomenal concepts. Standing phenomenal concepts have their origin in direct phenomenal concepts. A standing phenomenal concept is "a concept of [a] phenomenal [quality] as the quality it is, based on a lucid understanding of that quality, rather than on a mere relational or demonstrative identification", and is in that regard similar to direct phenomenal concepts (p. 239). However, where the direct phenomenal concepts have the limited life-span of the experience that constitutes part of their content, standing phenomenal concepts may persist in the absence of the experience.

2.2.1.2 PHENOMENAL BELIEFS AND ACQUAINTANCE

Corresponding to the formation of direct phenomenal concepts formed in the manner described above — concepts that are 'wholly based' on attention to the quality that forms the content of the concept — is a class of beliefs, the *direct phenomenal beliefs*. They arise "when a subject predicates the [direct phenomenal] concept of the very experience responsible for constituting its content" (p. 236). Beliefs of this type instantiate the belief-schema whose cognitive significance was discussed above, $\underline{E} = \underline{R}$ or 'this quality is $\underline{R}$' (p. 236). In these instances, the

quality being picked out by the demonstrative phenomenal quality is *the very same quality* that constitutes the content of the pure phenomenal concept.

For the present purposes, the important question is whether an account of justified phenomenal beliefs can be established with a basis in the framework Chalmers has built here. A noteworthy point in this regard is that on Chalmers' account, direct phenomenal beliefs are infallible:[25]

> A direct phenomenal concept by its nature picks out instances of an underlying demonstrated phenomenal quality, and a direct phenomenal belief identifies the referent of that concept with the very demonstrated quality (or predicates the concept of the very experience that instantiated the quality), so its truth is guaranteed (p. 242).

In *The Conscious Mind*, Chalmers argued that *having* phenomenal experiences is sufficient for being justified in judgements about phenomenal experience. This might make one think that infallibility is all that is required for justification of judgements about phenomenal experience. Not so, for a couple of reasons. First, direct phenomenal beliefs constitute a small (though not insignificant) subclass of phenomenal beliefs. Most phenomenal beliefs have either standing pure phenomenal concepts or relational phenomenal concepts as constituents, and these beliefs are not infallible (2003a, p. 242). Secondly, infallibility alone is not enough for justification, even for the subset of phenomenal beliefs that *are* infallible. Any necessary truth is infallible, but it is possible to judge a necessary truth to be true without being justified in so doing, for example if one judges a complex (roughly: non-transparent) mathematical or logical truth to be true on the basis that it is being considered on a Tuesday (2003a, p. 245).

[25] Chalmers uses the terms 'incorrigible' and 'infallible' (and their cognates) interchangeably (p. 241), but, as the following quote makes clear, he is referring to what Alston (and I) call infallibility.

So further work is needed to account for justification of phenomenal judgements. To this end, Chalmers discusses a feature of direct concepts: they are, on his view, only supported by phenomenal properties:

> This conclusion is apparently revealed by an examination of cases; but it would be preferable not to leave it as a brute conclusion. In particular, it is natural to suggest that the conclusion holds because we bear a special relation to the phenomenal properties instantiated in our experience: a relation that we do not bear to ... other instantiated properties ..., and a relation that is required in order to form a direct concept of a property.... This relation would seem to be a peculiarly intimate one, made possible by the fact that experiences lie at the heart of the mind rather than standing at a distance from it; and it seems to be a relation that carries the potential for conceptual and epistemic consequences. We might call this relation *acquaintance* (p. 248).

Acquaintance is, then, on Chalmers' picture the relation that allows formation of direct phenomenal concepts. Direct phenomenal concepts are pure phenomenal concepts, and on Chalmers' account pure phenomenal concepts have a special feature; they involve or allow a "*fully lucid knowledge* of the referents of the concepts in question" (p. 231, emphasis added). Acquaintance is thus a relation that allows for this fully lucid knowledge of concepts' referents to arise. No further explanation is given of what is meant by 'fully lucid' knowledge, but, as we shall see, it plays a significant role in his argument.

Chalmers argues that the particular certainty phenomenal beliefs have traditionally been held to enjoy, as well as the fact that, in sceptical scenarios, the phenomenal experiences of the deluded are nearly always held invariant between the two scenarios, confers independent plausibility on the view that "phenomenal properties and beliefs have a distinctive epistemic character" (pp. 248-49). He argues that the acquaintance relation best accounts for this. In particular, he wants to defend the following *Justification Thesis*:

> When a subject forms a direct phenomenal belief based on a phenomenal quality, then that belief is *prima facie* justified by virtue of the subject's acquaintance with that quality (p. 249).

Again, the traditional appeal to acquaintance by philosophers attempting to account for the particular certainty with which we seem to know phenomenal beliefs is offered as a further justification for the thesis. Chalmers argues, however, that his view is to be preferred because it is "more constrained", and because the acquaintance relation, on which the view relies, has been justified independently (2003a, p. 250).

To sum up, the structure of the argument is as follows: first, a conceptual framework is constructed. A special role is played in the framework by *pure phenomenal concepts*, which fix or pick out their referents "directly, in terms of [their] intrinsic phenomenal nature" (2003a, p. 225). Secondly, an infallibility thesis is put forth, the truth of which is guaranteed by meaning of the concepts as defined. Thirdly, it is argued that the reason that the infallibility thesis is limited to beliefs about phenomenal properties only is that one type of the special concepts needed in the infallibility thesis — the direct phenomenal concepts — is only supported by *phenomenal* properties (not by any other properties). Fourthly, it is argued that the class of concepts to which direct phenomenal concepts belong — the pure phenomenal concepts — have a special feature; they allow for '*fully lucid*' knowledge of the *referents* of the concepts. Fifthly, it is argued that the stipulated relation of *acquaintance* between a subject and phenomenal qualities or phenomenal experiences can *unify* our account of why only phenomenal properties allow for direct concepts to be formed and why these concepts imply 'fully lucid' knowledge of their referents. In the first case the explanation is that we are acquainted only with phenomenal properties, in the second it is that only acquaintance allows for this sort of knowledge, so the unifying element is the relation of acquaintance. (Alternatively, we can explain the second fact in terms of acquaintance, and the

first in terms of relying on the second; this subtle difference will not be important.)

The upshot of all this is that the subclass of phenomenal beliefs to which the infallibility thesis applies — the *direct* phenomenal beliefs — are thought to have acquired not only infallibility but *justification*. The (defeasible) justification obtains because the subject stands in a relation to the quality she or he forms a belief about that allows that quality to be *lucidly understood* by the subject.

2.2.2 CRITICISM

At this point, one might raise suspicions about the position in a number of places.

2.2.2.1 A TENUOUS DISTINCTION

Firstly, as we have seen, the relation of acquaintance is argued for by an argument to the best explanation, and the limited scope of the infallibility thesis is one of the purported *explananda*. However, the substantiveness of the infallibility thesis itself relies on the distinction between the demonstrative phenomenal concept $\underline{E}$ and the direct phenomenal concept $\underline{R}$. There are several reasons why that distinction may be thought to be tenuous.

Chalmers takes demonstrative concepts to "have a reference-fixing 'character' that leaves their referent open" and notes that this description fits $\underline{E}$ (p. 227). He then goes on to argue that "$\underline{R}$, on the other hand, is a substantive concept that is tied *a priori* to a specific sort of quality" (p. 227). That, however, seems odd. If our experiences had been radically different, then so, too, would any concept depending on those experiences be, and pure phenomenal concepts depend on experiences. So it seems that phenomenal concepts cannot have determinate content *a priori*, and $\underline{R}$ cannot be as described.

Chalmers attempts to strengthen the case for the distinctness of the two concepts and for the cognitive significance of beliefs such as $\underline{E} = \underline{R}$ *via* a discussion of the primary or epistemic intensions picked out by the two

concepts. Assuming that a single quality is being ostended, the primary intension of E picks out the quality that is ostended by the person at the centre of that world. The primary intension of E will thus vary across centred possible worlds (considered as actual), across different locations of the centre, and according to which quality is being ostended by the being at the centre. Chalmers argues that the epistemic intension of R, in contrast, will pick out phenomenal redness in *all* worlds, so this concept would have the same entry in *every* cell of the two-dimensional matrix.

If that is so, then there must be a severe restriction on who could possibly have the concept R, and about whom we could confidently assert that they had the concept R. Consider Mary, still imprisoned. Regardless of her theoretical proficiency, she could not have the concept R. To have confidence that a concept picks out phenomenal redness across worlds we must stipulate that the concept is the concept of a person who is then ostending phenomenal redness. So, it is a precondition of having the concept R that the concept E has the content it is stipulated, in the situation, to have. It is not, strictly, a precondition of having the demonstrative concept E that the concept R be had, on Chalmers' account, for the identity conditions of the demonstrative concept is not supposed to rely on its contents. But for E to have content at all, a quality must be ostended, and for R to have the epistemic intension it is argued that it has, *phenomenal redness* must be the ostended quality. One might think that the concept R need not be formed, for perhaps the phenomenal redness is not being attended to, perhaps the subject merely is in a position that *allows* it to form R, without *actually* forming R. Plausibly, however, the amount of attention necessary for ostending suffices for the formation of R, too. So it is effectively, though not strictly, a precondition for E to obtain that R obtains, too.

The demonstrative phenomenal concept E is said to fix its reference *relationally*, "with the referent characterized in terms of ... acts of ostension" (p. 225). A crucial question here is how robust the sense is in which this really is a

relational reference fixing. Conversely we can ask of the pure phenomenal concept, how robust the sense is in which that concept fixes its referent *directly*. These two questions point to a third, namely of whether there is a robust difference between the two concepts $\underline{E}$ and $\underline{R}$.

Is it possible, when ostending phenomenal properties, to conceive of the property *as* 'an object of demonstration'? It seems more likely that the *character* of the concept would be, as it were, 'overwhelmed' by its content, to such an extent that the conception of the quality as 'the quality I now ostend, whatever it happens to be' would give way for the quality itself. (One might loosely say that $\underline{E}$ would 'collapse' into $\underline{R}$.) Conversely, it is not obvious that there really are pure phenomenal concepts, that in each case "characteri[ze] the phenomenal quality *as* the phenomenal quality that it is", without also containing some demonstrative element (p. 226).

It is useful to consider Chalmers' 'Inverted Mary' thought-experiment, in which Inverted Mary is just like Mary, except that her colour vision is red/green inverted. As Chalmers points out, where Mary acquires the beliefs $red_C = \underline{R}$, $red_I = \underline{R}$, Inverted Mary acquires the beliefs $red_C = \underline{G}$ and $red_I = \underline{G}$. Further, Chalmers argues that Inverted Mary acquires the belief $\underline{E} = \underline{G}$ where Mary acquires the belief $\underline{E} = \underline{R}$. The first claims are sufficient to establish that Mary and Inverted Mary acquire belief with different content. But that is insufficient to say anything about the cognitive significance of the two last identities, for in these it is not only what is to the *right* of the identity sign that is different; the content of the demonstrative phenomenal concept on the *left* side is, of course, different too.

All of this casts doubt on the distinctness of the two concepts $\underline{E}$ and $\underline{R}$, and suggests instead that there may be a single concept that incorporates elements from each. It is of course possible that these problems could be explained away, but until they are, some doubt is cast on the cognitive significance of the identity. Once there is doubt that the identity is cognitively

significant, that doubt is passed on to the question of whether the infallibility thesis is substantive, and finally onto the argument for acquaintance as the best explanation for the limited application of the infallibility thesis.

2.2.2.2 OTHER DIFFICULTIES

There are other difficulties. The most obvious one pertains to the concept of 'lucid understanding'. As Chalmers points out, "[m]any have held that phenomenal properties can ... be known with a distinctive sort of justification" (p. 248). It is not clear, however, that the claim that acquaintance allows for lucid understanding really *advances* our understanding beyond that traditional thought.

Chalmers argues that "whenever a subject has a phenomenal property, the subject is acquainted with [it]" (p. 250), so his suggestion does not limit the *range* of circumstances in which 'special justification' obtains. Any improvement of our knowledge must, is seems, have its origin in the concept of lucid understanding. But the understanding is lucid only in name. What lucid understanding *is*, how it is afforded by acquaintance, and how the possession of lucid understanding confers justification on belief, is still rather obscure.[26] As Chalmers quite rightly points out, it is one thing to argue that there is *some* special justification for phenomenal beliefs (in fact, most people do not even seem to require much argument to be convinced of this), and quite another to

[26] The concept of 'lucid understanding' is — especially in the context of visual phenomenal experience — reminiscent of the concept of 'revelation'. Russell, e.g. writes: "[S]o far as concerns knowledge of the colour itself ... I know the colour perfectly and completely when I see it, and no further knowledge of it itself is even theoretically possible" (1912/1967, p. 25), and Strawson argues that "there is ... a fundamental sense in which colour words are *words for properties which are of such a kind that their whole and essential nature as properties can be and is fully revealed in sensory, phenomenal-quality experience, given only the qualitative character that that sensory experience has*" (1989, p. 224, emphasis original). If lucid understanding of phenomenal experiences implies knowing "exactly what they are" and gaining knowledge that "reveals the essence" of a phenomenal experience, then the thesis that lucid understanding is afforded by experience seems to coincide with what Lewis calls the "Identification Thesis" (1995, pp. 141, 42). (He rejects that thesis there and elsewhere (1997, section VII).) It does not seem to me the question of how *revelation* confers justification on belief has any less bite, however, so I think the difficulty remains, despite these similarities. (See also discussion in 'How to Speak of the Colors' (Johnston 1992, pp. 223-25).)

explain "what this justification consists in" (p. 249). The discussion in Chalmers' paper contributes toward that goal, but it is clear that more needs to be done.

This becomes even clearer once the claim of an independent case for the relation of acquaintance is questioned. Acquaintance is introduced as the relation "that makes possible the formation of direct phenomenal concepts" (p. 248). Its postulation is meant to explain why direct concepts fail to be sustained by properties other than phenomenal ones, by reference to the fact that we are not appropriately acquainted with the properties in question. That feature of the relation of acquaintance, however, must be seen together with what else is said about the relation. Acquaintance is said to obtain *whenever* a phenomenal property is had by a subject, and it is said to allow for a *lucid understanding* of the referents of the phenomenal concepts employed by a subject — i.e. of the phenomenal experiences of that subject — in such a way as to *confer justification* on the subject's beliefs about those experiences. Acquaintance, in other words, appears to amount to little more than a placeholder for whatever gives us a justification for beliefs about phenomenal properties which we lack for other properties. That we have such justification is widely recognised. The relation of acquaintance, as presented in this paper, does not, I suggest, enlighten us further.

If this is right, there is reason to doubt that the account offered in the article under discussion can do much to offset the advantage interactionism holds over epiphenomenalism in the area of phenomenal belief generally, and in relation to the paradox of justified phenomenal judgement specifically. Even if the justification thesis should be granted, further questions attach to the extension of the thesis beyond direct phenomenal beliefs. After all, most of our phenomenal beliefs are *not* direct, and the paradox of justified phenomenal judgements arises in full force also for beliefs involving what in the article is called *standing* phenomenal concepts.

Having noted that direct justification of beliefs involving standing phenomenal concepts would require a different account than that developed for direct phenomenal concepts (since the latter account relies on constitution), Chalmers goes on to say that:

> [i]ndirect justification for such beliefs can be secured by virtue of the plausible claim that any belief of the form $\underline{S}$ = $\underline{R}$ is (*prima facie*) justifiable, where $\underline{S}$ and $\underline{R}$ are standing and direct phenomenal concepts with the same epistemic content. ... Such beliefs are plausibly justifiable *a priori* .... If so, then beliefs involving standing phenomenal concepts can inherit justification by *a priori* inference from direct phenomenal beliefs, which will be justified in virtue of the Justification Thesis (pp. 253-54).[27]

Setting aside the worries raised previously, another problem arises that is specifically related to the 'inheritance' of justification. It is true that, if a subject is in a position to know that $\underline{S}$ and $\underline{R}$ have the same epistemic content, beliefs of the form $\underline{S}$ = $\underline{R}$ would gain *prima facie* justification on the basis of an inference that could be carried out *a priori*. It is not plausible, however, that a subject will be able to know the content of *either* concept *a priori*, and, it is not plausible, *a fortiori*, that the justification could actually be so inherited.

A final comparative point is worth making about the epiphenomenalist and the interactionist accounts of justification of phenomenal beliefs and their responses to the paradox of justified phenomenal judgement. Chalmers considers two arguments against phenomenal realism reconstructed from Shoemaker's 'Functionalism and Qualia' (1975).

The first reconstructed argument runs as follows:

(1) If phenomenal realism is true, experiences are causally irrelevant to phenomenal beliefs.

[27] Chalmers does not italicise the two last occurrences of the standing and direct phenomenal concepts $\underline{S}$ and $\underline{R}$ in this passage. I take that to be a mistake, since the letters here clearly refer to *concepts*, rather than to the referents of those concepts, cf. his n. 2, p. 223.

(2) If experiences are causally irrelevant to phenomenal beliefs, phenomenal beliefs are not knowledge.
(3) If phenomenal realism is true, phenomenal beliefs are not knowledge (Chalmers 2003a, p. 255).

Here the epiphenomenalist following Chalmers would deny (2). The interactionist would, of course, deny (1), so here there is a clear difference between the positions. (In fact, (1) is only plausible on the assumption that interactionism is not a viable option; for interactionism clearly qualifies as a species of phenomenal realism.)

The second reconstructed argument runs as follows:

(1) If phenomenal realism is true, then every conscious being has a possible zombie twin.
(2) If zombies are possible, they have the same phenomenal beliefs as their conscious twins, formed by the same mechanism.
(3) If zombies are possible, their phenomenal beliefs are false and unjustified.
(4) If it is possible that there are beings with the same phenomenal beliefs as a conscious being, formed by the same mechanism, where those phenomenal beliefs are false and unjustified, then the conscious being's phenomenal beliefs are unjustified.
(5) If phenomenal realism is true, every conscious being's phenomenal beliefs are unjustified (Chalmers 2003a, p. 256).

Interestingly, in response to this argument, Chalmers says that "the most obvious reply is to dispute premiss 2. There is no reason to accept that zombies have the same phenomenal beliefs as their conscious twins, and every reason to believe that they do not" (p. 257).

That regular subjects and their zombie counterparts would have different beliefs is, of course, something an interactionist would readily accept.[28] It is,

[28] The status of premise 1 on the interactionist account is not straightforward; the most natural response for an interactionist seems to be to deny the possibility of zombie twins. If zombies do not share the phenomenal experiences and (thus the) beliefs of their conscious 'twins', it seems that the interactionist should claim that they will be disposed to act differently in the same external circumstances. But then, of course, they would no longer be twins. This conclusion seems unavoidable where the laws of nature are

however, a somewhat startling admission from an epiphenomenalist. It is already difficult to believe that the phenomenal experience of a subject could vary without a concomitant variation in behaviour; that epiphenomenalism accepts this puts, I have argued, the position at considerable disadvantage to interactionism. (The rescue-plank for epiphenomenalism has been to claim that the causal efficacy we *think* that phenomenal experiences have is an illusion brought about by common causal ancestry.) It is harder still to believe that a difference in *belief* would fail to lead to a variation in behaviour.

Suppose that Mary, before being let out of her room for the first time, is unknowingly administered a drug that causes red/green phenomenal inversion. She is then taken back into her room, where the drug wears off. Before the second release it seems obvious that she would now be disposed to *behave differently* than she would have, had she not been administered the drug before her first release. The difference in anticipated behaviour is straightforwardly traceable to differences in phenomenal belief, and to differences in phenomenal experience. An interactionist that favours a causal account of belief and justification can account for this intuition; the causal history of Mary and her counterfactual counterpart will be distinct. But the epiphenomenalist that follows Chalmers cannot account for the intuition, and is presumably forced to deny it.

Now, it is of course true that everyone's a critic, and that it is harder to build a theory up than to attempt to tear it down. It is also true that causation — which interactionism relies on — is still ill understood; perhaps some would argue that the epiphenomenalist picture discussed here is no worse off for relying on the relation of acquaintance. Nevertheless, a causal theory of justification has a very high degree of intuitive appeal and is widely held. The application of this theory to the case of phenomenal belief is therefore not only

---

the same as in our world, so it seems that an interactionist should at least say that zombie twins are not nomologically possible.

best in tune with commonsense intuitions, it also has the significant advantage of merely extending the scope of widely accepted theory, instead of having to construct a new theory to account for the justification of phenomenal beliefs, so to speak, 'from scratch'. So, I submit, the advantage that interactionism was argued to hold over epiphenomenalism is untouched by the considerations offered in this paper.

Chapter 3: Two Objections

3.1 Introduction

In this chapter I consider two objections that are sometimes brought to bear against interactionism. Unlike the objections to interactionism considered in the rest of the thesis — which take results from science or trends in scientific history as their starting points — the objections considered in this chapter are of a more conceptual or 'analytic' nature, and depend less on extra-theoretical features.

The first objection is that interactionism suffers from a conceptual problem that bars it from yielding causally efficacious phenomenal properties. In section two I discuss a statement of this problem and offer a very simple counter-argument. I then offer some further speculations that might serve as starting points for investigation, should the simple argument not prove persuasive.

The second objection is that interactionism has a *definitional* problem. This objection is related to recent discussion on the difficulty of non-vacuously defining physicalism. Interactionist *dualism*, it might be thought, depends on the contrast with physicalism to have definite content, so if there is a difficulty in defining physicalism, there will be a parallel difficulty for interactionism. In section three I discuss the definitional problem, and explain why I do not think it poses a serious challenge to interactionism.

3.2 A Conceptual Problem

In his book, Chalmers briefly considers interactionism as a response to the woes of the epiphenomenalist, but claims that interactionism "raises more problems than it solves" (1996, p. 156).[1] He offers three considerations in support of this

[1] Chalmers has since changed his opinion on interactionism considerably. He now thinks that interactionism, epiphenomenalism and 'Type F monism' — a view according to which consciousness "is constituted by the intrinsic properties of fundamental physical entities" (2002, p. 265) — are the three best candidate views on the mind-body problem; see his 'Consciousness and its Place in Nature'

claim: an optimistic meta-induction on the history of science, a criticism of using quantum mechanics as a method for 'carving out room' for the mind to interact with the body, and lastly a conceptual challenge against interactionism. The first two are common but unconvincing criticisms of interactionism that I consider and reject in chapter five. The last, however, will be considered briefly here.

3.2.1 THE ARGUMENT

The claim is that interactionism has a conceptual problem such that it must fail to ensure that phenomenal experience has causal relevance. If the claim were defensible it would constitute a serious difficulty for interactionism, for the chief advantage interactionism enjoys (over epiphenomenalism) is precisely that it accords better with our common sense intuitions *about the causal relevance* of experience. Here it is:

> We can always subtract the phenomenal component from any explanatory account, yielding a purely causal component. Imagine (with Eccles) that "psychons" in the nonphysical mind push around physical processes in the brain, and that psychons are the seat of experience. We can tell a story about the causal relations between psychons and physical processes, and a story about the causal dynamics among psychons, without ever invoking the fact that psychons have phenomenal properties. Just as with physical processes, we can imagine subtracting the *phenomenal* properties of psychons, yielding a situation in which the causal dynamics are isomorphic. It follows that the fact that psychons are the seat of experience plays no essential role in a causal explanation, and that even in this picture experience is explanatorily irrelevant (pp. 157-58). [2]

(Chalmers 2002). He has, moreover, indicated (in personal communication) that he now considers the conceptual problem less serious than he did in his book, and that he sees it first and foremost as bringing out "a respect in which [T]ype-D [dualism, i.e. interactionism] and [T]ype-F [monism] are really continuous with one another" in that "both have phenomenal properties serving as the categorical basis for certain fundamental causal roles in the causal network" (2006).

[2] The reference is to (Eccles 1986). There may be two distinct issues here: (i) a competing position (what Chalmers calls 'type F-monism') appears to have some of the theoretical advantages that interactionism proclaims for itself without carrying what is widely taken to be its greatest burden: the denial of causal closure, and (ii) interactionism may appear unable to actually secure what it takes to be its greatest theoretical benefit: casual relevance for phenomenal experience. A note elsewhere (n. 26 in Chalmers

3.2.2 A COUNTER-ARGUMENT

Remarkably, this line is at first not unconvincing; it actually manages to make it sound plausible that we should always be able to subtract the phenomenal properties *even of psychons* in a causal account, leaving it otherwise intact.[3] If that were true, then the claim that phenomenal experience is 'explanatorily irrelevant' would come out very nearly tautological.

That fact gives a very good clue to what is wrong with the argument. Whether or not phenomenal experience is causally efficacious is *not* a question we can settle *a priori.* The strength of our intuition that phenomenal experience *is* causally efficacious gives us reason to treat empirical arguments to the contrary with caution, but the intuition does not decide the matter once and for all. If a strong account emerges that accounts for our actions without reference to phenomenal experience we should be forced to acknowledge that the intuitions were wrong all along.

By parity of reasoning, the question should not be foreclosed the *other* way either. Should a situation emerge where we have empirical reasons to believe in the causal efficacy of experience but difficulty reconciling the empirical data with features of our conceptual scheme, it is of course not given that the *data* must be wrong. On the contrary, the data could lead us to discover problems or inaccuracies in our conceptual scheme, in our 'ways of thinking'. The question of whether phenomenal experience is causally efficacious is an empirical one.

Then the next step is straightforward. If the causal efficacy of phenomenal experience is an empirical question, the *explanatory* relevance of

---

2002) suggests reading the section as concerning (i) and not (ii), contrary to my interpretation. I think that reading is less well supported by the text in the book, but some humility is obviously appropriate, since the interpreter and the interpreted coincide. It is therefore acknowledged that it is (ii) and not (i) that is the focus of criticism in this section, and that it may be that it was (i) and not (ii) that the author was originally concerned with. In any case; (ii) is succinctly formulated in the passage and interesting in its own right; it will be the sole focus here.

[3] I do not here discuss the 'subtle' forms of causal efficacy that Chalmers discusses in his book (1996, pp. 153-56). Whatever the merit of those proposals, they do not reflect the causal efficacy that Eccles has in mind. The causal efficacy of the psychons is of the more straightforward 'pushing around' type.

phenomenal experience in our accounts of the world — accounts that are, of course, largely causal — *cannot* be a matter to be settled *a priori*. If phenomenal experience turns out to be causally relevant it just follows that any account where phenomenal experience does not figure causally is *not the right account*. An account that fails to mention causally efficacious features minimally departs from the correct account by either containing 'blank spaces' where the excluded features should have been or by positing the wrong entities at some nodes. Either way, it is the wrong account. (It could of course depart from the correct account in more dramatic ways.)

The mistake of positing at some nodes the wrong entities need not be more conspicuous or spectacular than positing, at those nodes, entities such that their causal properties are primitive facts about them. What Chalmers' argument draws on — and a point about which it is, of course, entirely correct — is that an account which posits entities whose causal properties are primitive facts about them in place of our supposed psychons, can be isomorphic to the account with psychons in a great number of respects. *But not in all respects!* If the interactionist is right, the causal properties of psychons are *not* primitive facts about them; they depend on intrinsic, phenomenal properties. If interactionism is true, any account that fails to mention the phenomenal properties will not be the correct account. Since explanation requires giving correct accounts, phenomenal experience is explanatorily relevant if it is causally relevant.

3.2.3 FURTHER CONSIDERATIONS

This simple argument seems to me decisive, but I fear that others may not share my view. For those not yet convinced, are there alternative arguments to be found? Even if the simple argument fails, it seems to me that another must be waiting to be found, for whether phenomenal experience is causally efficacious

seems certain to be an empirical question, and from that explanatory relevance appears to follow.

There are interesting, difficult and subtle issues in this vicinity. The main focus of this thesis is to consider challenges to interactionism stemming from scientific results (or trends in scientific history), so a thorough treatment of conceptual problems cannot be undertaken. In place of a fully rigorous argument — to replace the simple argument for those which that argument leaves unconvinced — I will offer some speculations that I hope might serve as starting points for further investigation. If the argument in the other chapters is successful, the scientific results that are available to us today make for poor starting points for arguments against interactionism. It may be, therefore, that conceptual challenges constitute the most serious of the current objections to interactionism, and they may thus be well worthy of attention in future research.

3.2.3.1 ATTEMPTED REASSERTION OF EFFICACY

When searching for another way to counter the conceptual challenge, a simple-minded thought that springs to mind is that the description Chalmers gives may not be accurate. Perhaps psychons are not the *seats* of experience. Perhaps psychons *are* phenomenal experience; it could be that phenomenal experience *itself* pushes around physical processes in the brain. If that is so, we cannot, of course, subtract the phenomenal properties from our explanation and still keep isomorphic causal dynamics. We would be subtracting the very entities that are causally efficacious! That is precisely what an interactionist thinks you *would* do if you were to tell a causal story without giving mention to phenomenal experience, so thus far the picture fits well with the interactionist account.

The problem is that the properties which the interactionist believes are causally efficacious appear to be prime candidates for being *intrinsic* properties of experience. If an intrinsic property is a property such that it is possible both for lonely and accompanied things to either have or lack the property, as Lewis

and Langton argue (1998), then properties such as 'being green' and 'being blue' are intrinsic, for an experience can either be, or fail to be, green or blue, both when it is accompanied and when it is lonely.[4] (We all know of accompanied green and blue experiences — we have had them — and there is arguably a possible world (an odd one!) where all that exists is a green experience, and likewise one where all that exists is a blue one.)

But if phenomenal properties are intrinsic properties of psychons it seems that there must be other properties — *causal* properties — that ensure that psychons behave as they do in causal interactions with other parts of the world. The causal properties of an experience are, one might say, to psychons like cogs are to a gear; without them nothing moves, *they* are the efficacious properties, they transmit motion.

We can imagine varying one property or category of properties without feeling forced to imagine varying another. The conceptual challenge asks us to imagine an experience with the very same *causal* properties but with different *intrinsic* properties. Since we seem to succeed in imagining this (just like we can imagine the same cogs on different naves) the claim arises again that it is the causal properties that matter for our explanation, not the intrinsic ones; we can *abstract away* from phenomenal properties without loss of explanatory power.

3.2.3.2 A RESPONSE

The claim that we can abstract away from phenomenal properties without loss of explanatory power obviously rests on the assumption that causal powers *can* remain the same through a change in intrinsic properties. That seems hard to deny, if the relevant modality is logical possibility. There is, surely, a possible world in which there are entities that have causal properties just like those of psychons but completely different intrinsic properties, and there may even be

[4] Lewis and Langton offer this definition with qualifications, but the qualifications can safely be ignored here.

one with entities in it that have causal properties just like those of psychons, but no intrinsic properties at all.

The interactionist thesis, however, does not purport to hold with logical necessity; the interactionist thesis is a contingent claim. It is the claim that in the actual world *and* in nearby worlds, experiences with different intrinsic properties do *not* have the same causal properties. According to the interactionist, counterfactual claims like "had the experience been blue instead of green, that person would have acted differently" are *true*. But if such claims are true, phenomenal experience is not explanatorily irrelevant.

Since counterfactual dependence relies on *nearby* worlds, there is no threat to counterfactual dependence from the existence of a *remote* or *far fetched* world in which other entities play the roles psychons actually play without sharing their intrinsic properties (or without having any intrinsic properties at all), nor from (what amounts to the same thing) the *logical* possibility of subtracting intrinsic properties while keeping the causal properties intact.

The supposition, therefore, that we can subtract "the *phenomenal* properties of psychons [to yield] a situation in which the causal dynamics are isomorphic" is either irrelevant or begs the question. It is irrelevant if it is a claim about a far fetched world, for interactionism is a contingent thesis about the region of logical space that constitutes the 'vicinity' of our own world, and it begs the question if it is a claim about this world or a nearby world; then it presupposes what interactionism denies.

3.2.3.3 LEWIS AND ROBINSON

In 'Epiphenomenalism, Laws & Properties', Robinson (1993) discusses an argument from Lewis's 'What Experience Teaches' (1988/1999, pp. 282-85). Amalgamating features from both the original argument and Robinson's paraphrase of it, we can represent the content of the argument in yet another paraphrase, as follows. For (non-physical) phenomenal properties to be causally

efficacious, they must make a difference to something physical. Some physical states must therefore *depend* on phenomenal properties, in the following sense: a different phenomenal property would result in a different physical state. Take the phenomenal property to be that which results from tasting Vegemite, and the physical state to be a verbal response to it. Imagine that we knew that Vegemite-tasting could only bring about two different phenomenal properties, and also that the verbal response would always be either "Yum!" or "Yuk!". Suppose that the dependence relation we envisage is such that if Vegemite brings about the first taste a "Yum!" response is yielded, and if it brings about the second a "Yuk!" response comes about instead. We can then imagine an *inverted* dependence relation, yielding the opposite outcome in both cases; yielding, that is, the "Yuk!" response to the taste that under the first dependence relation elicited a "Yum!" response, and *vice versa*. Since both dependence relations are *possible*, so the argument goes, we cannot discover which phenomenal experience actually obtains, given the physical result. So, given a verbal response, nothing can be discovered about the phenomenal properties of the experience. So the argument goes.

Robinson's discussion of this argument is interesting and penetrating, and I agree that the issues are "subtle, interesting, and well worth further exploration" (p. 28). Here I wish to make just two remarks to his discussion.

First Remark

Considering various possible responses to Lewis's argument, Robinson wonders what the correct question to ask would be, if one were interested in finding out about counterfactual dependence between the phenomenal experience and the physical state. He suggests that the right question to ask might be whether, if a *different phenomenal aspect* of an experience had been the usual result of Vegemite-tastings than that which actually *is* the usual result, then would the response (the physical state) have been a different one than that

which actually usually follows? This is a question it is "difficult to evaluate", he says (p. 25).

Robinson goes on to ask which world is closer: that in which "the entire nomic profile associated with the phenomenal aspect of [the actual taste] rather than just the physical–to–phenomenal half of it, is transferred to the phenomenal aspect" of the other taste, so that the physical result that actually results from Vegemite-tasting also results there (even though the phenomenal aspect is different), or one in which the physical result would be a *different* one (Robinson 1993, p. 26). He goes on to say that "the overall pattern of laws" in the world where a different taste results in the *same* physical outcome "seems more similar to the pattern we are imagining for the actual world" than the alternative, but again admits that the intuition is unclear (p. 26).

My first remark is that the interactionist is likely to experience intuitions that are rather more clear, here. A world where a different taste leads to different physical (behavioural) consequences is most certainly closer, according to the interactionist, to our world than is a world where a different taste leads to *the same* (behavioural) consequences. The interactionist story yields, as has already been argued, very clear counterfactual dependence relations between the (phenomenal aspect of the) taste and the physical consequence. Different views may answer the question of whether counterfactual dependence suffices to avoid epiphenomenalism differently, but it would seem likely that a counterfactual theory of *causation* should force a positive answer. So, without begging the question against the interactionist, it is unclear in which sense Lewis's argument can make the phenomenal aspects of experience come out as epiphenomenal.

Second Remark

Robinson later attempts to establish *explanatory* irrelevance for phenomenal properties. In favour of explanatory irrelevance he says that

> the underlying intuition is simply the idea of separability between nomic profile and phenomenal aspect: it is inherent in the idea of a property's nomic-profile aspect that *it alone* suffices to determine the causes and effects of instances of that property, hence that any aspect of that property which is metaphysically separable from its nomic-profile aspect is redundant in the explanation of them (p. 27, emphasis added).

My second remark is the following. I think the analysis Robinson presents here is correct; it is precisely this idea that lies behind the argument that interactionism has a conceptual problem. That, however, does not affect the point that was made in the simple argument above. The fact that it is the 'nomic profile' of an entity that figures directly in the causal network, and the fact that we can speculate that entities with different intrinsic natures (or with none at all) might have exactly the same causal properties as the entities which, according to interactionism, *actually* occupy certain nodes in the causal network, does *not* establish the explanatory irrelevance of the latter.

A benefit of Robinson's analysis is that it enables a clearer statement of what the interactionist response to the conceptual challenge should be. It should be to deny that our intuition about the separability between the causal properties and the intrinsic properties of phenomenal experience is veridical in the actual and nearby worlds; to deny, in other words, that the concept of a 'nomic profile-aspect' separate from the phenomenal aspect of experience has applicability in this region of logical space.

3.2.3.4 SYMMETRY

I have argued that the supposition that we can subtract "the *phenomenal* properties of psychons [to yield] a situation in which the causal dynamics are isomorphic" is either irrelevant, or, if it is a claim about our world and nearby worlds, it begs the question against interactionism, presupposing what interactionism denies. The situation, however, is symmetric in what regards

question-begging, for the supposition that we *cannot* so subtract seems to beg the question against the epiphenomenalist position. A reasonable conclusion to draw is one which has already been endorsed, namely that whether or not phenomenal experience is causally efficacious cannot be finally decided by conceptual analysis, it relies on empirical investigation; or, more modestly, that the question cannot be 'cut short' by an analysis of this simplicity. This area of inquiry is still sufficiently ill understood to warrant caution, and perhaps we should not let superficial analysis lead us to deep conclusions. The conclusion that interactionism fails to provide what is usually thought to be its greatest advantage — causal efficacy — would be a deep conclusion indeed; so we should resist being lead to it by an analysis at this level of simplicity.

3.2.3.5 THE GROUNDING RELATION

The interactionist thesis has *inter alia* been characterised by mention of counterfactual claims which are true according to that thesis, claims such as "had the experience been green instead of red, that person's behaviour would have been different". Such claims are claims about a set of worlds which includes our own world and worlds near to it. On the interactionist account, phenomenal experiences are intrinsic properties, but they cannot be eliminated from causal explanations.[5]

It would be good to have more to say about what phenomenal experiences are like, according to the interactionist. Interactionism claims that phenomenal properties are ineliminable properties of experience (with respect to the causal properties of experience), so to shed some light on the interactionist position, let us ask what a property would have to be like for it to be *eliminable* without any effect on causal properties.

[5] It does not matter here, by the way, what we count phenomenal experience as intrinsic properties *of*; be it (say) of psychons or of experiences conceived as physical entities. For simplicity I continue to talk of them as intrinsic properties of psychons.

Characteristics of Eliminable Properties

First, the property can obviously not *itself* be causally efficacious. If it were, the causal properties of the entity it pertained to would *ipso facto* be changed by its elimination. So, a property we can eliminate without changing the causal profile of an entity must itself be causally inefficacious.

That appears to entail that the property be intrinsic. With the exception of loneliness — which on the Lewis–Langton view is extrinsic, but which we should still hesitate to call relational — and epiphenomenal properties — which *by stipulation* are caused without themselves having causal efficacy — it appears that all other extrinsic properties must be accorded at least *potential* causal efficacy by default. A property capable of standing in relations must be assumed to be capable of standing in *causal* relations as well, and in both positions of causal relations; unless it is stipulated to always be on the 'receiving end' of a one-way relation (epiphenomenalism), or in some other principled way barred from having causal efficacy. No such stipulation about phenomenal experience may form part of an *argument* that purports to *show* that interactionism fails to solve the problems of other dualist positions. Thus, the property which is claimed to be eliminable cannot be extrinsic, and must therefore be intrinsic. That is in accord with the received view of what phenomenal experience is.

It is, however, important to realise that a claim of eliminability of a property needs to say *more* about that property than that it is causally inefficacious and intrinsic. It is not enough for the purportedly eliminable property *itself* to be causally impotent; there must also be an absence of relations such that it is *in virtue of* having the *intrinsic* properties they have[6] that psychons (say) have the *relational* properties they do. In other words, if properties of a certain type are to be eliminable without loss of explanatory power they must

[6] If we suppose that psychons *are* phenomenal experiences, instead of being the 'seats' of phenomenal experiences we might say: "in virtue of *being* the phenomenal experiences they *are*".

not only be intrinsic and causally inefficacious. The relational properties of the entity from which we are supposed to be able to eliminate the intrinsic properties, must moreover not be *grounded* in its intrinsic properties.

The question thus becomes whether we can assert that phenomenal experience, in our world and in nearby worlds, is (i) causally inefficacious, (ii) intrinsic, and (iii) not such that it *grounds* relational properties. Above I argued that a feature of the dialectic situation — a symmetrical 'impasse' of mutual question-begging — should deter us from drawing such conclusions. I now wish to suggest that there may be good theoretical reasons for avoiding those conclusions as well.

Why Believe in the Grounding Relation?

Specifically I wish to suggest that it is hard to believe of *any* entity that it should have intrinsic properties that are 'disconnected' from its relational properties. This is so even for the entities posited by physics, but especially for entities such as the hypothesised psychons, whose intrinsic properties our commonsense view of the world certainly takes to be causally efficacious.[7]

Consider electrons and protons. If psychons have an intrinsic nature it is natural to think that electrons and protons do, too, unless one wishes to adopt the implausible view that some entities have, but other entities lack, an intrinsic nature. Such a view seems more implausible still than the view that the world is an insubstantial 'pure causal flux', which Chalmers rejects (1996, p. 153). Moreover, accepting that view without independent motivation would be *ad hoc*. In what follows, let us therefore suppose that psychons, electrons and protons all have intrinsic properties.

It is widely held that we are in principle barred from direct knowledge of the intrinsic properties of particles. That is not to say that they are in principle

[7] I do not mean to say that the commonsense view has an opinion one way or the other about *psychons*; it obviously does not. It does hold that phenomenal properties are causally efficacious, however, and that is enough.

unknowable *simpliciter*. If knowledge can result from inference to the best explanation, then knowledge about the intrinsic properties of particles might fall out of our best explanation of the world. Such knowledge would, however, be more indirect than knowledge resulting from, for example, measurements; direct knowledge of particles seems to have to be due to the *extrinsic* properties of those particles. Thus, even though some knowledge may be had about intrinsic properties, there is a real sense of lacking knowledge about intrinsic properties. This lack of knowledge, however, does not make it reasonable to believe that the intrinsic properties of electrons and protons are *irrelevant* to electrons and protons 'behaving' as they do. 'Playing the electron role' and 'playing the proton role' is something it is hard to believe that electrons and protons do independently of their intrinsic character.

Consider our belief that electrons and protons will continue to behave in the way we have noted that they have behaved in the past. That belief relies on induction from past to future experience, so we cannot be *certain* that our belief will be vindicated by future experience, but that admission falls well short of accepting that we have *no reason at all* to hold those beliefs. Inasmuch as we do have such reason, 'ensuring' constancy of particle behaviour seems to be a job best carried out jointly by the constancy of intrinsic properties and the grounding relation.

Without belief in a relation between the *intrinsic* properties of electrons and protons and their respective *relational* properties such that it is *in virtue of* having the intrinsic properties they have and *in virtue of* this relation obtaining that the particles have the relational properties they have — without, that is, the belief that the relational properties are *grounded* in the intrinsic properties — our belief that particles will continue to behave as they have in the past seems oddly unsubstantiated.[8]

[8] Perhaps a similar worry could be raised about our belief that the intrinsic characters of entities are stable. When I say that I believe that, in the future, particles will 'behave' like they have in the past, and

Similarly, consider our knowledge of the *difference* in behaviour between electrons and protons. How would we explain that the two types of entities behave differently, if not by reference to differences in intrinsic properties? Without the belief that the relational properties are *grounded* in intrinsic properties we will have no analysis of the difference between the relational properties of protons and electrons; no analysis that is, of the origins of the *variety* of the relational properties that entities possess in our world.[9]

Even if we should wish to allow that the 'pure causal flux' view of the world *might* provide an analysis of the constancy and variety of particles' behaviour we should be forced to moderate the above argument only slightly. Perhaps (although it appears unlikely) some other account could vindicate our belief. In that case, however, we would lack any reason to believe in the existence of intrinsic properties at all. If the properties can be varied and stay constant over time regardless of whether or not they are grounded in a variety of constant intrinsic properties there does not seem to be a role left to play for intrinsic properties. None, that is, apart from avoiding the "strangely insubstantial view of the physical world" that results (Chalmers 1996, p. 153).

It is easy to agree with Chalmers when he argues that "it is more reasonable to suppose that the basic entities that ... causation relates have some internal nature of their own, some *intrinsic* properties, so that the world has some substance to it" (1996, p. 153). However, the view that results from positing intrinsic properties, while allowing that they do not *ground* the relational properties of the entities they are properties of, seems more counterintuitive and is certainly more *ad hoc* than the 'pure causal flux' view, the implausibility of which intrinsic properties are supposed to remedy.

---

justify this by saying that I believe that the particles have stable intrinsic characters, perhaps I could just as well be asked why I believe that the intrinsic characters of the particles will remain stable. Our belief in the stability of the intrinsic characters of things is, however, entrenched and widespread, and by no means restricted to particles (or psychons). In contrast, entities whose relational properties were not grounded in their intrinsic properties would stand out as rare exceptions, among all the other entities in the world.

[9] As Lewis (1983) points out, not all accounts are analyses, so we could still have an *account* of that variety if it were taken as a primitive fact. My suggestion is that an analysis is wanted.

3.2.3.6 CONSEQUENCES OF GROUNDING

Elsewhere in the book Chalmers rejects the suggestion that our concept of electronhood relies on something other than the relational properties of electrons. The account he rejects is one according to which, in addition to sharing the relational properties of electrons, an entity would also have to share a "hidden essence of electronhood" in order for it to count as an electron in a counterfactual world:

> Arguably, physical predicates apply even *a posteriori* on the basis of extrinsic relations between physical entities, irrespective of any hidden [intrinsic] properties. This is a purely conceptual question: if electrons in our world have hidden protophenomenal properties, would we call an otherwise identical counterfactual entity that lacks those properties an electron? I think we would. ... The notion of an electron that has all the extrinsic properties of actual protons does not appear to be coherent, and neither does the notion that there is a world in which mass plays the role that charge actually plays. The semantic account given above predicts that these notions should be coherent, and so gives a false account of the concepts (pp. 135-36).

This objection does not pose a problem for the present proposal. The view urged here is not that an entity would have to share *the very same* intrinsic properties grounding its relational properties in order for it to count as an electron in nearby counterfactual worlds, nor even that nothing could ever count as an electron unless it had an intrinsic nature. Rather, the claim is that our concepts of electronhood, protonhood etc., entail that in all the worlds that are close to ours, there is *some* categorical basis for the relational properties of electrons such that the intrinsic properties ensure the stability and variety of the relational properties. It is perfectly consistent with what has been said that an electron in a nearby counterfactual world has *different* intrinsic properties grounding its relational properties, so long as it has *some* intrinsic properties that so ground them. What is not consistent with the present view is that

intrinsic properties in nearby worlds be disjoint from relational ones in the sense that the relational properties are *not* grounded in the intrinsic ones.

Thus I resist the conclusion that we would call a counterfactual entity an electron if it 'resided' in a nearby possible world and 'behaved like' an electron, but lacked grounding intrinsic properties. That applies, I maintain, both when the entity lacks intrinsic properties altogether *and* when the entity has intrinsic properties but lacks the relation of grounding between the intrinsic and the relational properties. That is just to say that worlds where electrons, protons and such lack intrinsic natures, or have intrinsic natures but the grounding relation does not hold between the intrinsic and the relational properties, are far fetched worlds.

Thus, talk of a 'hidden essence' of electronhood misses the mark here. The present view is silent on the question of whether there is an essence of electronhood. Indeed, it follows from what has been said that a number of different intrinsic properties may ground the electron role in various instances. The view urged here opens for multiple realisability of electronhood across possible worlds. Indeed, we cannot rule out that the possibility electronhood is multiply realised in our own world, too. It may be that many different intrinsic properties ground a smaller number of different relational properties, so that the grounding relation between intrinsic and relational properties is many–one. Such an account would not offer quite as neat an explanation of the variety of causal profiles that we observe.

An inference to the best — the most parsimonious — account, would lead us to suppose only that there is *one* property (or set of properties) such that on *our* world it is in virtue of an entity having *that* (*these*) properties that the entity has the relational properties of electronhood. That conclusion, however, is at present still uncertain.

The present suggestion does *not* predict that the notion of an electron playing the proton role (or vice versa) should be coherent. The current

suggestion is that our concept of electronhood is a 'package deal', containing both (i) the relational properties of electrons, and (ii) the existence of *some* intrinsic properties such that it is *in virtue of* the intrinsic properties obtaining that the relational properties obtain, i.e. such that the intrinsic properties *ground* the relational properties. Alternatively, we can describe the suggestion by saying that, on this analysis our concept of an electron relies on (i) intrinsic properties, (ii) relational properties, and (iii) a relation of *grounding* between (i) and (ii). An entity in a nearby world without one of the (possibly various) intrinsic properties that *ground* the electron role would not be an electron, but neither would an entity with one of the intrinsic properties (or sets of properties) that *usually* ground the relational properties of electrons, but where the relation of grounding did not obtain; where the intrinsic property did not *in fact* ground the relational properties.[10]

One might sum up the situation in this way: the correct conceptual analysis of electronhood according to Chalmers is that all it takes for an entity to qualify as an electron is for it to *behave like* an electron. I have suggested that an alternative view may be better, according to which, in our world and in nearby ones, something does not qualify as an electron unless its relational properties are *grounded* in stable intrinsic properties.

3.2.3.7 ANOTHER PROBLEM?

It seems that the problem discussed in this section is spurious; but another may be lurking in the offing. It was argued that a problem for any view which denies

[10] Perhaps one might argue that, if the intrinsic properties of *all* entities were *identical*, one could subtract them from the explanation without loss of explanatory power. Would such a position be better? No. Above it was noted that, even granting that a view which denies the grounding relation might somehow come up with an explanation of the stable and differentiated behaviour of the different entities, such a position would leave us without any reason to believe in intrinsic properties of entities. There would be no role left to play for them. Similarly, a view which posits that the intrinsic natures of all entities are identical also leaves us without any reason for believing in the intrinsic properties of entities in the first place. If these properties are not explanatorily linked to the varied behaviour of entities, the question still remains of why we should believe that they exist. What reason remains for resisting the conclusion that the world is a 'pure causal flux'? The sole remaining justification for positing intrinsic properties would again be to avoid the uncomfortable feeling that the world is 'insubstantial', and that is hardly enough.

the grounding relation while affirming the existence of intrinsic properties, even if somehow able to account for the stability and variety of relational properties, will be unable to provide any further reason for believing in the intrinsic properties in the first place, apart from the mere 'uncomfortableness' of a 'pure causal flux' view. It was also argued, however, that the view that some entities have intrinsic properties while other entities lack them, is implausible. In combination, these two considerations open a route to a different conclusion than the one I have suggested we should accept.

Someone might justify the existence of intrinsic properties by insisting that we have direct knowledge of *phenomenal* properties, which are intrinsic, only to invoke the consideration that if some entities have intrinsic properties, it is plausible to think that all do. Such an account would justify belief in intrinsic properties all round while denying that the grounding relation obtains, either only for psychons, or generally. It would either have to take the stability and variety of the relational properties of different entities as a primitive fact, or explain this in some other way than by appealing to the grounding relation.

It is hard to imagine how such an explanation could proceed. It is therefore reasonable to conclude — perhaps pending further developments — that the stronger view is that the constancy and variety of relational properties is 'guaranteed' by the fact that the relational properties are *grounded* in a variety of stable intrinsic properties. This view affords the intrinsic properties of psychons the causal efficacy our intuitions would have us believe that they have, yields a world which is not insubstantial, and on this view the constancy of relational properties is analysed in terms of the constancy of intrinsic properties, it is not taken as a primitive fact.

3.2.3.8 CONCLUSION

If the view I have suggested is correct, the distinction between causal and explanatory relevance collapses. Chalmers claims that the *explanatory*

irrelevance of phenomenal experience — so problematic for epiphenomenalism — will arise nearly as strongly for any view that takes phenomenal experience to be causally efficacious. That claim relies on the assumption that this region of logical space contains worlds in which the intrinsic properties of an entity are disjoint from its relational properties. Interactionism rejects that assumption.

I have, moreover, suggested that it is part and parcel of our concepts even of what we take to be some of the basic physical entities in our world that their relational properties are both varied, and for each variety constant, and that, unless we are prepared to suppose that such constancy and variety could obtain in a world of 'pure causal flux', the best explanation for the constancy and variety is that relational properties are *grounded* in variety of constant *intrinsic* properties. It seems likely that is *in virtue of* the intrinsic properties being what they are *and* the grounding relation obtaining that the constancy and variety obtains.

If that is so, and it is true that the notion of intrinsic properties that do not ground relational properties is unsustainable in this region of logical space where basic physical entities are concerned, it must be doubly so where psychons are concerned. Psychons are meant to be entities whose intrinsic properties are phenomenal, and in contrast to the case with basic physical entities, the intrinsic nature of psychons is one we *do* have knowledge of. Our strong intuition is that phenomenal experience is causally efficacious; the phenomenal character of a humiliating experience is, for example, very much responsible for what normally ensues: subsequent avoidance behaviour of similar situations. Then the claim that the phenomenal component can be subtracted from any account without loss of explanatory power must be rejected. The conceptual problem has been resolved.

Thus, if considerations along the lines of what has been suggested in this section are anything to go by, interactionism remains a dualist position which avoids the problems that face epiphenomenalism on account of the explanatory

irrelevance of phenomenal experience. According to interactionism, phenomenal experience is causally relevant, and it is explanatorily relevant as well.

Even if the arguments suggested in this section do *not* go through, there is still the simple argument for the explanatory relevance of phenomenal, given in the previous section, to fall back on. That argument still seems decisive. And even for someone who rejects that argument as well as the suggestions in this section, it would be difficult to deny that these issues are far from clear, and that there is much work left to be done in this area is. And that realisation is, in a sense, *itself* sufficient to back what this thesis attempts to establish. We do not *yet* possess decisive evidence against interactionism, and, *as of yet* interactionism has an important advantage over epiphenomenalism in that it establishes causal and explanatory relevance for phenomenal experience, in accordance with our commonsense view of the world. That much stands firmly.

3.3 Defining Interactionism

Interactionism depends on the following thoughts. There are two sets of entities, one of which is a proper part of the other. The smaller set is that which physicalists intend to pick out when they gesture toward the set of physical things. The larger set is the set of all the entities that causally interact. Some entities that are members of both sets sometimes stand in causal relations with entities that are members only of the larger set, and the causal relations go both ways.

This work undertakes the project of demonstrating that no evidence we currently possess excludes interactionism. I think this is a promising project. It seems, however, that it is an unintelligible one, unless it is supposed that there is some way of referring to the smaller of the two sets mentioned just previously.

To complicate matters, I am sceptical about the prospects of articulating physicalism in such a way that the view that emerges picks out a set of all and only the causally interactive entities as small as that which I believe physicalists wish to pick out.

A typical way for physicalists to attempt to articulate their view is by referring to what they take to be paradigm examples of physical entities or processes, and then to say that all the things there are (or all the processes there are) are similar *enough*, or *very* similar, or *relevantly* similar, or ..., to those things (or processes). What the physicalists are gesturing toward here is a *comparative* similarity relation of the form that *b* is more similar to *a* than *a* is to *c*, where (i) the supposition is that for any entity *b*, known or to be discovered, that entity will be more similar to the paradigmatically physical entity or process *a* than *a* is to *c*, and where (ii) *c* is a *foil*, that is, the 'nearest miss' there is for counting as physical: the first entity we should count as non-physical, going 'outward', gradually acquiring 'distance' to the paradigmatic case.[11]

The problem is that no straightforward way to give content to this relation is immediately available. Some argue that there is *no way at all* to give content to the relation (Crane and Mellor 1990). It appears, however, that interactionism is just as dependent on there being a way to give this relation content as is physicalism; if there really is "no divide" between the mental and the physical, as Crane and Mellor claim, interactionism comes out just as vacuously true as does physicalism (p. 206). So it seems that scepticism about the similarity relation in question sits uncomfortably with an interactionist position, at least so long as the position portrays itself as a genuine *alternative* to physicalism.

I think, however, that a more careful analysis reveals that there is no real problem for the interactionist here.

[11] For a discussion of this similarity relation, see Quine's 'Natural Kinds' (1969).

First notice that the position that no evidence as yet available excludes interactionism does *not* depend on there being a non-vacuous way of picking out the smaller of the two sets. An interactionist can 'piggyback' on the expressions used by the physicalist, and say that talk about the smaller set should be understood as talk about whichever set the physicalist is talking about, *provided* that the physicalist is talking about some determinate set at all. Claims about the smaller set, then, can be interpreted as being conditional. Whichever set the physicalist is referring to, I refer to it as well, and about it — provided that it can be picked out — I say that we have no decisive evidence as yet to the effect that it is causally closed.

Secondly, my suspicion is that when the relation in question is cleared up, it will not deliver as small a set as the physicalist thinks it would. I think, in other words, that there are surprises in store for a physicalist willing to clarify the similarity relation and see what science will tell us that it encompasses. Depending on the degree to which this suspicion is well founded, it may be that the set that the relation, when cleared up, *actually* points to *does* turn out to be the set of all and only the causally interactive entities. An interactionist stance toward *that* set would, of course, be false. That however, does not entail that interactionism as a stance toward the set which *today's* physicalists *think* they gesture toward is false. It may well be that among the entities that surprise the physicalists of today by ending up within the set they were gesturing toward are *all* those that the interactionist of today wants to insist are part of the causal network.

So, my suspicion is that when we get a better grip of the similarity relation physicalists rely on without cashing out, we will see that it yields a (for the physicalist) surprisingly large set.

It may not. My suspicion may turn out to be unfounded, and physicalists may look at the set their similarity relation in the end picks out with satisfaction, congratulating themselves on their foresight in loosely gesturing toward a

relation that picks out exactly what they wanted. It may even turn out that the physicalist claim that *that* set is the set of all and only the causally interactive things is correct. Even *that* result is compatible with the view defended in this thesis, for the view defended here is merely that have insufficient evidence to assert this conclusion *yet*. (If the claim argued here is right, physicalist could then congratulate themselves on *foresight*, not on their ability to draw conclusions from evidence.)

Dualism is often taken to be a claim about the make-up of ultimate reality, as revealed by 'ultimate science'. The dualist claim, it is said, is the claim that we shall ultimately find that reality is made up of two fundamentally different *kinds of things*: the mental and the physical. John Bigelow and I argue elsewhere, however, that reflection on the dualist intuition reveals that dualism is *not* best understood as a position about ultimate reality.[12] Dualism is, rather, a position about the *difference* between the science of the day and the science believed to be a plausible candidate for one that can account for the mind. Understood in this way, the proper question to ask is not whether someone is or is not a dualist, but how much of a dualist they are.

On this view, a defence of interactionism does not commit one to a view about a certain make up of reality as it will ultimately be revealed by a perfect science. Interactionism only commits one to a claim about *how different* from today's science a science would have to be, to be a plausible candidate for one that can account for the mind. Quite different, I say. Not very different at all, says the physicalist. There is a plethora of positions to be had; Bigelow and I suggest that we may systematically categorise them in terms of reference to various upheavals in the history of science.

This explains how an interactionist can allow for the possibility that the similarity relation the physicalists are gesturing toward may actually turn out

[12] (Bigelow and Koksvik Unpublished); Appendix A.

to yield the set of all and only the causally interactive things. All the interactionist should insist on is that if it does, then the set it yields will contain elements which it is surprising to physicalists that it does contain. *How* surprising, and *how many* (types of) such elements there would have to be, is again a question of *how much of a dualist* the interactionist is. If there are *no* surprising elements in the set, interactionism is false. What this thesis argues is that we cannot yet ascertain that that is how it will turn out. (We can still take bets, of course!)

Interactionism depends on there being two sets of entities — one a proper part of the other — and the smaller being that which physicalists *intend* to pick out when they gesture toward the set of physical things. By presuming to be able to talk about these matters at all I am obviously presuming that I have some not too-far-off idea of which set the physicalist *intends* to refer to. Inasmuch as I am wrong about this, I may also be mistaken about the degree to which physicalists will be surprised by the contents of the set of all causally interactive entities.

The physicalist may be mistaken, too. The set of things picked out by the relation they gesture toward when they talk about *relevant* similarity and difference and other things in that vein may be larger than what they expected. My reference is parasitic on what physicalists (attempt to) refer to. If the physicalists are mistaken about the contents of the smaller set, there will be some elements of that set whose inclusion will be surprising to physicalists. Clearly, the mistake can be small or large, there may be large or small surprises in store for the physicalist. As the mistake gets larger, there will be a point such that it makes the interactionist claim, defined parasitically, mistaken, too.

It should be emphasised, however, that *that* mistake would be entirely unthreatening to the interactionist. If the account Bigelow and I defend is correct, what really matters for the interactionist is that certain entities participate in the causal network of things. If it turns out that the physicalist

agrees with that, those who were interactionists should generously allow those entities to be called physical, if the physicalist should so wish. Nothing much hinges on the label. A lot hinges on their inclusion in the causal network.

Let us put this point a different way. Interactionists claim, about some of the things they *believe* that physicalists wish to exclude from the causal network (by reference to a similarity relation which physicalists *think* they indicate), that they belong in the causal network, and not outside it. If physicalists find, to their surprise, that many more entities are picked out by the relation they gesture toward (and they still think it is the right relation), the disagreement may be dissolved. If interactionists find that they were wrong in their assumption about the entities the physicalists wanted to exclude (and they agree with the inclusions and exclusions), no tear should be shed.

The position defended in this work can therefore be said to be *fragile* to an improved understanding of what the similarity relation which physicalists gesture toward actually is, and to an improved understanding of what that relation encompasses. If it turns out that it encompasses all the important entities, interactionism toward that relation will be false.

Furthermore, the position is also fragile to talk at cross–purposes, where what interactionists thought that physicalists were saying is not, in fact, what they were saying at all. So the claim made in this work — that we do not *yet* possess evidence that should lead us to exclude interactionism as a viable alternative — depends on the degree to which the assumptions that are made, about meaning and understanding, are correct. That is, I think, as it should be. While interactionists should obviously acknowledge and identify *unwelcome* ways to be wrong (and they do; a completed account of our behaviour that does not include reference to phenomenal experience is an unwelcome outcome), *this* way of being wrong is not unwelcome. It is a welcome way of being wrong, because by being wrong about things that do not matter, interactionists can be right about things that do.

Chapter 4: Conservation of Energy

4.1 Introduction

Interactionist dualism is sometimes dismissed on the grounds that it is supposed to be at odds with certain well established results in science. In this chapter I discuss, evaluate, and ultimately reject what I take to be a particularly influential claim of incompatibility, namely the claim that interactionist dualism cannot be true because it is incompatible with the Conservation of Energy law in physics (CoE). Since CoE is almost certainly true, so the argument goes, the incompatibility between interactionism and CoE should persuade us to deny the former.[1]

Section two of this chapter considers a likely explanation for the widespread belief that CoE is true and universally applicable. Section three discusses an important question concerning the interpretation of CoE; section four explains the appearance of incompatibility between interactionist dualism and CoE. While I think that interactionist dualism can be shown to be compatible with CoE, it is nevertheless important to understand what lies behind the prevalent belief that the two are incompatible.

The majority of the chapter is concerned with the study and evaluation of various *defences* one might mount against the incompatibility claim. In section five, two different ways of ordering the defences into groups are briefly considered. In sections six, seven and eight, the defences themselves are discussed. From the discussion it emerges that some defences must be rejected, but also that effective defences against the incompatibility claim are very much

[1] Alternatively CoE could be interpreted as a *reason* one might give for belief in the principle of causal closure of the physical. I take it that no argumentative force could be mustered against interactionism from CoE using an indirect route, *via* causal closure, that cannot be mustered directly against interactionism; CoE would, I think, be seen as a reason for belief in causal closure only *in virtue of* being seen as a reason to deny interactionism, so by discussing it as such, I do not think I am taking any shortcuts.

available. Hence it is concluded, in section nine, that the incompatibility claim that is anchored in CoE is an ineffective argument against interactionism.

4.2 The Status of CoE

In a recent book, Penrose takes CoE to be the claim that "total energy is conserved in any isolated system" (2004, p. 690). In *Thinking About Consciousness*, Papineau explains that "when all forces are ... 'conservative', then the sum of actual kinetic energy ... plus the potential to generate more such energy ... is conserved: when the particles slow down, this builds up 'tensions', and if those 'tensions' are expended, the particles will speed up again" (2002, p. 244). A force is considered conservative if it is "independent of the velocities of the interacting particles and of the time", and this requirement, he says, is one which all "fundamental forces" are now taken to satisfy (p. 244).

It is clear that the law enjoys very widespread acceptance. Many prominent philosophers believe that the principle excludes the possibility that interactionist dualism is true.[2] They would not believe that if they did not think that the principle will turn out to be true, not just in a great many cases or in a large portion of the world, but at least in a set of cases such that it includes the brain (where interactionists suppose that the interaction takes place), and more likely simply in *all* cases *everywhere*. How do we explain that belief?

The explanation is not that we direct have experimental support for the universal conclusion. We have not yet examined the brain and found its workings conservative, and we have still less investigated all types of energy-transactions in all relevant circumstances and found them conservative. The belief in the universal applicability of CoE must have a different source.

It is very likely that the belief derives largely from the knowledge that there are noteworthy examples of forces that were initially thought to be non-

[2] See Montero (forthcoming, p. 2), who lists Leibniz, Dennett, Fodor, van Inwagen, Putnam and Crane.

conservative, but that have been found, upon closer investigation, either to be conservative after all, or else to be *reducible* to conservative forces. Frictional forces, for example, depend on the speed the body has relative to its surroundings. Since a force is considered conservative only when it is "independent of the velocities of the interacting particles and of the time", frictional forces are *not* conservative. Indeed, the failure of conservation is conspicuous where friction is concerned. A rock thrown upwards decelerates and finally stops, but will then accelerate again, converting the potential energy it built up in its ascent back into kinetic energy during the decent. In contrast, a body which decelerates due to friction is *not* prone to accelerate again due to the release of any built up 'tension' or potential energy. So frictional forces appear to constitute a counterexample to CoE.[3]

It has been shown, however, that the forces involved in friction are conservative after all, contrary to initial appearances. Even though it looks like there is no stored up tension waiting to accelerate the decelerated body, energy is nevertheless conserved on the micro-level. The right amount of *heat* — which is, of course, energy in a different form — is generated to compensate for the decrease in kinetic energy. So the frictional forces are not basic, but they supervene on forces that are, and the forces they supervene on are conservative. When a falling body accelerates more slowly than it otherwise would have due to resistance from the surrounding medium, the apparent loss of energy is 'made up for' by an increase in temperature both in the falling body itself and in the surrounding medium. The conservative processes are not apparent, but they do occur.[4]

It is likely that such discoveries have played an important role in encouraging the prevalent belief that CoE should be regarded as a particularly certain and universally applicable result from science. The tendency has been

[3] See Papineau, (2002, p. 244).
[4] See Penrose (2004, p. 690).

that forces that seemed to be non-conservative turn out to be conservative after all, so a widespread view is that we should expect this pattern to repeat itself. Even if we discover new forces that at first appear to not conserve energy it is thought that these will almost certainly turn out either to be conservative themselves (which just means that our initial observations were inaccurate), or they will turn out to be reducible to a set of more basic forces which *are* conservative. Apparent failures of conservations will turn out to be spurious.

This way of justifying the belief that CoE is universally true is aptly described as an (optimistic) *meta-induction on the history of science*. Chapter five argues extensively for the claim that no argument on this schema can add independent weight against interactionist dualism, and section 4.8.2 below applies the same strategy against the claim that CoE holds universally. In this chapter we shall, however, also see that the claim that interactionism is incompatible with CoE must *itself* be rejected; interactionism and CoE can peacefully co-exist. Thus the conclusion that interactionism is a perfectly tenable position to hold — also for scientifically–minded philosophers of mind — is strengthened, both via a rejection of allowing a meta-induction on the history of science to result in an argument against interactionism *and* independently of this route.

4.3 What Is a Closed System?

The conservation of energy law says that in any *closed* system, energy is *conserved*. Energy is *conserved* in a system if the total quantity of energy in the system is unchanged over time; if "there is at every moment in time the same total amount of energy" (Dieks 1986, p. 90). But what is it for a system to be *closed*?

For the purposes of CoE it is natural to think that a closed system is such that energy *does not flow* or *is not transferred* into or out of it (Dowe 2000, p. 95;

Halliday *et al.* 1997, p. 167). Call such a system an *e-closed* system. A good reason to think that e-closure is the right notion of closure for the purposes of CoE is that using e-closure in CoE yields a result that accords well with the 'slogan version' of the law. The slogan version says that energy is neither created nor destroyed; it "cannot magically appear or disappear" (Halliday *et al.* 1997, p. 168). If energy is neither created nor destroyed, and if a system has 'energy-tight' borders, then energy will remain constant in that system. Thus, if a system is e-closed and energy is not created or destroyed within it, the total quantity of energy within the system will remain unchanged, which is what CoE says.

There is, however, a complication.[5] Talk of *transfer* or *flow* of energy seems to rely on the idea that energy is identifiable and re-identifiable through time. If there is no way to determine that *this* is the same energy as *that*, a case where energy has been transferred (has flowed) from one entity to another will be indistinguishable from one where the energy level of one entity *simply rose* shortly after it had fallen in another. To be substantive, CoE (formulated in terms of e-closure) relies on the distinction between these two cases.[6] Thus, when a closed system is taken to be an e-closed system, it is necessary to be able to identify and re-identify energy in order to maintain the substantiveness of CoE.

But the idea that energy has identity over time is problematic (Dowe 2000, pp. 55-59; Dieks 1986, pp. 87-91). Dowe, for example, concludes that "energy generally doesn't have identity through time" (p. 59), and Dieks states

---

[5] I am grateful to Toby Handfield for bringing the complication to my attention, and for helpful discussion on the matter.

[6] If one were considering a purportedly e-closed system while looking for experimental falsification of CoE, discovering that energy had been transferred to an entity in the system would lead one to conclude that the system was *not e-closed* after all (although perhaps an extended system that was e-closed could be found in the vicinity), and one would *not* then think that a violation of CoE had been discovered. On the other hand, if the energy level of an entity inside the system *just rose*, that would be a case of a e-closed system where energy was *not conserved*, and would therefore constitute a violation of CoE.

that "the idea that energy ... possess[es] an identity of [its] own which is retained throughout physical interactions is already foreign to classical physics", and in quantum mechanics "[t]he mere *ascription* of an identity to the energy elements ... entails consequences which are observably wrong (pp. 88, 89).

If these difficulties with accounting for the identity of energy over time are serious, the notion of e-closure is threatened. Then a definition of a closed system *independent* of the notion of transfer or flow of energy would be needed. Formulating such a definition without begging the question against interactionist dualism is not easy.[7,8]

---

[7] Here is one attempt:

> *A closed system is a system such that no eligible pair of entities, one of which a part of the system and the other not, form a conservative system.*

This goes some way toward capturing the intuition that a closed system is one that does not get energy 'given to' it and that does not itself 'give away' energy to something outside of it, but without talking about 'flow' or 'transfer' of energy. Call a system e-closed* if it is closed in this sense.

What might CoE amount to if it is taken to make a claim about e-closed* systems? It might amount to the claim that *all systems such that no eligible 'cross-border pair' of entities conserve energy, conserve energy.* Does this strange expression still express a substantive claim? Let us ask what it would be for the claim to be false; what it would be for a system to be e-closed* but not conservative.

For CoE to be false on this understanding of it, the energy of some entity *within* the system would have to be distinct at times t_1 and t_2, *without it being the case* that the difference in the energy level between times t_1 and t_2 of some *other* entity such that the two would form an eligible pair — either *inside* or *outside* the system — were equal but of opposite sign. (*Mutatis mutandis* if the equal but opposite change in energy pertained to a *set* of entities.) If there was an equal but opposite difference in the energy level of some entity *inside* the system, the system would be *conservative* after all; if such a change obtained in some entity *outside* the system such that the two would form an eligible pair, the system would not be *e-closed**.

What is an *eligible pair* of entities? Suppose you had two systems, both containing a large number of entities with energy, both purportedly e-closed*. Then one could easily construct a number of cross-border pairs — pairs such that one member was an entity in the first system and the other an entity (or set of entities) in the second — which conserved energy. It would be too easy to counter a system's claim to be e-closed*; the only system protected from the challenge would be the set of all entities with energy. Thus the notion of eligibility is introduced to restrict the range of admitted counter-examples to claims of e-closure*.

How the notion of eligibility could be cashed out is unclear. (Perhaps counterfactual dependence is important: the energy of *this* entity would not have changed had the energy of *that* one not changed.) However, absent identity of energy over time it is clear that some such notion is needed. Our definition of e-closure* should make the question of whether there are e-closed* *subsystems* to the system of all entities with energy come out contingent. Without a restriction on the eligibility of pairs of entities it looks like the question could be settled *a priori*, for it seems likely that it would be possible to construct counterexamples to any purportedly

Call someone wishing to mount an argument against interactionism on the basis of CoE a 'critic'. Here I wish to give the critic the best shot possible. For the purposes of this chapter I therefore assume that a closed system can be defined in terms of flow or transfer of energy into and out of a system, or in terms of some reconstructed notion that does the same job.[9] This has the effect of making a defence against the critic *harder*, not easier, so no unfair dialectical advantage is gained.

4.4 The Relevance of CoE to Interactionist Dualism

In the recent article 'What Does the Conservation of Energy Have to Do with Physicalism', Montero (forthcoming) observes that "many ... hold that the conservation of energy law is inconsistent with interactive dualism", an inconsistency which "would lead to an argument for physicalism" (p. 3).[10,11] Montero argues, however, that no valid argument against interactionism can be constructed without also taking on board, as extra premises, contentious claims about causation and the nature of entities that have energy. Thus she argues that, although the argument that results from such augmentation is valid, CoE becomes a redundant premise in it, and the title question of her paper must be

---

e-closed* subsystem, just by carefully selecting a pair of entities. All that would be required is for a change in the energy level of an entity *outside* the subsystem to take place at the same time (between t_1 and t_2) as a change in the energy level of an entity *inside* it, and for it to be equal but of opposite sign. Thus the notion of eligibility of pairs would have to be brought in to capture a 'residue' of the *relevance* that the notion of 'flow' ensures that the members of conservative pairs have to each other.

[8] I argue below that the critic's intuitions are not adequately accounted for without defining a closed system for the purposes of CoE as an e-closed system (see section 4.4.4), so I do not think that the difficulties with e-closure should persuade one to accept a different notion of closure instead.

[9] For a suggestion, see (Dieks 1986, p. 90).

[10] The paper has not yet appeared in print. The page numbers refer to the 'online early' version of the paper, available from http://www.blackwell-synergy.com/toc/dltc/0/0.

[11] I will speak as if Montero considers an argument against interactionism throughout, even though she often frames the discussion in terms of an argument in favour of physicalism. This inaccuracy is inconsequential since, on the premise that there are causal relations between the mental and the physical — a premise that figures in all the arguments Montero considers — an argument for physicalism is also an argument against interactionism (and *vice versa*).

answered with "nothing whatsoever": "the conservation of energy law ... is irrelevant to the question of whether the mental is physical" (p. 13).

It is certainly true that arguments for physicalism (or against interactionism) anchored in CoE are seldom enunciated. It is true, also, that the incompatibility claim has been allotted importance and weight that it does not have, and that, consequently, the enterprise of reconstructing and evaluating these arguments is an important one. This notwithstanding, Montero's argument for the outright *irrelevance* of CoE to interactionism fails to adequately account for the intuitions behind the incompatibility claim, and is, therefore, unsatisfactory. I summarise Montero's argument, consider a general argument for the *prima facie* relevance of CoE to interactionism, and finally attempt to explain where I think that Montero's argument goes wrong.

4.4.1 PURPORTED IRRELEVANCE

Montero spends some time justifying her preferred reconstruction of the argument for physicalism from CoE. Reconstructed in this preferred "general form" (p. 6), the "Argument from the Conservation of Energy" (ACE) appears as follows:

1. Energy is conserved in any closed system.
2. The universe is a closed system.
3. There are causal relations between the mental and the physical.
4. Causation involves the transference of energy.

Thus: The mental is physical (p. 3).

To make the argument conclusive, Montero points out that another premise has to be added: "anything with energy is physical" (p. 11). When this last premise is appended, however, the first two premises are no longer needed to establish the conclusion.

With the first two premises subtracted and the remaining premises appropriately renumbered we get what Montero calls the "Real Argument for Physicalism" (RAP):

1' There are causal relations between the mental and the physical.
2' Causation involves the transference of energy.
3' Anything with energy is physical.
Thus: The mental is physical (p. 12).[12]

This argument, which according to Montero "lies behind the supposed argument for physicalism from the conservation of energy" is valid, but, of course, CoE is conspicuously absent from it (p. 12). So, Montero concludes, CoE turns out to be "irrelevant to the question of whether the mental is physical" (p. 13).

4.4.2 THE *PRIMA FACIE* RELEVANCE OF COE

The effectiveness of Montero's tack here obviously depends on the degree to which this reconstructed argument is the 'real' argument for physicalism. If Montero's criticism is to be effective, the 'real' argument must represent the result of our best efforts to reconstruct an argument on behalf of the critic that accounts for the critic's intuitions. It seems implausible, however, that the closest we can get to a argument in this vicinity should involve making contentious claims both about the nature of causation (that it involves transfer of energy) and about the mind (that it has energy).

Consider the possibility that an omnipotent being exists. Call this being, about which we need make no other assumptions, '*It*'. If *It* exists, *It* can bring something about in any logically possible manner whatever. The way in which *It* chooses to bring something about might be so alien to us that if we observed *It* bringing something about, we should be forced to characterise the process as

[12] For ease of reference I have distinguished the premises in RAP from those in ACE by adding a prime symbol to their number.

miraculous or, perhaps, magical. Suppose that *It* uses *It's* powers to move a single particle in a brain. Then the total level of energy in the brain after this occurred would be different to what it was before *It's* interference. On the assumption that *It's* (magical, miraculous) power does not go along with the usual changes in adjacent physical systems (equal but of opposite sign) — an entirely unproblematic assumption — CoE has been violated.[13]

Now notice that nothing salient about *It* is assumed here that, according to the interactionist, cannot also be assumed about the mind. The only crucial ability the example allocates to *It* is the ability to cause a particle to move *without* it being the case that any of the changes we usually observe in adjacent physical systems (in particular: changes in energy that are equal but of opposite sign) occur as well. That, of course, is *precisely* the ability that I have argued that the minimal interactionist position holds that the mind *does* have.

So, in order to arrive at the conclusion that interactionism would violate CoE, the 'minimal' counterfactual theory of causation is quite enough. Interactionism is committed to the claim that the non-physical mind sometimes causes a change in the properties of a particle in the brain (where A caused B iff B would not have occurred if A had not occurred) *without* it being the case that the usual changes in adjacent systems take place.[14]

---

[13] I have been told that some (sub-atomic) particles are such that it does not make sense to talk about their identity. If I have not been misinformed, we can rewrite the example so that *It* simply *destroys* a particle only to *create* another one exactly like it, and therefore identical to it, in a different location in the same brain afterwards. Then CoE has been violated. Some might be worried that something like commonplace physical causation has somehow been smuggled into the example when *It* uses *It's* (magical, miraculous) power to *move* a particle. If so, I hope this revised version serves to soothe their intuitions.

[14] It need not, strictly speaking, be a core claim of interactionism that what I have called the usual concomitant changes in the physical system do *not* occur. It is consistent with interactionism that the mind, as well as causing such changes in the brain as have downstream behavioural consequences *also* causes a change in the adjacent systems, for example by increasing the heat of one of those systems. This is what I call a local compensation view below; the mind compensates for its own activity in such a way that CoE is not violated by it. Interaction would still constitute a violation of CoE (albeit an undetectable one), on some interpretations of that law, since both the original change and the concomitant changes the mind brings about (the latter 'to disguise itself') would be caused by the mind, and not in ways

So an argument for physicalism from CoE can get started without the supposition that causation involves *transference* of energy. All that is required to get the argument started is that the type of mind–brain causation that interactionist dualism entails somehow *makes it the case* that a change occurs in the brain such that the energy level of the brain after the change differs from what it was before, without any accompanying and compensating change elsewhere in the physical world. Thus, Montero's claim that "without any of these three premises [the valid argument for physicalism] would not succeed" (p. 13) is *false*, and CoE has *prima facie* relevance here.

As Montero points out, the transference theory of causation is disputed. A charitable interpretation of the argument from CoE to physicalism must allocate it the minimal theoretical burden possible. In particular, a disputed theory of causation must not be attributed to it when that is not necessary. Since 2' is stronger than is has to be, the argument that contains it is *not* the 'real' argument for physicalism.

4.4.3 DIAGNOSIS

I argued above that e-closure is the most natural concept to employ when considering CoE. An important aspect to notice about Montero's argument for the irrelevance of CoE to interactionism is that she operates with a very different concept of closure. About her notion of closure, Montero writes:

> [T]he notion of being a closed system here is distinct from the notion of being causally closed as it is used in the [causal] argument for physicalism ... [where] the physical world is said to be causally closed, roughly, if every physical effect that has a sufficient cause has a sufficient physical cause. Here, 'a closed system' means that the system *neither affects nor is affected* by anything from outside of it (p. 3, n. 12, emphasis mine).

---

that conserve energy. (More about interpretations of CoE below.) I think this is an uninteresting and unattractive option for the interactionist. Since there are far better defences available, I set this one to one side.

Call this the principle of *affective* closure. Montero indicates that she takes it to be distinct from causal closure, and it seems that it may be stronger. Perhaps one can imagine that a system can affect another without there being causal relations between the two. An epiphenomenalist, for example, might subscribe to both the transference theory of causation and CoE without seeing a contradiction therein; perhaps thinking that the brain 'gives rise to' the mind without *causing* it to arise. However, if the physical world were *affectively* closed, even epiphenomenalism would be excluded.[15]

With this notion of closure in hand, Montero argues that the critic cannot suppose that the *physical world* is closed. On her account, claiming that the physical world were affectively closed would be to beg the question outright against interactionist dualism: if the physical world is not affected by anything outside of it, and if the mind is not physical, it follows that the mind does not affect anything in the physical world, and interactionism is false. Thus Montero concludes that the only closed system that the critic can rely on is the universe as a whole (p. 4).

Could the critic grant that the only closed system is the universe while also holding that the conservative system constitutes a *sub-system* of the closed system, and in particular the *physical* sub-system? Even given the universe as the only closed system, an argument against interactionism on the basis of CoE could still get off the ground if the mind was thought to be *inside* the closed system but *outside* the conservative one.

---

[15] It seems to be this strong notion of closure that leads Montero to argue that ACE is "just as much an argument against epiphenomenalism ... as it is an argument against interactive dualism" (p. 6). That is, it seems to me, further reason to think that the notion of affective closure is the wrong one here. Proponents of the incompatibility claim do not worry that energy would "disappear" if epiphenomenalism were true (as the quotations in Montero's paper themselves show), and epiphenomenalists do not take the brain to have to expend energy to give rise to the mind. Campbell, e.g, argues that "for the production of spiritual effects by material causes, it is no part of the conservation principle that the production of non-material effects requires physical energy" (1970/1980, p. 53).

No. Montero goes on to argue — in my view unconvincingly[16] — that the empirical evidence can only legitimately be taken as evidence for the claim that energy is conserved in the *whole* of the closed system, and not as evidence for the claim that energy is conserved in the physical sub-system of the closed system.

Thus the critic's hopes are quashed. If the mind is not only inside the closed system but inside the conservative system as well, as Montero argues

---

[16] Montero notes that there are two distinct interpretations of CoE: one 'restricted' version, where energy is said to be conserved among *the physical components* of the closed system (she calls this 'the principle of physical conservation'), and one unrestricted version, where no such qualification is made. She then goes on to claim that "while physics gives us reason to believe the [unrestricted version], it does not seem to give us reason to believe [physical conservation]" (p. 4).

By way of backing this claim up, Montero claims that restricted version cannot be accepted "without question" while the unrestricted version can, because the former is "a philosophical principle rather than a law of physics" (pp. 5-6). But that is not relevant; what the philosopher is in a position to point out is that any evidence we have for CoE is only evidence for the restricted version, and not for the unrestricted one.

Montero considers this possibility: "One might argue that as all close observations that support the conservation of energy have been of entirely physical systems, all of our evidence for the conservation of energy *is also* evidence for [the restricted version]" (p. 6, emphasis mine). She goes on to object that "when evidence for a theory is consistent with another theory ... [it does not follow that] the other theory is justified by this evidence" (p. 6).

That is getting things the wrong way around. Since the experimental evidence in favour of CoE indeed has resulted from observations of entirely physical systems, we must insist that it is the *generalisation* to the *un*restricted version of CoE that is unwarranted by our observations. The evidence we have for CoE is *only* evidence for the restricted version; it is *not* evidence for the unrestricted version.

It is true, of course, as Montero also points out, that if an experiment with mind-body interaction showed that energy was not conserved among the physical components of the system, this would provide a counter-example to the restricted version of CoE, and not to the unrestricted version. One could still, in that case, retain the unrestricted version of CoE by stipulating that energy was conserved in the entire system, just not among the physical parts. That, however, does not show that the unrestricted version of CoE is what is supported by our present evidence; it just shows that CoE is not as well established as many might have thought that it was, since it has not been tested in the full range of circumstances. CoE, being a thesis for which empirical evidence is adduced, seems certain to be a thesis about the physical part of a system.

The only way to resist this conclusion would be to insist that the class of the physical components of a system is such that it cannot figure in a law whose instances confirm it (cf. e.g. Quine 1969). That avenue is not open, however. What is at issue here is charitable reconstruction of an argument, such that a proponent of the incompatibility claim could be expected to accept it, and the claim that the class of physical objects is in any way like the class of *grue* objects is certain to be one which a critic would reject.

that we must take it to be, no valid argument can be mounted to show that interactionism is incompatible with CoE without denying the possibility that the mind could have energy of its own. If it did, its participation in causal transactions could still conserve energy.

It is this that leads Montero to conclude that the 'real' argument for physicalism has as its premises the claims that mind and body interact causally, that causation involves transference of energy and that anything with energy is physical. That is a valid argument for the conclusion that the mind is physical, but CoE is not in it.

4.4.4 RELEVANCE REASSERTED

A deep problem with Montero's strategy is that it fails to adequately account for the central intuitions of the critic, namely that that *the mind lies outside a conservative system* and that *the mind's interaction with the conservative system would negate conservation. Those* are the intuitions that almost certainly drive the critic, but Montero's reconstruction fails to address them.

Once the notion of closure in CoE is understood as e-closure (and not as affective closure) it becomes possible to reconstruct a different argument on behalf of the critic, one in which both the crucial intuitions form part. Thus the Alternative Argument for Physicalism (AAP) reads:

1* Energy is conserved in any e-closed system (assumption).
2* The physical world is an e-closed system (assumption).
3* Thus: Energy is conserved in the physical world (from 1* and 2*).

4* The mind is not a part of the physical world (dualist claim).
5* The mind interacts with the body by bringing about such local physical changes in the brain as have downstream behavioural consequences (assumption).
6* Local physical changes involve a change in the energy level in the local system (assumption).
7* Thus: Energy is not conserved in the physical world (by 4*, 5* and 6*)

Thus, when conjoined with the assumptions in 1*, 2*, 5* and 6*, the dualist claim in 4* leads to a contradiction. So the dualist claim is false, and we can conclude:

8* Thus: The mind is a part of the physical world

1* is CoE on what I have already argued is a natural understanding of what a closed system is.

2* is certainly a substantive assumption. However, it seems to be an assumption the critic would be happy to make, and unlike the claim that the physical world is *affectively* closed, the claim that the physical world is *e-closed* does *not* beg the question against interactionist dualism (the transfer theory of causation is by no means an integral part of interactionist dualism) so it is in that sense unproblematic.

Note that 2* is implied by the claim that "[a]nything with energy is physical", which was 3' in Montero's 'real' argument for physicalism. If anything with energy is physical, no non-physical entity would *have* energy, so energy could never flow from a non-physical entity to a physical one. So there could not be flow of energy across the border between the physical and the non-physical, so the physical world would be e-closed.

It is perhaps surprising that the argument I present contains a claim implied by one of the premises in Montero's 'real' argument, given that I have criticised her reconstruction of that argument above. There are, however, very significant differences between RAP and the argument presented here. While 2* is implied by 3' it does not imply it, for it is compatible with 2*, though not with 3', that there be non-physical entities with energy, so long as they do not transfer energy to, nor receive energy from, physical entities. So 2* is a weaker claim than 3', and a much less problematic one. Another difference, and an important one, is that the alternative argument presented here does *not* assume

the transference theory of causation, and so allots the critic a significantly smaller theoretical burden.

From 1* and 2*, 3* follows.[17] 4* is the dualist claim. 5* is the claim that mind and body interacts in what on an interactionist dualist account is surely among the most plausible formulations of it, so, in arguing against interactionist dualism, it is unproblematic for the critic to assume 5*. Implicit in the formulation of 5* is also the assumption that the mind, in causing local physical changes in the brain with downstream behavioural consequences, does not *also* cause other changes in the system, such that those extra changes 'compensate' for the change in the level of energy in the brain that would otherwise have occurred. (Perhaps this could be made explicit by appending a phrase such as "and in the process it causes no other changes".) This is not an unreasonable assumption here; the *minimal* interactionist claim is that the mind causes changes in the brain that cause changes in behaviour, not that it also causes additional changes in the brain to, as it were, 'disguise' its influence, almost so as to take on a physical appearance.[18]

6* is a very natural assumption about what would happen to a physical system if a change like that which interactionists claim take place in the brain were to occur (and no other, 'compensatory' changes occurred). One might object to 6* by claiming that the mind could cause changes in the brain by *reconfiguring* the brain in such a way that its level of energy is *not* changed.[19] But again, that is very far from the minimal interactionist claim. The set of reconfigurations of the brain such that the energy level would not be changed by an implementation of that reconfiguration is certain to be vanishingly small compared to the set of changes such that an implementation of that reconfiguration *would* cause a change in the energy level of the brain, so

[17] Montero denies that 3* is supported by evidence, however see n. 16 for reasons to reject that claim.

[18] See also n. 14.

[19] See e.g. (Campbell 1970/1980, pp. 52-53), (Broad 1925, pp. 103-09).

assuming that this is the way that the mind interacts with the brain is laying a very heavy theoretical burden on the interactionist. So 6* is unproblematic for the critic here.

What I claimed are the critic's central intuitions — that that the mind lies outside a conservative system and that the mind's interaction with the conservative system would negate conservation — both play a part in the Alternative Argument for Physicalism. By 1* and 2* we have 3*, which claims that energy is conserved in the physical world. Since 4* states that the mind is not a part of the physical world we have that the mind lies outside a conservative system. The conjunction of 5* and 6* implies that the mind's interaction with the brain would violate conservation.

Thus the above alternative argument for physicalism from CoE seems to be a much more reasonable reconstruction of what lies behind the belief that interactionism is incompatible with CoE. While I will go on to argue that effective defences against the incompatibility claim anchored in CoE are forthcoming, it is clear that Montero's case for the outright irrelevance of CoE to interactionist dualism fails.

4.5 Taxonomy of the Defences

In what follows, various *defences* against the incompatibility claim anchored in CoE are evaluated. As we shall see, this leads to a negative answer to the question of whether the alternative argument previously reconstructed is evidence of a real conflict between interactionism and CoE. First, however, we stop briefly to consider how the various possible defences against the incompatibility claim share features that facilitate their ordering into groups.

4.5.1 A TRIPARTITE TAXONOMY

In his 'Mind-Body Interaction and the Conservation of Energy' Larmer (1986) considers various defences the interactionist could mount, and suggests the following grouping of them:[20]

First we have the *compensation* defences, which are arguments to the effect that the energy the mind creates by moving a particle is compensated for by a *disappearance* of the same amount of energy, so that its effect fails to bring about a violation of CoE. Secondly we have what we might call *approximation* defences, which are arguments claiming that the creation of energy by the mind would not violate the principle after all, since CoE is only statistically — understood as *approximately* — valid. Thirdly we have defences that argue that the formulation of the principle employed by the critics of interactionism begs the question against that theory (p. 278).

4.5.2 AN ALTERNATIVE TAXONOMY

An alternative strategy for dividing up the available defences against the incompatibility claim anchored in CoE is suggested by the discussion above: one could classify the defences according to which of the premises in the Alternative Argument for Physicalism it denies. An obvious disadvantage of such a taxonomy is that it hinges on the acceptance of AAP as a good reconstruction of the argument the critic can reasonably be expected to put forward. On the other hand, this taxonomy has the significant advantage of making it very clear how the defences work, and how they relate to one another. Below I characterise the defences both in terms of Larmer's taxonomy and according to the premise in AAP which is denied.

[20] The groups are his, the names are my suggestion.

4.6 Popper's Defence

We start our survey of defences against the incompatibility claim by considering a suggestion put forward by Popper in *The Self and Its Brain* (1977):

> One possibility that would suit us extremely well would be that the law of the conservation of energy would turn out to be valid only statistically. If this is the case, it might be that we have to wait for a physical fluctuation of energy before World 2 [the mental world] can act on World 1 [the physical world], and the time-span in which we prepare for the "free-will movement of the finger" may easily be long enough to allow for such fluctuations to occur (p. 541).

4.6.1 CONTEXT

Popper's work is infrequently referred to in the (admittedly scarce) literature on the *prima facie* violation of CoE by interactionism. That is a little surprising. The book is itself an argument for interactionism, so it would seem to be a reasonable port of call for anyone seeking a defence against such an influential attack on interactionism. The reason, one suspects, is that no *single* solution to the apparent problem is really proposed here. Instead, what might be called a 'family' of solution-proposals is presented, none of which is by itself defended vigorously or developed in great detail.

To understand why Popper takes this problem — now often seen as a serious challenge to interactionism — as lightly as he does, it is useful to consider some of the comments that surround the solutions he proposes. Central to Popper's perspective is the view that at various stages during the recent scientific advancement, the physical world as then conceived has been discovered to be '*open*' to influence so different from what had until then been countenanced that it is best characterised as openness to a *new world*. The physical world as governed by Newtonian mechanics was e.g. discovered to be open to "the world of electricity" (p. 542). Later, Popper argues, followed a reduction of mechanics to electromagnetism so successful that it nearly led to

belief in 'electrical monism'; to the belief, that is, that the world could be completely described in terms of the electro-magnetic forces. But again, the worlds of mechanics and electromagnetism were found to be open *inter alia* to nuclear forces. This trend shows, Popper suggests, that "modern physics is pluralistic"(p. 542). CoE has had to be generalised several times before, so the fact that interactionism *appears* to violate the law should not, he argues, worry us too much. There will probably be a way to reconcile the conflict.[21]

Opponents of interactionism sometimes describe the *prima facie* conflict between CoE and mind-body interaction as an irreconcilable conflict between a well-founded and very nearly complete science on the one hand and a set of beliefs grounded in little more than wishful thinking on the other (cf. quotes in Montero forthcoming, p. 2). Popper's point of view stands in sharp contrast to such a position. On his view we have, on the one hand, extensive and reliable evidence for interaction between the non-physical mind and the brain; and on the other we have a science that, while it may be well substantiated, nevertheless has been subject to the kind of change that acceptance of interactionism would demand of it not just once, but several times before. The accuracy of this description can be debated, but seeing that Popper held this view allows us a better understanding of his approach to the issue. Seen with Popper's eyes, it is not natural to see the *prima facie* conflict as a crucial problem for the interactionist position, possibly dooming it to failure. It is more natural to see CoE as a hurdle to overcome; a tall one, perhaps, but one we can be reasonably confident *will* be overcome. His proposals are most charitably regarded as interesting lines of thought to pursue farther, not as definitive and well worked out solutions, ready to stand up to rigorous scrutiny.

[21] See *The Self and Its Brain* (Dialogue X, especially pp. 539 - 543).

4.6.2 INTERPRETATION

What does it mean for CoE to be is a "valid only statistically"? The suggestion can be understood in at least three different ways.

4.6.2.1 PROPENSITIES

The first two ways of understanding the suggestion rely on realism about *propensities*. What is a propensity? Consider the example of a fair coin. A realist about propensities would say that a fair coin has the *propensity* to land on heads and tails equally often when tossed. If you toss a fair coin one thousand times, it is very unlikely that you will get heads every time, and it is just as unlikely that you will get tails every time. During a much longer run of tosses, on the other hand, it is likely that there will be some long stretches with only heads, and some long stretches with only tails. Such stretches would not skew the overall frequency much in very long runs. Thus, in very long runs it is very likely that the actual *frequency* of heads — the proportion of heads to total tosses — will be *very close to* 0.5. For our purposes we get enough precision by saying that what it means for a fair coin to have a propensity to land equally often on heads as on tails, is that a coin tossing system has a *tendency* or *disposition* to produce a frequency close to 0.5 for both heads and tails. (It is quite unlikely that the frequency of heads will be *exactly* 0.5, and more unlikely the longer the run is.)[22]

Imagine a plotted graphic representation of how the frequency of heads and tails change over time, with dots representing the frequency at any given time, and imagine a straight line representing the 0.5 mark. Then it is very likely that there will be many more dots close to the line than further away from it. The representation will probably look like a thick line, gradually fading away on either side.

Thus, for CoE to be 'valid only statistically' might mean that physical systems *have the propensity to maintain the same level of energy over time*. We

[22] This is so for an even number of tosses in the run; the likelihood of the frequency being exactly 0.5 with an odd number of tosses is zero, obviously.

should expect the results of measurements of the energy level of e-closed systems to remain close to some particular value, just as the frequency of a fair coin in a long run is expected be close to 0.5 for heads and 0.5 for tails. Although it would be very *unlikely* that the energy level of a system turn out to be *exactly* the same after a change in the system as it was before, it would also be very unlikely that it would vary very much from what it was before. Imagining a graphic representation of the measurement of the energy level of an e-closed system, we could draw a straight horizontal line through the graph, such that it would be very unlikely for the results to fall very far away from that line.

Questions can be raised concerning how much substantiveness CoE retains when it is conceived of in this way. What constitutes a counterexample will vary with how much variation is accepted, and for each counterexample it would be possible to reformulate the law slightly less stringently to avoid the counterexample. There is, however, another question that requires more immediate attention, for the understanding of CoE is still not clear enough to allow us to ascertain whether interactionist dualism is compatible with it.

4.6.2.2 FIRST INTERPRETATION

On the one hand, CoE could be taken to be a claim just about the sorts of *results* we will get when measuring the energy level of e-closed systems. CoE would then amount to the claim that e-closed systems will give a certain characteristic output; the energy level of the system will not vary more than some specified amount.

How would this understanding of CoE be of help to the interactionist? It may be that the mind's influence on the brain changes the energy level of the brain very little. If so, then *after* the mind's influence the energy level of the brain may still be within the range that CoE allows; it might still be in a range of values such that it would have been within that range anyway, had the mind

not exerted any influence on it. Thus, if CoE is a law about the *results* we will get from measurements of the energy levels of e-closed systems, then interactionism may be compatible with CoE.

This interpretation does not imply that the brain-state itself would be the same as if the mind had not exerted any influence; the interactionist claim is that the mind *makes a difference* to the physical world. However, the difference that the mind makes in the brain may cause only very, very small changes in the energy level of the brain, and if CoE allows the energy levels of a system to stay within a range of some width, interactionism may well be compatible with it.

4.6.2.3 SECOND INTERPRETATION

On the other hand, the result that the energy level will not vary more than some specified amount could be taken as a *consequence* of CoE, and not as itself constituting the *content* of CoE. Then CoE itself could be understood as *whatever generated* this output. For example, CoE might be the law that "*for any t, given the energy level in the system at* t_i*, the energy level in the system at* t_{i+1} *is the result of statistical propensity x*", for some specification of x.[23]

On that reading, mind–brain interaction is *incompatible* with CoE. If the mind interacted with the brain, the actual level of energy in the system would *not* be the result of the statistical propensity x; it would be the result of the statistical propensity x *as modified by the mind's activity*.[24] It might still be the case, of course, that the energy level of the brain would be inside the range the statistical propensity predicted — in which case the violation of CoE would not be detectable — but that does not change the fact that interactionism is

[23] For example: "The energy level of the system has a 45% likelihood of rising up to 0.5%; a 4% likelihood of rising 0.5% to 1%; a 0.9% likelihood of rising 1% to 2%; a 0.05% likelihood of rising 2% or more; a 45% likelihood of falling up to 0.5%, a 4% likelihood of falling 0.5% to 1%; a 0.9% likelihood of falling 1% to 2%; a 0.05% likelihood of falling 2% or more, and a 0.1% likelihood of not changing measurably at all."

[24] This assumes that the mind's activity is not already a part of x, an assumption we must make if we are to give the incompatibility claim a fair hearing.

incompatible with CoE, on this reading of it. Since Popper presents his suggestion as a defence against the incompatibility claim anchored in CoE, we can discount this interpretation.

4.6.2.4 THIRD INTERPRETATION

Perhaps what Popper is suggesting here is not that the principle concerns only an *approximately* conserved quantity of energy. Perhaps he is suggesting that the quantity of energy is *exactly* conserved over time, while nevertheless allowing for deviances to occur at any given instance. The mind might have to wait for a deviance in one direction — a slight fall in the total energy, say — which it can then 'exploit' by causing a particle to move. The increase in total energy that this movement may otherwise have caused would then not obtain, given the opposite fluctuation that was already manifest. That, it seems, may well be what Popper had in mind.[25]

4.6.3 EVALUATION

4.6.3.1 FIRST INTERPRETATION

The first understanding of Popper seems to be what Larmer has in mind in his article 'Mind-Body Interaction and the Conservation of Energy' (1986). About Popper's proposal that CoE may be valid only statistically, he writes:

> Presumably, if this were the case then the action of the mind upon the body could be viewed as relatively minor fluctuations which do not really violate the Principle of the Conservation of Energy, provided we remember that this principle is really a statistical one, not absolute (p. 279).

Larmer then goes on to argue that the number of interactions posited by interactionism is so large that even if each instance caused only a very small

[25] This interpretation is corroborated by a brief comment in the last dialogue: About "...the possibility that the first law of thermodynamics may ... be valid only statistically" he writes: "[a]ny violation in one direction may be statistically levelled out by one in the opposite direction" (pp. 564-65).

deviance from the conserved quantity of energy, the accumulated effect of these numerous interactions would have a large effect: "Clearly, although the net effect of any one mind will be relatively small, and although there will be some kind of averaging effect, the net fluctuation of energy can be quite great" (p. 279, n. 5).

This criticism is not convincing. It is clear from his choice of words ("... any one mind") that Larmer is not considering a small system under the assumption that it is e-closed. He is, rather, considering the possible effect mind–brain interactions can have on the total energy level of the one system we can be certain is e-closed: that of all entities with energy. In that perspective, however, it seems overwhelmingly likely that the amount of energy in all the physical systems that we think may be under the influence of minds — the energy of all the *brains* put together, that is — is vanishingly *small*; unless, perhaps, brains are very much more numerous or much larger than it looks like they are, or there are other large or numerous 'mind-sensitive' systems. But if that is true, it is surely up for grabs whether the impact of mind–brain interaction would be large enough to amount to a violation of CoE, on this understanding of it. There does not seem to be any reason to accept Larmer's claim that "it is very improbable that energy will be even approximately conserved"(p. 279), especially since he provides no specification of what might and might not count as an 'approximate conservation'. Absent such a specification we just cannot tell, one way or the other.[26]

Indeed, this lack of specificity as to how large a deviation would have to be before it would constitute a counterexample to CoE can be seen both as a strength and as a weakness of this interpretation. There will be a number of ways in which CoE could be made precise, and it not clear how one could decide which level of stringency is appropriate. This seems to rob CoE of

[26] Furthermore, the 'averaging effect' may well be rather more significant than Larmer recognises; see section 4.7.1 below.

determinate content, to some extent. However, if this really turns out to be the correct way to interpret CoE, as Popper suggests that it might, it seems very unlikely that an argument from CoE to the falsity of interactionism could succeed.[27]

Above it was argued that the conjunction of 5* — the claim that the mind interacts with the body by bringing about local physical changes in the brain, which have downstream behavioural consequences — and 6* — the claim that local physical changes involve a change in the energy level in the local system — in AAP implies that the mind's interaction with the brain would violate conservation. The discussion here suggests an interpretation of CoE such that this is not the case. CoE was given in 1* as the claim that energy is conserved in any e-closed system. To form part of a valid argument it should strictly speaking have read that energy is *exactly* conserved in any e-closed system. On the first interpretation of what it means for CoE to be valid only statistically, energy will only be *approximately* conserved in all e-closed systems. This would block the inference to 7*, and interactionism would be compatible with CoE.

Thus this first interpretation of CoE yields what seems to be an effective defence against the incompatibility claim. When the 'simple levelling out view' is considered in section 4.7.1 below, we shall have more to say about the efficacy of this defence.

4.6.3.2 THIRD INTERPRETATION

We have already seen that the second interpretation does not yield a defence against the incompatibility claim, so we can leave that to one side. Does the third? Consider what the universe would have to be like if CoE were to hold true statistically, if this is understood as exactly preserving an average while still allowing for fluctuations from moment to moment. At any time t, we could not be assured that the total amount of energy in the universe would be the

[27] Popper refers to Schrödinger (1952) for reasons to think so (p. 541).

same as the total amount of energy at a different time. So, for any n, it is possible that $e(t_n) \neq e(t_{n+1})$, where e(x) is the function from times to total amounts of energy in the universe. Given the stipulation that the *average* total level of energy does not change over time, the variations or spikes that occur at individual moments must always, over time, be matched by variations or spikes in the opposite direction, either one by one or cumulatively; otherwise the total energy of the universe would either rise or decline over time. That would mean, of course, that there are no 'surplus' variations left for the mind to utilise to 'innocently' exert its influence. Then this interpretation does not yield a defence either, *unless* it is supposed that more such variations or spikes occur in a universe where there are minds present than in one where there are no minds. Supposing *that* is to suppose that the effects of the mind on the total amount of energy in the universe is compensated for.

In a universe which worked the way this interpretation suggests, there would be small fluctuations in the total amount of energy in the universe, fluctuations which the mind takes advantage of to interact with the brain. Suppose that in such a universe, the minds suddenly ceased to exist. If the small variations or spikes in the total energy of the universe were such that the mind could utilise them to interact with the brain without violating CoE, the result of suddenly subtracting the influence of all the minds must be that the average total energy level of the universe would cease to be stable, since the spikes the minds previously used would now go unmatched by the minds' activity. Thus, if a nuclear war broke out, and all the minds in the universe ceased to exist, CoE would be violated. That result is untenable. Interactionism should not defend itself against the charge that its theory entails the violation of CoE by making mind–body interaction necessary in order to *avoid* such a violation.[28]

[28] The argument could also be recast in terms of individual minds, with the improbable result that the energy level of the universe responds to the emergence and destruction of individual minds.

To avoid this conclusion one might suppose that the relevant spikes would cease to occur after (in response to?) the destruction of all minds. That must surely mean that the spikes came into existence in the first place as a result of (for the benefit of? to accommodate?) the emergence of the minds that were in need of the spikes to 'innocently' exert their influence on the brain. Again, this is just to suppose that the mind's influence on the brain is compensated for by an opposing counter-reaction elsewhere in the system.

Thus, this third interpretation of Popper's suggestion yields a defence against the incompatibility claim only if it is seen as a 'compensation view'. Such defences are considered in more detail in what follows.

4.7 Compensation Defences

We have seen that a proponent of Popper's defence on the third interpretation is forced to admit that the fluctuations that allow the mind to 'innocently' exert its influence came into existence in the first place in response to, or to allow for, the mind's interaction with the physical world. The other option for the proponent is to accept the rather more implausible conclusion that the existence of minds that affect the physical world is necessary for CoE to hold true.

4.7.1 TWO 'LEVELLING OUT' VIEWS

4.7.1.1 THE COMPLEX LEVELLING OUT VIEW

That unwanted result can also be avoided in a different manner. One might suppose that the mind is just as likely to utilise negative 'spikes' — small *decreases* in the total energy level of the universe — as it is to utilise positive spikes — small *increases* in the total energy level of the universe. If this were the case, then — supposing no systematic difference between the amplitude of the positive and the negative spikes — the sudden subtraction of all minds would neither cause a gradual increase nor a gradual decrease in the total energy of the universe after all. The subtracted effect would be likely to be roughly equal

in either direction; its net influence over time would be very close to zero. The view that the mind affects the physical world without violating CoE by waiting for positive and negative spikes before exerting its influence will be dubbed *the complex levelling out view*.

4.7.1.2 THE SIMPLE LEVELLING OUT VIEW

There is a simpler and therefore preferable candidate view in the vicinity. Clearly we need not stipulate that the mind has to wait for small spikes in the energy level of the universe before it interacts with the brain. Instead we could just accept that *each individual* interaction between the non-material mind and the brain is prone to cause a slight increase or decrease in the total energy level of the universe.[29] We stipulate, however, that the combined net effect of all the interactions would, at any given time, be likely to be close to zero.

To achieve this result we need just two assumptions, and both are plausible. First we must suppose that an immaterial mind's interaction with the brain may not only cause an increase in the total energy of the brain, it may also cause a *decrease*. This is plausible, since it certainly seems possible for the mind to exert influence on the brain not only by *moving* a particle or changing its trajectory, but also by slowing it down. The slowing down of a particle without any concomitant increase of energy in any form elsewhere in the system — such as what we would see when a rock falls through air — would cause a net drop in the total energy of the system of which it were a part.

Secondly we assume that such increases and decreases are equally likely results from the mind's activity. This second assumption needs no independent motivation given the first; the burden of proof would be on whoever might wish to argue that this was not so, to instead give reasons that we should expect the mind to influence the brain in one way more often than in the other.

---

[29] But it is not guaranteed to, since it might happen that another interaction–event elsewhere caused an opposite reaction of exactly the same proportions at exactly the same time, in which case the result would be no net difference at all.

This view will be called *the simple levelling out view*, for obvious reasons. Unlike its predecessor, this view does not stipulate that the mind monitors the total energy level of the universe for a 'spike' it can use to 'innocently' exert its influence; it stipulates only that the mind causes small deviances in the energy level, but that it is equally likely to cause them in both directions, resulting in a net effect very close to zero.

4.7.1.3 THE INTUITION BEHIND THE LEVELLING OUT VIEWS

Characterised in terms of the alternative argument for physicalism above, we get closest to really describing the motivating intuition behind the levelling out views if we say that they work by rejecting premise 6*, the claim that local physical changes involve a change in the energy level in the local system. The complex view posits that the mind uses 'superfluous' fluctuations in the energy level of the universe, so that, in each individual case the mind could be said to *prevent* a change in the energy level of the brain that would otherwise have occurred naturally. The added stipulation is that the mind has equal propensity to utilise positive spikes as it has to utilise negative ones, the purpose of this stipulation being to ensure that the energy level in the brain does not change much over time.

What the simple levelling out view suggests is that 6* fails *over time*. Since the mind is just as likely to cause small increases in the energy level of the brain as it is to cause small decreases, the thought is that the various increases and decreases will average out over time, resulting in no net effect, no 'disturbance' of the energy level of the universe over time. So both the levelling out views can be said to be motivated by a rejection of premise 6*.

4.7.1.4 WHAT THE VIEWS ENTAIL

It is important to note, however, that stipulating that the mind is as likely to use positive spikes as it is to use negative ones, as the complex levelling out view does, *still* allows for comparatively long stretches in which a particular mind

only uses positive spikes. Similarly, the fact that the simple levelling out view posits that the mind is as likely to cause a small increase as it is to cause a small decrease in the energy level of the universe is no guarantee against there being comparatively long stretches where a particular mind causes only increases.

Both the levelling out views stipulate that events of two different kinds are equally likely to obtain. There is thus an obvious parallel between this feature of the views and the example discussed in section 4.6.2.1 above: a fair coin being tossed. To say that a coin is fair does not guarantee that there will be no long stretches of only heads or only tails; and to say that the mind is just as likely to cause a small increase in the total energy level as it is to cause a small decrease does not guarantee that there will not be long stretches where a mind, in its interactions with the brain, only causes increases in the energy level.

4.7.1.5 THE EFFICACY OF THE LEVELLING OUT VIEWS

This makes the question of whether the levelling out views are effective defences against the incompatibility claim anchored in CoE a much more complex one to answer.

There are many minds, and on the interactionist view there are many interactions between each mind and its brain. Given that each mind is stipulated to have an equal propensity to 'add' as to 'subtract' a little energy, the existence of many minds that frequently interact with the physical world already makes it unlikely that there should be a detectable difference in the *overall* energy level of the system of all entities with energy. Add to that the considerations about the proliferation and sizes of mind–sensitive systems above, and the chances of a measurable difference in the overall energy level seem abysmal indeed, even on the quite unlikely assumption that we should ever be in the position to carry out such measurements.

It is a different matter, perhaps, with smaller *sub-systems* of the system of all entities with energy, on the assumption that some of them are, or can be

made to be, e-closed. Would we, if interactionism is true, be able to measure disturbances in the energy level of a single, perhaps furiously thinking, brain? We cannot know. The answer depends on a number of questions we lack the answers to, such as how frequent interactions are, and how big the effect on the system is each time (relative to effects we can measure), and so on. If the interactions are extremely frequent, there will be a higher likelihood of long stretches of accumulated effects in one direction; if the frequency is lower, it may turn out to be extremely unlikely that we will ever get a 'piling up' of (say) 'additions' to the energy level in the brain large enough to measure, in the lifetime of a person.

It is clear that accepting the simple levelling out view comes at the price of rejecting CoE *if* it is taken to be a law about *each and every* energy transfer, or each individual interaction such that the energy level of the interacting entities is different after the interaction from what it was before. But it is not clear that that CoE read in *that* way is a law we have good evidence for.

That each energy transfer or transaction conserves energy is at most a stipulation of physical theory, and it is not reasonable to demand of an interactionist account that it should be compatible with all features of physical theory. It is the *data*, and not the theory, that requires explanation. Thus, if the interactionist can come up with an explanation that both accounts for the available data and has the same (or very nearly the same) utility for scientific investigation, the objection that it conflicts with certain features of physical theory carries little force. Insisting on compatibility with physical theory under these conditions would be to beg the question.

In the discussion of the first interpretation of Popper's suggested defence against the incompatibility claim, an interpretation of CoE was suggested which would guarantee its compatibility with the simple levelling out view: e-closed systems will give characteristic output, in which the energy level of the system varies very little. Whether CoE as thus interpreted has adequate utility for

scientific investigations remains to be established, and it is not something which could easily be established conclusively.

If CoE *does* have equal or near-equal utility for science when read as the claim that energy is *very nearly* conserved in e-closed systems, however, it is clear that the combination of Popper's view (on the first interpretation) and the simple levelling out view yields a very strong defence against the incompatibility claim. Given that the mind could influence the brain in ways that would 'subtract' a little energy from it just as easily as it could influence it in ways that would 'add' a little energy, and given that we have no reason to think that one of these should be more common than the other, this combined solution does not only withstand the incompatibility claim. It also carries light theoretical baggage, and it may, as we have seen, even turn out to be open to empirical testing. With sufficiently fine-tuned equipment it is conceivable that the (presumably minute) changes caused by mind–brain interaction could be detected.

It was mentioned above that Popper, rather than presenting a single, well worked out solution to the incompatibility claim anchored in CoE, presents a *family* of defences, the intended cumulative effect of which must have been to bolster his view that the incompatibility claim presents a hurdle, but not an insurmountable challenge, to interactionism. The other members of the family are not without interest, but they are unlikely to produce an effective defence against the incompatibility claim, and in particular a theoretically attractive one.[30] Having seen how Popper's suggestion can, on one interpretation of it, be

[30] In one suggestion he argues that "the fact that a vessel or a vehicle *can* be steered from the inside without violating any physical law" so long as it comprises also a source of energy, and so long as it compensates for a change in direction by "pushing some mass ... in the opposite direction" may give new life to Descartes' solution to the problem, which stipulates that that the mind's influence is worked by changing the directions of particles already in motion, not by actually causing them to move (p. 180). (Leibniz argued convincingly against Descartes' proposed solution, by replacing the conservation of motion with conservation of linear momentum; see for instance Papineau (2002, p. 236).) In another defence Popper argues that quantum indeterminacy, in addition to causing the genetic mutations on which natural

combined with the simple levelling out view to yield a strong defence against the incompatibility claim, we now move on to the next class of defences.

4.8 Defences that Allege Question-Begging

Defences in this group are by Larmer characterised as

> [a]ttempts to argue that the Principle of Conservation of Energy is merely the defining-postulate of a wholly closed physical system and therefore any attempt to rule out the theory that an immaterial mind acts upon the body on the grounds that such a theory violates the Principle of the Conservation of Energy is only to beg the question (1986, p. 278).

Larmer claims that this objection is based on a misunderstanding. Those who argue that mind-body interactionism cannot be true because it would violate CoE are according to him not begging the question, but only relying on an "initial presumption" in favour of CoE (p. 280). The presumption relies, he argues, on a balance of probability evaluation, comparing the "enormous body of experimental evidence" that CoE has in its favour, to the "not ... as great an amount of evidence in favour of mind-body interaction" (p. 281).

It is not at all clear that a balance of probability equation is what is driving the critic here.[31] Setting that question to one side, however, it appears

---

selection works, might similarly also allow the brain to bring about a "range of possibilities", from which the mind then selects: "In the brain there may at first arise purely probabilistic or chaotic changes, and some of these fluctuations may be purposefully selected ... in a way similar to that in which natural selection quasi-purposefully selects mutations" (pp. 540-41).

[31] It is also not clear that such an equation would tip obviously in favour of CoE. The commonsense thesis that how things *feel* matters for what we *do* is supported by a truly *massive* and ever-growing body of evidence: the evidence from our own daily experience. Where sheer size is concerned it is surely second to none. Consider again what someone might answer if asked why they did something. As was argued in chapter one, their ultimate answer is likely to be that they did it because it (or some downstream consequence) *feels* good somehow. It is assumed, in this work, that dualism is true about phenomenal experience. On that assumption, the position that can account for this enormous body of evidence is interactionist dualism. Thus, if Larmer's interpretation of this defence is correct, the claim that there is a balance-of-probability evaluation at play which *clearly* leans toward CoE (and away from interactionism) does not seem to stand up to scrutiny.

that Larmer misinterprets the strategy of those who attempt to resist the critic. As the discussion has already suggested, the crucial allegation this defence makes is that the principle of conservation of energy is formulated (or interpreted) in a more restrictive way than is necessary for the continued progress of science, or than the evidence dictates, or both, and that the *choice* of the more restrictive formulation is biased against mind–body interactionism. The allegation is, in other words, that formulation or interpretation of CoE makes interactionism incompatible with the principle when a formulation without this result is available, equally well supported by the evidence and equally well suited to allow for progress in science. If this is true, then, by opting for a more restrictive version instead of a less restrictive one, the critic can properly be said to be begging the question after all. Assuming that there really is an alternative formulation that is equally well supported by the evidence and that is not incompatible with interactionism, the question of whether it is CoE or interactionism that is better supported by evidence does not even come into play.

4.8.1 QUESTION-BEGGING I: AVERILL AND KEATING

Just such an argument is put forth in 'Does Interactionism Violate a Law of Classical Physics?' (Averill and Keating 1981). They argue that "[i]nteractionists and their opponents have thought that ... [interactionism is incompatible with certain laws of physics] ... because they have used statements of the laws of physics that are stronger than is necessary to develop physics and which are

---

The outcome of an evaluation of which thesis is better supported by evidence would obviously depend to a very large extent on which types of evidence were considered admissible. Someone who wished to argue, however, that the evidence we have for claiming that phenomenal experience plays an important role in motivating our actions is *not of the right type* would owe us an account of why this is so, and the account would have to be one that does not beg the question against interactionism. One might also think that the balance of probability might work in the manner Larmer envisages if cast in terms of *types* of evidence. Our phenomenal experience is just *one* type of evidence for interactionism, it might be argued, but CoE is corroborated by many *different* types of evidence. I doubt that this claim could be substantiated without begging the question, but I do not pursue this question here.

question-begging against interactionism" (p. 102). This is a good example of this type of defence, so we shall consider it in a little more detail.

4.8.1.1 EXPOSITION

The authors frame their discussion in terms of (1) the law of conservation of linear-momentum: "If the total external force is zero, the total linear-momentum is conserved".[32] They grant (2) that: "[i]f the mind exerts a force F on the brain .... [that moves a particle in the brain] then the total linear-momentum of the brain is changed due to F" (p. 103). However, they reject the move to (3), the claim that: "[i]f the total linear-momentum of the brain is changed, then 'some net external physical force' affects the brain".[33]

Thus they argue that since (3) is not entailed by (1) — "[t]his law holds for all kinds of forces ... and not just ... '*physical* forces' " — an interactionist can both accept the law of conservation of linear-momentum and reject that (3) follows from it (p. 103). The change of the total linear-momentum of the brain as a result of the mind exerting a force on it does not constitute a violation of (1), since in such an instance the antecedent of the law is false; the total external force is *not* zero (p. 103).

The authors are clearly attempting to show that the conclusion is stronger than warranted by the premises. The allegation is that the physicalists are helping themselves to (3) on the basis of the law of conservation of linear-momentum with the aim to argue that, since the mind according to dualism does not exert a *physical* force on the brain, the antecedent of (3) must be false (from the negation of the consequent of (3)). But if the mind exerted a force on the brain that moved a particle in it, then the antecedent of (3) would be *true* (from (2)). So it cannot be that the mind exerts a force on the brain that moves a particle in it, so interactionism is false.

---

[32] (Goldstein 1980, p. 5), quoted in (Averill and Keating 1981, p. 103).

[33] (Averill and Keating 1981, p. 103) The account they are criticising is in (Cornman 1978).

If, however, the move from (1) to (3) is blocked, this argument does not get off the ground. The move from (1) to (3) is justified by demanding that the only forces up for consideration are the physical ones, but this, Averill and Keating argue, begs the question against the interactionist. Thus it is open to the interactionist to hold that "when a force due to the mind acts on a particle in the brain, the sum of that force plus all of the other external forces on the brain is equal to the time rate of change of the total linear-momentum of the brain", a position which is, they argue, "consistent with all the momentum laws of physics" (pp. 103-04).

This defence can be adapted, *mutatis mutandis*, to repel arguments where CoE has replaced the law in (1), simply by assuming that the mind has energy of its own which it transfers to the brain in such a way that an extended version of CoE is upheld.[34] Such a defence would stipulate energy-flow between the mind and the brain. It therefore blocks the argument against interactionist dualism by denying premise 2*, which says that the physical world is an e-closed system.

4.8.1.2 EVALUATION

It is clear, however, that the evidence we have for CoE is evidence only for the claim that energy is conserved in the physical world. It is *not* evidence for an extended version of CoE, also encompassing transactions with mental energy as a participant. A criticism that alleges that critics are begging the question when phrasing their argument in terms of physical entities (or physical forces) therefore seems misdirected.[35] This is the *only* way the critics can legitimately phrase their claim, for a claim about physical entities is all there is evidence for. While it is not impossible that evidence could be found for an extended version of CoE — also encompassing mental energy — it is clear that we do not possess

[34] When I discuss such a defence in what follows I do not mean to imply that Averill and Keating would accept the reformulation of their argument in terms of CoE.
[35] See also n. 16.

such evidence yet. Rejecting 2* by claiming that it must be reformulated in more general terms is not a reasonable criticism. All the evidence we have is only evidence for the restricted claim, not for the unrestricted one.

4.8.2 QUESTION-BEGGING II: REJECTING 1*

We have already seen that a defence against the incompatibility claim anchored in CoE can be built on an interpretation of CoE that takes it to say that energy is *approximately* conserved in all e-closed systems. This may on its own be sufficient to withstand the incompatibility claim, or it may be coupled with the simple levelling out view.[36] It was also argued that if this interpretation of CoE has the same, or very nearly the same, utility for scientific investigation, the insistence that interactionism must be compatible with the feature of physical theory that demands that energy be conserved in each and every transaction would beg the question.

A related approach — foreshadowed at the start of this chapter and discussed at length in the next — is to contest the claim the evidence we have amounts to conclusive evidence for CoE *stated in full generality*; that is, to *accept* 2*, the claim that the physical world is an e-closed system and instead to reject 1*, the claim that energy is conserved in *any* e-closed system.

As it stands, 1* applies *inter alia* to systems that the interactionist believes are under the influence of a mind. However, given that the evidence we have for CoE stems exclusively from systems that are *not* supposed to be under the mind's influence,[37] it is clearly open to the dualist to argue that the assertion that the evidence we have suffices to establish that CoE holds universally —

[36] See sections 4.6.3.1 and 4.7.1.5, above.

[37] Counting of calories taken in through food and measurement of energy expended through heat, excretions and so forth may be thought to constitute evidence for the operation of CoE over systems containing the human body. I think it is fairly clear, however, that such data is nowhere near the level of precision that would be required to amount a credible incompatibility claim against interactionism. For further reasons to reject such evidence, see *The Mind and its Place in Nature* (Broad 1925, p. 106). (Broad accepts such evidence for the sake of argument.)

and in particular also in physical systems influenced by minds — is unconvincing.

Some would no doubt object that, although we lack direct experimental evidence to suggest that energy is conserved in systems including brains, we have extensive *indirect* evidence for this claim, and that this is sufficient to establish the conclusion. After all, all the examples of failure of conservation we have thus far encountered have turned out to be spurious. So, it might be argued, we should expect all e-closed physical systems, including those under the mind's influence, to be conservative.

To this the dualist can retort that the dualist claim that the mind is *unique* in nature is being ignored. If the mind is truly unique it is very far from clear that the evidence we have gathered so far should lead us to *expect* energy to be conserved even in systems that are under the influence of a mind. On the contrary, such systems may very well be the first examples of physical systems that *fail* to conserve energy. Accepting the evidence we do have as conclusive evidence for the claim that CoE applies in full generality would therefore amount to begging the question after all.

It is true that *if* the general formulations (and not the restricted ones) were supported by our evidence Averill and Keating's defence would be effective; by choosing the restrictive version the critic would be begging the question against the interactionist.[38] Then we would have yet another good reason to reject the incompatibility claim anchored in CoE, for the interactionist could deny that the human brain is a closed system in the required sense. But it is the restricted formulations, and not the general ones, that are supported by our evidence.

---

[38] Even if the general formulations were supported by our evidence, Montero's argument would still not represent a charitable reading of the critic, for the critic clearly need not subscribe to the transference theory of causation to mount an argument on the basis of CoE. If the general formulations were those that our evidence supported, my criticism in n. 16 would be unwarranted, however.

It should be noted, also, that there is a certain similarity between the thought Averill and Keating put forth, a thought that seems to motivate parts of Montero's argument and the thought I defend. Averill and Keating claim that the critic is unwarranted in excluding the possibility of mental forces; Montero — wielding the notion of affective closure — asserts that the closed system must include the mind; I claim that the critic is unwarranted in asserting that the (narrowly formulated) laws of physics hold in full generality, inasmuch as they are *prima facie* incompatible with interactionism. In all cases the claim is that there is an unwarranted exclusion of the effects which the interactionist supposes that the mind has on the body by certain formulations of the laws of physics.

There is, however, an important difference between Montero's and Averill and Keating's lines on the one hand, and the defence I present here (and in the next chapter), on the other. Unlike the defence that rejects 2*, my suggested defence, which rejects 1* instead, does *not* rely on stipulating non-physical forces or non-physical energy. That is a significant advantage, for many believe that our concepts of force and energy gain nearly all their content from their applications to physical entities and their places in physical theory, and that the appeal to a mental force or mental energy consequently is a definite non-starter.

Premise 1* of AAP claims that energy is conserved in any e-closed system; the defensive strategy outlined in this section rejects that 1* holds in full generality. As a result, one might suggest the following reformulated version of CoE:

> CoE* If the total external influence *of any kind* on a system is zero, energy is conserved within that system.[39]

[39] Given her notion of closure, this is a formulation with which Montero would presumably agree, however, as I have argued, for the wrong reasons.

Whenever the non-physical mind exerts influence on the physical brain by moving a particle, the total energy of the brain is, in all likelihood, changed. But in those cases the interactionist claims that the total external influence on the system is non-zero, and finds that CoE* is not violated. Given our data, it is entirely plausible that energy is conserved only in physical systems that are not influenced from the outside, but that it is *not* conserved when it *is* influenced from the outside, by, for instance, a mind.[40]

Whether the effect of a mind on a part of the physical is detectable is up for grabs. Whether it would affect the total energy level of the universe over time is up for grabs too, since the effect could very well be evenly distributed between tiny additions to the energy level and tiny subtractions from it. Supposing the latter is effectively to adjoin to the rejection of 1* a rejection of 6* as well.

4.9 Conclusion

We have considered various defences against the incompatibility claim anchored in CoE, and found that several prove effective. Among the compensation views, the *complex* and the *simple* levelling out views are both effective, though the latter is preferable, precisely because of its simplicity. The intuition behind these defences is best described by their denial of 6* in the Alternative Argument for Physicalism, the claim that local physical changes involve a change in the energy level in the local system.

---

[40] See also 'The hidden premiss in the causal argument for physicalism', where Bishop (2006) makes essentially the same point for the totality of physical laws. All physical laws rely on *ceteris paribus* clauses, he argues, and "[a]mong these idealizations and qualifications is that no *non*-physical influences are present" (p. 50). (While I have sympathy for this claim, I reject his claim that the "hidden premiss — that the only efficacious states and causes are physical ones" is "largely indistinguishable from physicalism", since *epiphenomenalist* dualism is, of course, perfectly compatible with that premise (pp. 47, 45).)

We also saw that the combination of the simple levelling out view with Popper's suggested defence under one of its interpretations — that which results in a restatement of CoE as the claim that energy is *very nearly* or *approximately* conserved in e-closed systems — yields a particularly strong defence. By combining that interpretation of CoE with the claims that the mind could just as easily exert its influence on the brain in ways that would 'subtract' a little energy as it could in ways that would 'add' a little energy and that we have no reason to expect one over the other, the incompatibility claim is effectively countered.

In contrast, the attempt to block the incompatibility claim by alleging that the restrictive formulation of CoE (or the law of conservation of linear-momentum) begs the question against the interactionist, was seen to be misguided. Thus, denying 2*, the claim that the physical world is an e-closed system, does not seem to be a promising defensive strategy. However, it *is* true that the critic begs the question against the interactionist by taking the evidence we have for CoE as conclusive evidence for the claim that CoE holds *in full generality*, also in systems under influence of a mind. Thus, a good defence rejects 1*: CoE stated in full generality. The claim is that CoE must be reformulated in terms of external influence *simpliciter*, such that energy is said to be conserved in an e-closed system *only when* that system is under *no* external influence. This strategy has the significant advantage over the suggested strategy of denying 2* of not requiring the postulation of mental energy (or a mental force).

We may conclude, therefore, that there are effective defences against the incompatibility claim anchored in CoE, and that the argument from CoE to the falsity of interactionism is ineffective.

Chapter 5: Other Arguments

5.1 Introduction

In the previous chapters it was argued that monism derives a significant proportion of its support from the elimination of dualist alternatives; that aside from epiphenomenalism and interactionism the dualist alternatives are indeed quickly eliminated; that epiphenomenalist dualism is an unattractive position; that interactionist dualism is, in contrast, a very attractive position; and that it therefore is important to investigate very carefully whether interactionist dualism is a viable position. It was further argued that though interactionist dualism is obviously as vulnerable as any dualist position to attacks levelled at it *qua* dualist position it is, as a matter of fact, almost always criticised *qua* interactionist position and specifically because it is incompatible with the thesis of causal closure of the physical. (The general view, it was said, appears to be that the thesis of causal closure of the physical is obviously true, so interactionism must be obviously false.) As a result it was argued that it is very important indeed to investigate whether or not there is good reason to believe that the thesis of causal closure of the physical is true, and in doing so it is warranted to assume that dualism about certain aspects of experience is true.

It was suggested in the first chapter that the widespread recent acceptance of the thesis of causal closure without supporting argument violates the *proportionality constraint* on accepting a thesis. Two reasons were given. First, it is not true that not much hinges on accepting the thesis; given the way in which monism gains its acceptance, the unattractiveness of epiphenomenalism and the crucial role the thesis of causal closure plays in discrediting interactionism, quite a lot hinges on whether one accepts or rejects the thesis. Secondly, it is not the case that we have overwhelming reason to believe that the thesis is true. A significant part of the case for this latter claim was made in

the previous chapter. There it was shown that the principle of conservation of energy does not rule out interactionism.

In this chapter the argument for the claim that we lack overwhelming reason to accept the thesis of causal closure — and thus also for the claim that the proportionality constraint is violated where the principle is concerned — is concluded. This comes about by a demonstration of the inconclusiveness of a number of other arguments that are sometimes put forth in support of the causal closure thesis.

5.2 Papineau

The first examined argument is due to David Papineau. It was first published in his paper 'The Rise of Physicalism' (2000). In personal communication Papineau (2006) has, however, confirmed that he prefers the version of the argument found in the appendix to his book, *Thinking about Consciousness* (2002), so that is the version that will be discussed here.

In that appendix Papineau discusses the 'completeness of physics', the claim that "[a]ll physical effects are fully caused [or have their probabilities fully determined] by purely *physical* prior histories" (2002, p. 17). This is also the claim which in this thesis has variously been called the causal closure thesis and the thesis of causal closure (of the physical); those terms will continue to be used here.

Papineau sets out to do two things. First, he wishes to "rehearse the history of scientific attitudes to the completeness of physics" (p. 233); to show *how* belief in the completeness of physics *actually* became widespread. Secondly, however, it is also clear that Papineau wishes to show that to *resist* belief in the causal closure thesis is now futile, given what he takes to be overwhelming evidence for its truth: "I see no virtue in philosophers refusing to accept a

premiss which, by any normal inductive standards, has been fully established by over a century of empirical research" (p. 256).

The historical project is an interesting one, and there are many valuable insights in the text. However, while the spread of belief in the causal closure thesis is an undeniable fact, it does not on its own show that such belief is *warranted*. That part of the job is to be done by two arguments extracted from the historical account: the argument from *fundamental forces* and the argument from *physiology*.

Papineau states the argument from fundamental forces in the following way:

> [A]ll apparently special forces characteristically *reduce* to a small stock of basic physical forces which conserve energy. Causes of macroscopic accelerations standardly turn out to be composed of a few fundamental physical forces which operate throughout nature. So, while we ordinarily attribute certain physical effects to ... 'mental causes', we should recognize that these causes, like all causes of physical effects, are ultimately composed of the few basic physical forces (p. 250, italics in the original).

On its own, Papineau thinks that this argument is insufficient to show the truth of the causal closure principle. The problem he sees is essentially that there might also be *mental* forces that conserve energy; they might "operate in such a way as to 'pay back' all the energy they 'borrow', and *vice versa*" (p. 252).[1] Another argument is therefore needed: the argument from 'direct

[1] Papineau recognises that conservation of energy fails to show that the causal closure thesis is true. However, he still takes conservation of energy to be incompatible with 'spontaneous' or 'indeterministic' 'animate forces', just not with deterministic ones (pp. 243-49). It seems to me that the argument for incompatibility between 'spontaneous' mental forces and the principle of conservation of energy is also inconclusive: "Why shouldn't [a 'spontaneous' mental] force simply respect the conservation of energy by not causing accelerations which will violate it? But this doesn't really make sense. The content of the principle of the conservation of energy is that losses of kinetic energy are compensated by buildups of potential energy, and *vice versa*. But we couldn't really speak of a 'buildup' or 'loss' in the potential energy associated with a force, if there were no force law governing the deployment of that force. So the very idea of potential energy commits us to a law which governs how the relevant force will cause accelerations in the future" (p. 249). As was argued in the previous chapter, however, an interactionist need account only for the *data*, not the theory. Finding an explanation that 'makes sense' in the

physiological evidence'. That argument is simply that our extensive knowledge of the fundamental physical forces would enable us to recognise 'anomalous' physical events — events not caused by physical (conservative) forces — if they *did* occur. Since there is, as yet, no evidence of this, we can conclude, he argues, that no such events occur. He elaborates:

> "[B]y the 1950s it had become difficult ... to continue to uphold special vital or mental forces. A great deal became known about biochemical and neurophysiological processes ... and none of it gave any evidence for the existence of special forces not found elsewhere in nature. ... If there were such forces, they could be expected to display some manifestation of their presence. But detailed physiological investigation failed to uncover evidence of anything except familiar physical forces (pp. 253-54).

This argument can according to Papineau "be viewed as clinching the case for the completeness of physics against the background provided by the argument from fundamental forces" (p. 254).

5.3 Meta-Inductions on the History of Science

5.3.1 THE ARGUMENTS

Both Papineau's arguments fit a schema of arguments one might call *meta-inductions on the history of science*. Such arguments go, in broad strokes, as follows. First, reference is made to a perceived *trend* or *tendency* in the history of scientific development. Secondly, it is claimed (though usually not argued) that we *should expect* that trend to continue; that it is unreasonable not to expect this. Finally, it is argued or implied that when one supposes that the trend will continue, one can see that the perceived anomalousness of the mind will turn

---

theoretical framework is not necessary; one need only come up with a theory that shows how mental forces could intervene in the physical realm in such a way that the data we in fact observe would obtain. No special problem obtains for 'spontaneous' mental forces, since whether the *onset* of a mental intervention is law-governed or not is beside the point. What matters is the *manner* in which the influence is exerted.

out to be illusory. Once the described tendency has developed further, the mind will be revealed as a 'garden variety' physical phenomenon after all.

Although we *lack* the scientific story that would constitute incontrovertible evidence that interactionism is false — the story, that is, that gives a full physical account (or prediction) of all physical events, including those in the brain — we do *not*, according to these arguments, have good reason to believe that the story will not be completed. We have every reason to believe, it is argued, that the project of finding physical causes for physical effects will continue to its completion. It is true that the last blank spaces on the map are in the brain; they will, however, not stay blank for too much longer, and the belief that something *out of the ordinary* is going on in the brain amounts to nothing more than die-hard superstition. So the arguments go.

5.3.2 THE RESPONSE

The strategy in responding to these arguments is very simple. It is quite true that the interactionist claims that there will sometimes be events in the brain that are inexplicable by ordinary physical standards. The interactionist is indeed minimally committed to the claim that the mind sometimes brings about a physical event that would not otherwise have occurred (or at least that it brings it about that such an event occurs *differently* than it otherwise would have) and the most likely location for such an event is the brain. However, how one should evaluate the plausibility of that claim, and in particular whether one should see it as die-hard superstition or not, depends *entirely* on *how out of the ordinary* one thinks that the mind, or a certain aspect of the mind, is. That, in turn, depends on how convinced one is that dualism is true.

To someone convinced by the arguments in favour of dualism it is not at all irrational or superstitious to believe that there may be events in the brain that ordinary physics will not be able to explain; events explicable only when it is posited that they are caused by the mind. Most dualists believe that some

aspect or other of the mind is extremely *special*; it constitutes the *only* exception to the norm that all phenomena supervene on the physical, as it is often put. In that light, the claim that something can occur wherever the relevant aspect of the mind is present that does not occur anywhere else is unsurprising.

That shows that no independent work is done by the arguments that try to show that 'everything points to' a completed physical account. There is simply no reason for a dualist to accept this. Dualist have good reason to say that there are strong indications to the contrary: the arguments that convinced them that dualism is true. They show that the mind, or aspects of the mind, are *unique* in the world; so the view that the mind, as the only non-physical entity, can cause exceptions to the otherwise prevalent order by causally influencing the physical world is not misplaced or disproportionate.

It should be obvious how this general strategy applies to Papineau's arguments. Against the argument from fundamental forces, which states that we should expect what we take to be the manifestation of mental forces to be reducible to run-of-the-mill conservative physical forces, the strategy is just to resist the claim that that is what we should expect. To reiterate, a dualist has extremely good reason to think that the mind, or aspects of the mind, alone are *exceptional* in the natural order, so a dualist has very good reason to be sceptical about the possibility of the proposed reduction of 'mental forces' to physical (and conservative) forces.

In any case, only a weaker claim is really needed, namely the claim that, given the trend in scientific development on the one hand, and the arguments for dualism on the other, we *lack* a clear reason to expect one outcome over the other. Where the mind's influence in the physical world is concerned, we will just have to wait and see. That is enough to dispel that argument as an argument against interactionist dualism and for the causal closure of the physical.

Against the argument from 'direct physiological evidence' the interactionist position is to do exactly what Papineau derides:

> You could in principle accept the rest of modern physical theory, and yet continue to insist on special mental forces, which operate in as yet undetected ways in the interstices of intelligent brains. And indeed, there do exist bitter-enders of just this kind, who continue to hold out for special mental causes, even after another half-century of ever more detailed molecular biology has been added to the inductive evidence which initially created a scientific consensus on completeness in the 1950s (p. 256).

The insistence, in other words, will be precisely that in the brain there are events that physical explanation cannot account for. This is obviously an empirical claim, and it is not hard to see how it could be proven false: a completed physical explanation (or prediction) of all events in the brain would falsify interactionism. To see that such an insistence is nevertheless not yet evidence of lunacy, consider two problems with Papineau's characterisation above.

First, Papineau makes it seem that there are nothing but tiny little mysteries left in the operation of the brain, little sections of its working that we do not yet understand. That is very far from the truth. Our best understanding is most often limited to crude approximations of which *areas* are activated in different circumstances; they amount to guesses — highly educated guesses, but guesses nonetheless — of which areas of the brain are *'associated with'*, for example, rational problem solving and emotional arousal. Beyond that level of precision — a very modest level indeed — the problems with even framing the research-questions are many and deep.

Secondly, as has been stressed, it is not as if interactionist dualism comes unmotivated, with philosophers being 'bitter-enders' for the sake of it. It is not at all *ad hoc* for someone convinced by the arguments for dualism to insist that

events can take place in brains that do not take place elsewhere. The mind, we know, is associated with the brain. According to dualism the mind is truly an *exceptional* entity. So, whether or not it is unreasonable to suppose that events can take place in the brain that cannot take place in the absence of the influence of a mind depends *entirely* on how good a reason one has for believing that dualism is true. No independent work is done by the argument from physiological evidence, neither alone nor in conjunction with the argument from fundamental physical forces.

5.3.3 OTHER VARIETIES

Papineau's argument is perhaps the best developed recent exemplar of this type of argument. Additionally there are other, less developed arguments in the vicinity. One is found in Frank Jackson's John Locke lectures. There he argues that

> "it is reasonable to suppose that physical science, despite its known inadequacies, has advanced sufficiently for us to be confident of the *kinds* of properties and relations that are needed to give a complete account of non-sentient reality. They will be broadly of a kind with those that appear in current physical science, or at least they will be as far as the explanation of macroscopic phenomena go, and the mind is a macroscopic phenomenon (Jackson 1998, p. 7).

The distinctive meta-inductive claim here is that science has given us reason to be confident that the properties and relations needed in an explanation of the mind will be relevantly similar (in some sense) to those that are mentioned in current explanations of other, less puzzling phenomena. The trend of extending the scope of explanations *like ours* will continue, in other words, until the mind is covered as well.

Now, a reasonable point of criticism against such an argument is of course that it is not at all clear what the supposed *kind* of properties and

relations is.[2] However, even granting that there is some way to cash this claim out, a more direct reply is, as before, quite reasonably given by the interactionists, when they simply *resist* the claim that we have reason to be confident that the trend (assuming there is a determinate trend) will continue. The arguments for dualism gives us reason to believe the opposite. And again, even the more modest claim that we have no reason to be confident either way, is more than sufficient here.

Other arguments that fit this schema can easily be imagined. One might refer to the trend of moving from agent-based to non-agent-based explanations; from explanations involving gods with hammers to explanations involving forces and particles.[3] Another might refer to a development from explanations adducing both non-conservative and conservative forces toward explanations only adducing the latter type of force, and claim that mental forces must fail to conserve, and so that we should expect them to be reducible to physical forces. A third might point out that it was previously thought that the material in animals was somehow special, now we think that it is not. Since the same stuff makes up the brain as everything else, we should expect the behaviour of that stuff to be explained in the same way as we already explain the behaviour of stuff elsewhere.[4] A fourth variety might be motivated by several of these considerations, and say, in general, that science has historically been ever more successful in discovering physical causes for physical effects; a trend we should believe will continue until purported effects of *mental* causes have been included. In all cases the response from the interactionist is the same, and in all cases is it as justified.[5]

---

[2] See the section on defining physicalism in chapter three.

[3] Jackson and Braddon-Mitchell use this example in their recent textbook (1996, p. 8).

[4] This argument is in the lore associated with presentations given by Frank Jackson, but I am not aware of it appearing in print. Either way, the question he supposedly asks is: "Do you really mean to tell me that the atoms in my brain are under the influence of something more than the forces that control the behaviour of particles everywhere else in nature?" To which the interactionist should reply: "Yes".

[5] It is interesting to note that meta-inductive argument have been presented in much the same way for some time now. One early example of a meta-inductive argument is named 'The Shadow of Physiology'

5.4 Unity of Science

A general claim made in this work is that the causal closure thesis has been more influential than the strength of the arguments in its favour warrant. It has, it seems, often been thought a solid argument for the thesis of causal closure exists, and this belief may at times have been held with sufficient conviction to allow for the omission of a more careful investigation of the argument itself. Among the arguments *thought* to conclusively demonstrate the truth of the causal closure thesis it seems likely that the argument from conservation of energy may have been the most influential. The runner-up is harder to judge, but if meta-inductions on the history of science come second, perhaps considerations of its unity come third.

Among 'considerations of scientific unity' are included anything from a hope or dream of a unified 'theory of everything' — in which all parts fit with one another — to an *observation* that there is a *tendency* for our scientific knowledge to become more unified. An oft-cited paper in this area is 'Unity of Science as a Working Hypothesis' (Oppenheim and Putnam 1958). In it, the authors argue that "the assumption that unitary science can be attained through cumulative micro-reduction recommends itself as a working hypothesis"; on the authors' view it is a *credible* hypothesis, one we have *good reason* to believe (p. 8). In their terminology 'unitary science' is an expression for "the ideal of ... an all-comprehensive explanatory system" in which "the laws of science become reduced to the laws of some one discipline" (p. 4). Their claim, in other words, is that a state where the laws (and vocabulary) of all scientific disciplines has

---

by Campbell, in his *Body & Mind* (1970/1980, pp. 39-40). Campbell's argument for the claim that "the evidence suggests that nothing happens in a man save what conforms to physical and chemical law" is *only* that "research based on the assumption that the brain obeys only physiochemical laws has not yet suffered a damaging reverse" (p. 39). Once the first objection to Papineau's account above is taken into account — that there is *very much* we do not yet know about the workings of the brain — it is clear that the lack of a setback *as of yet* in the research programme that assumes no mind-brain interaction can, at best, only nudge the balance of probabilities the slightest of fractions in disfavour of interactionism. We may well have to become considerably more sophisticated in our understanding of the brain before it even becomes possible to look for counterexamples to the assumption.

become reduced to laws (and vocabulary) of a single discipline, is a credible goal.

In the article, the authors argue extensively for the *credibility* of the working hypothesis. For the present purposes it is, however, most natural to first investigate a different claim, also made in the article, namely the claim that "the only method of attaining unitary science that appears to be seriously available at present is micro-reduction" (p. 8).[6] Here is why that claim should be investigated first.

The claim that micro-reduction is the only way toward unitary science is either true or false. If it is true, interactionist dualism appears to have been posed a significant challenge, for it appears that interactionist dualism would then imply the unattainability of unitary science. The dualist stance (as traditionally conceived) is that the mind is *irreducible* to matter, and the interactionist stance is that the mind influences the behaviour of matter.[7] Interactionist dualism would block progress toward unitary science either by blocking the micro-reduction of the behaviour of living organisms to the behaviour of the cells that are parts of those organisms, *or* by blocking the reduction of the behaviour of the cells of the organism to the behaviour of the constituents of the cells, depending on where the interactionist theory supposed the interaction to take place.

Thus, if micro-reduction were discovered to be the only way to move toward unitary science it seems that interactionist dualism would imply that the process of unifying science would be in principle impossible to complete. At such a point it would be pertinent to investigate whether this should be

[6] Micro-reduction is defined as reduction in which the reducing theory "deals with the parts of the objects" that the reduced theory deals with: cellular biology deals with cells, and chemistry deals with the *parts* of cells, so chemistry is, in their terminology, a 'potential micro-reducer' of cellular biology (p. 6).

[7] I discuss, in this section, the matter of the supposed incompatibility of dualism *as traditionally conceived* with the goal of unity of science. If dualism is instead conceived in the way which Bigelow and I suggest that it should be (Unpublished, see Appendix A), there is not even an apparent conflict between dualism and the unity of science. The discussion should therefore be taken to apply only to traditional dichotomous dualism, a restriction I leave implicit henceforth.

considered a *disadvantage* of the interactionist dualist position. The way to find that out would be to examine the other main argument in the article; the claim that the working hypothesis is credible. After all, if it were incredible that unitary science should be achieved anyway, interactionist dualism could hardly be faulted for denying the possibility. Since the present purposes are to ascertain the force of considerations of scientific unity as an argument for the thesis of causal closure, or against interactionist dualism, the claim that the working hypothesis is credible need only be investigated if it turns out that micro-reduction probably is the only way to move toward unitary science.

Someone wishing to resist the claim that micro-reduction is the only avenue that leads to unitary science might at first be tempted to re-run the strategy that was applied to the meta-inductions on the history of science. This is natural; on one understanding of 'considerations of scientific unity', such considerations are simply observations to the effect that scientific knowledge tends to become more unified. As such they certainly qualify as meta-inductions on the history of science. Against observations such as these the strategy applied above *is* successful. However, one should resist the temptation to believe that that is all there is to considerations of scientific unity.

In the meta-induction cases — including one variety of considerations of scientific unity — the defensive strategy outlined above claims that whether one should expect a trend to continue or not depends on whether or not one is convinced that dualism is true. If one is convinced that dualism is true, one does *not* have good reason to expect the trend to continue to include the mind, so arguing that we should expect this begs the question against the dualist.

We now need to ask whether a question is being begged in the cases where 'considerations of scientific unity' means a *hope* that we are moving toward a certain state of affairs, or a statement that such a state of affairs constitutes a worthy *goal* to work toward. If dualism is true, it is either the case that science cannot be unified or that there is a different way to unify science

than by micro-reduction. Putnam and Oppenheim argue that there is no other way to move toward unified science than by micro-reduction. If that were true, the dualist would have to *admit* that accepting dualism means relinquishing the ideal of unified science; accepting that the stated goal is out of reach. Moreover, since the question of whether or not it is *true* that no other way to move toward unitary science exists is to be decided *independently* of arguments for or against dualism, there is no appearance of question-begging here. Consequently, the strategy that works against meta-inductions on the history of science *fails* on at least one understanding of what 'considerations of scientific unity' means.

This notwithstanding, I think that both the claim that the only way to move toward scientific unity is through micro-reduction, and the derivative claim that subscribing to interactionist dualism means relinquishing the goal of unified science, can be resisted.

Micro-reduction offers a genuine promise of unification of sciences. Dualism is incompatible with micro-reduction. The question becomes whether there is some *other* way to genuinely move toward unity of science. However, not just any old way of moving toward unity will qualify. As Oppenheim and Putnam point out, while it is difficult to cash out what it might mean for the laws of a scientific discipline to be "in some intuitive sense 'unified' or 'connected'", there is, on the other hand, a real threat of *vacuousness* of the notion of unity of science if *no* such criterion is imposed (p. 4). We need to address the question of whether interactionist dualism is compatible with a unification of science in a more substantial sense than that which would be implied by a simple *conjunction* of a theory of the mind with the other theories we need to explain the world. So we need to consider whether there is some way of understanding *neat integration* of theories that does not rely on micro-reduction.

This question of whether neat integration is possible can be understood in terms of the intuitive notion of being *'on the same level as'*. If neat integration

is to be possible, then, in some sense, the entities that require integration need to be on the same level. This way of thinking allows what drives the intuition that traditional dichotomous dualism is incompatible with unity of science to be analysed as a confusion of two very different ways in which entities can be *on the same level* as one another.

One of the senses in which two entities can be on the same level is if the entities are roughly or comparatively of the same *size* as one another. Different scientific disciplines are typically concerned with entities of different *size*. Micro-reduction, in the sense used by Oppenheim and Putnam, promises an integration of sciences through the 'decomposition' of higher-level entities, which we need to explain, into *smaller*, lower level entities, which we also need to explain. Because larger entities are thought to be *constituted by* smaller entities, one can imagine that the explanation of the higher level ones will be exhausted by the explanation of the lower-level ones. Thus, by bringing all the entities that need explaining down to *the same level* of magnitude one can imagine the integration of the all the theories into one; that which deals with the smallest entities there are. The entities are on the same level; they can be integrated.

The mind, however, is a macro-phenomenon, and according to dualism, the mind is *not* reducible to the micro-entities physics deals with. Thus it is clear that the mind cannot be *on the same level* as the basic entities of physics *in this sense* of being on the same level, if dualism is true. The mind, in other words, resists integration in this way, by resisting decomposition into smaller entities. (For those who think that the mind supervenes on, or *is*, a certain organisation of micro-entities, mind has no tendency to be seen as a threat to the integration of science.)

It is yet to be determined, however, whether the mind can be on the same level as those entities in *any* sense that allows for integration of science.

There is another sense of being on the same level that shows promise in this regard, and that is being on the same level of *primitiveness* or *fundamentality*.

Consider Newtonian mechanics. Newtonian mechanics is a unified system; it 'pans out' in a predictable way. Nevertheless, the system has four primitives: time, space, particular objects and forces, and none of these are reducible to any of the others. The dualist can be viewed as arguing for the need of adding an *additional* primitive to the list. Does the addition of an additional primitive entity threaten the integration of science?

No. There is no intuitive support for the claim that the mind cannot be unified with the other primitives of our final physical system, for there is no (obvious) non-question-begging reason to think that mental states cannot be *on the same level of fundamentality* as the other primitives we need in our explanatory system. (The fact that it would have been possible for minds never to evolve, and the fact they apparently evolved only after much time had elapsed, for example, has no tendency to disprove the mind's claim to fundamentality; narrow application-conditions make a law no less fundamental. For all we know, some conditions could be made to arise through experiment that *never* arise in the natural world; that does not make the outcome of such an experiment fail to be governed by fundamental laws.)

I am not suggesting that all the primitive entities in either Newtonian or modern physics have *size*; it makes no sense to think of a force having size. The point is rather that the primitive entities in physics are either thought to themselves be of a certain magnitude, or else thought to *be able to act on* entities on that magnitude. It is, it seems, in virtue of being thought unable to be integrated in *either* of these ways that the mind is thought to be excluded from a unified scientific theory. It is clear that the mind is not *itself* on the same level of size as the basic entities in physics; inasmuch as it makes sense to attribute any size at all to the mind, Jackson is right in calling it a macro-phenomenon. It is much less obvious that the mind cannot *act on* entities that are of the size of the

smallest particles. The claim that the mind could act on entities of the smallest size does not entail that the mind should be able to act on *all* such entities. It is very reasonable to think that a certain *configuration* of those entities would be necessary in order for the mind to be able to so act; and in that case there is some sense in which the mind could only act on entities somewhat removed from the basic level; perhaps on the level of neurons, or perhaps on the level of overall firing-patterns of neurons. But in either case, the mind would be on the level of fundamentality in the sense that it would be able to *act on* the smallest entities, either directly or indirectly.

What makes the Newtonian system unified is the simplicity of the laws that connect the primitives. If the new primitive that the dualist argues for is on the same level *of fundamentality* as the other primitives in the system, the question of whether or not the system is unified becomes one of the simplicity of the laws that connects the primitives of the system. Some may at first find it hard to imagine that there could be simple laws governing the interaction of the primitives, if mental states were included among them. However, there certainly are phenomena in the natural world that do not, on the face of things, betray a simple underlying order, as anyone who has seen falling leaves or waves meeting shore will know. Simple laws do not entail that the *applications* of the laws will be simple. Discovery of simple laws might lead to an appreciation of why the application of the laws has such complexity; and this might ease the tension caused by the intuition that interaction between mind and body must be exceedingly complex.

If interactionist dualism is true, science can still be unified, for simple laws may govern the interactions between the mind and the other primitives of the system. That suffices to show that the argument from the unity of science fails.

5.5 Lack of 'Wiggle Room'

An argument that is sometimes raised is that modern physical theory leaves no 'room' for the mind to interact. The thought here is that there must be some indeterminacy in the physical system to allow for the influence from the mind to be 'slotted in'. Chalmers has argued that the only way to carve out room for the mind to interact with the physical world that looks "remotely tenable in the contemporary picture is one that exploits certain properties of quantum mechanics", e.g. through giving it a role in collapsing the quantum mechanical wave function (1996, p. 156). He has further argued, however, that this is unlikely to give interactionists what they need (p. 157).

It seems to me that the interactionists should be bold in their claims, and in particular that they should not suppose that the only way we could conceive of the mind interacting with the brain is through quantum mechanical 'loop holes'. There are straightforward ways that the mind could interact with the brain; by causing some neurons to fire that would not otherwise have fired, or by causing a neuron that would otherwise have fired, not to. Of course, the thesis that this is what the mind does is defeasible in light of empirical evidence, but that is a strength of the suggestion, not a weakness. It is only *thought* to be a weakness of the account in light of meta-inductive considerations to the effect that it is *unlikely* that the mind could so interact, that it would be a break with a *trend* that we have every reason to believe will continue. Once those arguments have been rejected, as we have seen that they should be, the claim that the mind interacts in the straightforward way just described becomes much easier to accept. There is certainly enough we do not know about what goes on in the brain to show that the worry that interactionism will *soon* be proven false by the completion of the account of the brain to be entirely unfounded. There is *no need* to 'carve out' any room for the mind to interact with the physical world; there is, already, room to be had in the brain aplenty.

Conclusion

Interactionism is an attractive theory about the mind. It is, in particular, the most attractive dualist position there is. Commonsense demands, in no uncertain terms, that phenomenal experience be afforded causal efficacy. Interactionism delivers.

Whether dualism itself is true is still an open question. No sustained attempt has here been made to justify belief in dualism, and no defence has been offered against arguments aimed equally at all dualist positions. If, however, it is true that that monism derives a significant amount of its support from the elimination of dualist alternatives, and if it is true, furthermore, that interactionism is almost always excluded *not* on the basis of arguments that cut all dualist positions equally, but rather on the basis of arguments specifically levelled against the position *qua interactionist* position, then this suggests that interactionism is worth a second look. Whether monism or dualism is true is a deep, difficult and important question. We cannot afford to be swayed by rash dismissal of viable alternative positions.

This work has considered arguments against interactionism starting from results in science, trends in the history of science, and arguments of a conceptual nature. None were found to be decisive. If no significant argument strategy has been overlooked, and in the absence of decisive arguments levelled against all dualist positions equally, the discussion leads to the following, perhaps surprising, conclusion: interactionist dualism is alive and kicking; it is a fully acceptable position to occupy, also for scientifically–minded participants in the contemporary debate in philosophy of mind.

APPENDICES

Appendix A
Dualism by Degrees

John Bigelow and Ole Koksvik
Monash University

Abstract

Philosophers often ask one another whether they are monists or dualists. They very seldom ask: "On a scale from one to ten, how much of a dualist are you?" We hope that that will change. In place of a strict dichotomy, we argue that the difference between monism and dualism in the philosophy of mind should be seen as variation along a continuous scale. Where on the scale a given view belongs is best characterised in terms of how radical it says that a break from current physical science will have to be — how much modification or supplementation will be required — in order for the joint science that then emerges to be seen as a likely candidate for a theory that can account for consciousness, thoughts, and feelings.

A.1 Introduction

Philosophers are often asked whether they are monists or dualists. They are seldom asked "On a scale from one to ten, how much of a dualist are you?" The second question, the pertinence of which we argue for in this paper, is not intended as the question: "How confident you are that mind–body dualism is true?" Rather, the question is something more like the following: "How radical will the changes be, which you think current physical sciences will have to undergo, either by alteration or by supplementation, before the resulting joint sciences can account for the mind?" (We use 'the mind' as our catch-all phrase for perceptions and thoughts and feelings and whatever 'mind–body dualists' are dualists *about*.) If you think the changes will have to be substantial, you are quite a bit of a dualist; if you think they will be minor, you are not much of a dualist.

This 'spectrum-theory' is not, however, one that standardly governs debates in the philosophy of mind. Philosophers generally treat the difference between materialism and dualism as if it were an on/off distinction; either there is *one* kind of thing, or there are *two*; and that is all there is to say about characterising dualism. Since no sensible theory will that say the number of kinds lies somewhere *between* one and two; dualism can obviously not come in degrees. That is the received view.

Thus the materialist, it is said, believes that only one *kind of thing* will have to be invoked to explain everything in the world. She is said to believe, specifically, that the entities needed to explain the mental features of the world are in some sense *the same as* those that are needed to explain the physical features of the world. The dualist, in contrast, is taken to believe that when the chips are all down, when we have come as far down the road toward completing our inventory of things in the world and our account of how it works as we will ever come, we will have discovered that *two essentially different kinds of entities* are needed for the explanation.

(Here the term 'entities' is deliberately left vague. Some dualists believe mind and body are different *substances* or *individuals*, others that there is a distinction between mental and physical *events* or *processes*, yet others that there are both physical and mental *properties*. We stretch the term 'entities' to cover substances, individuals, events, processes, and properties, and perhaps even more besides. It should be acknowledged that different mind–body dualists might affirm dualities in different ontological categories. Nevertheless, standard accounts of what dualism is take all dualists to agree that entities of *some* relevant category fall into two 'essentially different kinds'.)

Here is one example of such an account. In their recent textbook, Jackson and Braddon-Mitchell (1996, p. 3) describe dualists as believing that "the ingredients we need in order to understand and account for [mental phenomena] ... are different in kind from those we need in order to account for

the [material phenomena]", but "[a]ccording to materialists the ingredients are essentially the same".

This paper argues that this textbook characterisation is less than ideal. One of the authors of this paper self-identifies as a dualist (the other is shifting ground), and Jackson and Braddon-Mitchell's description of the dualist does not comfortably fit his position. Must a dualist think that mental phenomena are 'different in kind', in order to deserve to be called a dualist? Perhaps, on *one* understanding of what 'different in kind' means, but there are different *kinds* of 'differences in kind'.

If a dualist says things divide into *two* kinds, does a materialist therefore have to say that there is only *one* kind of thing? It seems that a materialist would be unlikely to say that there is only one kind of thing. Some, at least, will say there are *many* different kinds of things. For instance, they may say there are both photons and protons, and that photons and protons are very different kinds of things. Even Lucretius, whose sparse ontology consisted of nothing but 'atoms and the void', said that there are *both* 'atoms' *and* 'the void', and that empty space is a very different kind of thing from the atoms that occupy space. When Jackson and Braddon-Mitchell say that according to the materialists the ingredients we need for understanding mind and matter are 'essentially' the same, it is not implied that they are of *exactly* the same kind, but only that the differences between them are not very *deep* or *ultimate* ones. Yet this draws into muddy metaphysical waters: how shallow does a difference have to be in order for two different things to be "essentially the same"?

We shall argue that a dualist's intuitions need not concern any such 'deep' sameness or difference in kinds at all.

A.2 Dualism by Degrees

A dualist view in the philosophy of mind arises to a significant extent from *intuitions*, that is, from judgments that are held with confidence even though it

is hard to articulate the reasons that support those judgments. (A guiding model is that of judgments of ungrammaticality by native speakers of a language.) A dualist's intuitions are, of course, subsequently carefully examined and questioned; dualism is a name given to an opinion that withstands fairly rigorous scrutiny, and to which people adhere because it is their considered, argued and justified belief that it most accurately describes reality. Nevertheless, intuitions often play a significant role both in getting the process of theorising started in the first place, and as a part of that process.

What is at the heart of the driving intuitions of the dualist? Are they concerned with the configuration of reality according to a 'final' scientific account, and with whether the things described in that account will fall into one or two 'essentially different' kinds? The configuration of reality according to a final scientific account (that is quite probably utopian) and questions of essential sameness or difference are, we submit, not good candidates for what the most central dualist intuition is about.

According to the family of characterisations to which Jackson and Braddon-Mitchell's description belongs, the distinction between different kinds of entities that dualists think will suffuse our final account of reality is not a superficial one. Talk of difference *in kind* and of *essential* similarity indicates that the difference alluded to here must be a deep difference, a difference in *natural* kind, the sort of difference Plato had in mind when he spoke of 'cutting nature at the joints'. But how much more precise can we be about this?

Imagine an isolated tribe of materialists living in a very dry region. Their biologists know about insects and plants and birds and mammals – but they do not know about fish. When they discover fish their worldview changes. They have included a *new kind* into their theory. Arguably, it is a new kind that is *essentially different* from the ones they knew about before. Perhaps there is some sense in which this new kind might also be described as 'essentially the same' as the kinds they knew about before, but it is not clear how to decide whether

the new kind is 'essentially the same' or 'essentially different' in the appropriate sense. That is because it is not clear what the appropriate sense should be.

Compare that imaginary case with our own, and imagine that psychology were to require the postulation of some new kind in the *same sense* in which fish count as a 'new kind' for the isolated tribe. Would *that* change in psychology be tantamount to an admission that dualism is true? Presumably not. But if monism is more than a trivial thesis that is true by definition, there must be *some* point at which such an admission should be made: *some* newly discovered mental entity would have to be, as it were, essentially different *enough* to vindicate dualism.

It is a familiar fact that a concept having vague boundaries does not preclude there being clear exemplar cases, both to which the concept does apply and to which it does not. Our point, however, is not merely to argue that the concepts of essential sameness and difference and of natural kindhood have vague boundaries. The point is rather that dualists have no way of knowing *which concept is being employed* when – treating all dualists as natural-kind-dualists – some describe their dualism as the view that mind and matter are of "essentially different kinds".

Some natural kinds are standardly taken to contain others. The question of sharing natural kindhood must therefore be made precise by the additional question of how restricted the shared kind has to be. And, since an entity can have more than one essential property, two entities can *share* one or more of their essential properties and, at the same time, also *differ* in one or more of their essential properties — and so they may be both *essentially similar* and *essentially different* to various different *degrees*.

Questions such as which level of generality to single out from a nested hierarchy of natural kinds, and *how* essentially similar is essentially similar *enough*, are – needless to say – highly esoteric ones. Dualist intuitions, on the other hand, are both ubiquitous and deeply rooted — among both philosophers

and the 'folk'. That fact is, we submit, an *explanandum* ill explained by reference to concepts such as similarity or difference in kind, and 'essential' similarity or difference, since the application of these concepts is obscure even to philosophers who routinely consider these matters. What we need is an account of what the dualist intuition is about that does not make it come out as either arcane or confused and ill-conceived.

When dualism is taken to refer to a strict dichotomy, the problem is not just how to apply a concept with possibly vague boundaries to various cases. It is not that we could not just *pick* one way of applying the term, and say that *that* is what dualism means. Here the point is that any on/off distinction is ill suited to pick from the wide range of concepts – each with slightly different scope – that operate in this vicinity. *How* essentially different is essentially different *enough*? Just picking an answer, more or less arbitrarily, would do *injustice* to a great number of positions – both dualist and monist – for the selected distinction would not capture the distinctions that those positions took to be the salient ones. It would fail to capture their content.

It is very likely, we submit, that part of what the core dualist intuition primarily concerns is not ultimate sameness and difference, but rather something as vague but familiar as "the kind of explanation science typically puts forward nowadays". This is, we submit, a much better candidate, being both ubiquitous in our thinking and accessible to the 'folk'. True, there are certainly aspects of today's science that are grasped by a very small number of people. Most of us are no more proficient with general relativity than we are with concepts like *natural kinds* or *essentially similarity*. Nevertheless, most of us do have, or at least think we have, a general grasp of what we take to be typical contemporary scientific explanations. (Someone who did not have *any* grasp of what current physical explanations are like, would, we argue, not really qualify as a dualist.)

For example, we hear that iron rusts because oxygen from the air combines with iron to make iron oxide, which is reddish; and when copper combines with oxygen it forms copper oxide, which is greenish. A magnet attracts iron because it emits a force field. A rainbow forms because droplets of water in the air act as little prisms that divide the sunlight into different colours. Metals expand when heated. Milk goes off when it is left in the sun. And so on.

Thus we submit that one main source of the driving intuition behind dualism comes from the perceived character of explanations invoked by the physical sciences at the present time: not the imagined character of scientific explanations in some utopian future, but the apparent character of those that are on offer right now. On this view, driving a dualist intuition is a hunch about *the extent to which a theory (or set of theories) would need to differ* from the physical sciences of the intuiter's current acquaintance in order for it to seem likely to him or her that such a science would be able to explain the mind.

This line of thought leads to interesting consequences. One is that it suggests new questions, like: "How much of a dualist are you?" Another consequence, and, one should think, a welcome one, is that it may turn out that positions that have been thought to be far apart are actually closer than we thought.

A.3 Units of Measurement

How might one go about answering the question "How much of a dualist are you?"? This is a request for some sort of a measure of "how much". One natural way to measure things that come in degrees is by choosing one or more salient examples as units of measurement, and then measuring other things by comparison with the designated units.

Applying this general technique in the current special case, we could proceed like this. We could study a *transition* or *upheaval* in the history of science, and then take that as our unit of measurement for amount of change. A

dualist could then point to this example, and say: "I think physical science would have to undergo a change or supplementation of *that* magnitude before it could be able to account for the mind".

This enables us to articulate a measure of how much of a dualist you are. The bigger the historical upheaval you point to as your unit of measurement, the more of a dualist you are. There will, of course, remain much room for discussion about how the various upheavals in the history of science compare to one another in magnitude. Our proposal is not designed to end discussion; it is designed to help give that discussion content and structure.

In this description above it is assumed (for simplicity) that a dualist will hold that we need just *one* upheaval in physical science, or *one* supplementation of physical science, in order to explain the mind. There are, of course, other possibilities.

Some dualists might think that a very small transition would be needed before we could explain *some* phenomena that fall under our "catch-all" phrase 'the mind', but that we would need a very large upheaval before we could make progress on others. The mind encompasses *inter alia* thoughts and feelings, bodily sensations, beliefs and desires, subconscious motivations, and consciousness. It is not clear that the upheavals in science that will permit us to explain some of these mental phenomena will automatically enable us to explain all the others.

It is natural, for example, to think that there would have to be some upheaval in science before we could fully explain mental phenomena like *beliefs* and *desires*, but that there would need to be a *different* upheaval before we could fully explain *visual experience* and other aspects of *consciousness*. One kind of dualist, then, will be one who thinks that there will be a need for several upheavals in physical science – quite distinct in both character and magnitude – before we fully understand the mind.

This suggests that even a conception of dualism as varying in degrees along a single dimension may yield too simple a picture: perhaps the correct classification is very complex. Those who self-identify as dualists would reasonably see their intuitions as vindicated even if some mental phenomena are explained in the wake of relatively minor upheavals in the physical sciences, provided that some other ones are left unexplained until we make much more radical departures from the scientific theories with which we are familiar at present. The possibility that even continuous variation along a single dimension will be too crude a measure certainly strengthens the view that dualism should not be conceived as a strict dichotomy.

A.4 Action at a Distance

Consider the triumph of the Newtonian theory of gravitation. We are inclined to underestimate how large a change was required, in order to become accustomed to a theory that countenanced *action at a distance*.

From ancient Greek times through to the present many people have believed in *magic*, and many scientists have set themselves in opposition to all kinds of magical explanations. From within the mechanistic worldview – where objects were thought to act on one another only when contiguous or when series of physical events bridged the gap between them – the idea that one thing could act on another at a distance, with no mediating chain of events connecting them, must surely have seemed magical. The idea that the waters of the Mediterranean move towards the Moon because they are 'attracted' to it was thought by Galileo to be magical thinking, and he tried to find a much more mechanistic explanation.

We still lack a proper understanding of *how* gravity can act at a distance, but for several centuries at least this failed to trouble the scientists. They had become accustomed to the lack of an explanation, and accepted that there *just is* a basic law about how events in one place will reliably be correlated with events

at some other place, with no explanation whatever of any intervening series of events that spatio-temporally connects one of those places with the other.

Twentieth-century physicists have not always been so complaisant: some have spoken of *gravity waves*, for instance, and searched for a unified field theory, in which there are explanations of the mechanisms by which gravitational forces are transmitted from one location to another. Our point, however, concerns the history of science, and not just current science. Past scientists, and many of us amateurs, in our vague understanding of current science, have become accustomed to accepting that there can be action at a distance, with no explanation of any intervening mechanisms. If we do imagine a mediating 'mechanism' it involves the action of invisible, disembodied forces. What would have sounded magical and unscientific to Galileo no longer elicits either hostility or curiosity; we no longer wonder *how* gravity manages to act at a distance without any intervening 'bodily' mechanism.

Current physicists may hold a different theory of gravity, but there are still astonishingly 'magical' elements in quantum mechanics – involving things which at least *seem* to involve action at a distance. Instantaneous correlations in the states of particles that are arbitrarily far removed from one another certainly seem very strange to the lay man, but they are nevertheless accepted by some physicists.

Imagine, then, that a change were to occur of a similar magnitude to that involved in the Newtonian introduction of the force of gravity or the introduction of quantum mechanical correlations, and that, as a consequence, we were to become accustomed to the lack of answers for some of the questions that bother us today. If the mind is sufficiently different from the body to permanently resist explanation in terms like the ones used by today's physical sciences, but if it *could* be described in a science that has undergone a transformation on the scale we are discussing, then unanswered questions – which in today's light appear deeply damaging to dualism – may lose their

impact on us. One such question is how something 'non-physical' can causally 'act on' something physical. This seems mysterious indeed to many, and the lack of even the beginnings of an answer to this question has surely made many lose patience with interactionist dualism.

Yet a dualist might have a hunch that in the light of a future science we will appreciate that the question was somehow misconceived. Perhaps we are now in the grip of a misleading conception of causation, comparable to the conception that causation required bodily, spatiotemporal contact. Perhaps we will come to see what now looks like a glaring hole in theories as a question of lesser importance that we can afford to leave unanswered. Indeed, some already argue that a lesson from Hume is that requesting an explanation of how a 'causal nexus' between the physical and the mental can obtain is misguided (Chalmers 1996, p. 170).

A.5 Who Were Right?

Consider the possibility that all mental phenomena could be explained in a science that had undergone a transformation of the indicated magnitude. How should we then, after the transformation, judge the views we hold today? Should we say that we had discovered the mind to be physical, and so that we had also discovered that the dualists were wrong and the monists right all along?

Many dualists would, we imagine, answer in the affirmative if asked to reflect on whether a change of this magnitude taking place *now* – starting from contemporary scientific theory – might result in a theory capable of explaining mental phenomenal. And, we submit, such an admission from a dualist should *not* prompt us to relabel their (current) view as a monist one. For, as we have argued, dualism, at least one of the most plausible brands of it, is best understood *not* as a view about the ultimate unity or disunity of reality as explanation at the end of inquiry would portray it, but as a view that considers

relative difference between adequate explanations and the explanations found in the scientific theories of the day.

We have argued that the core dualist intuition consists in seeing a break with current science as necessary, so a part of the dualist intuition is that the mind is to some degree mysterious to us *now*. It need *not* be any part of a dualist intuition or position, however, that the mind be *ultimately* mysterious, or somehow beyond the reach of scientific inquiry. When dualist positions are no longer taken to allege *ultimate* mysteriousness for the mind, dualists should also be released from the accusation of being disbelievers in science or in its ultimate unity. Although consistent with disbelief in those lofty goals dualism does not in itself entail such disbelief.

Thus, we submit, some modern dualists are most charitably interpreted not as unimpressed by science and intent on preaching its limits, but rather as concerned with impressing upon others the difficulty of the problem of understanding our thoughts and feelings and the actions they cause, the dangers of complacency, the importance of curiosity and the magnitude of the upheavals that science is still capable of going through. Modern dualist positions typically preach these points *not* to impress upon us the pointlessness of investigating an area that will ultimately remain mysterious, but rather to steer us down a fruitful road of inquiry, one that will eventually lead to a scientific account of the mind.

A.6 Contrasts

Some distinctive features of the thesis that dualism comes in degrees are best appreciated when the view is contrasted with other philosophical positions in the vicinity. One such position is what we might call the *no-issue* theory. This is a theory according to which the appearance of a deep difference between dualism and monism is an illusion, with no substance behind it. Influential versions of this theory have been advanced by Hempel (1980) and Crane and

Mellor (1990). The no-issue view serves as a useful contrast by reference to which the spectrum view can be elucidated.

There are various ways of advancing a *real-issue* claim about dualism and monism. Advancing such a claim usually entails proposing a *definition* of *the physical* designed to show that there is a non-vacuous thesis that "everything is physical", and hence that the difference between monism and dualism is no illusion after all.

One sub-group of such proposals contains views that claim that physicalism is amenable to a definition in terms of *supervenience*. Proponents of this kind of theory include Pettit (1993), Jackson (1998), Oppy (2001), Kim (2005) and others. Another proposal, '*realisation physicalism*', uses the notion of *realisation* instead of the notion of supervenience to formulate physicalism: Melnyk (1996) is a defender of such a view. A third proposal is the so-called '*via negativa*', which begins with a positive characterization for the 'mental', and then the monist view is defined to be the view that everything that exists is 'non-mental': this is a view defended by Spurrett and Papineau (1999) and Montero and Papineau (2005). There are sure to be other alternatives.

In this paper we put forth a real-issue claim about physicalism, but it is *not* our project to discuss and evaluate all the other real-issue theories. We are articulating and defending a spectrum theory; but we are not (here) undertaking the longer task of articulating and attacking all the rival, dichotomous, theories.

We take it that real-issue theories either explicitly defend or at least presuppose the existence of a sharp dichotomy between physicalism and dualism. Our arguments encompass all dichotomous theories in ways that are logically independent of detailed differences among those theories. We have argued that at least *some* dualist intuitions will not be catered for by *any* dichotomous theory. The views of these dualists, as well as the views of many who take themselves to be monists, are, we argue, done justice only if we

abandon the way of thinking that leads to a dichotomy rather than a spectrum. A two-party system in politics runs the risk of steering a country in a direction almost half of its populace disagrees with. Any strict dichotomy between physicalism and dualism runs the risk of disregarding, for a great number of positions, weighty considerations that pull those positions away from the extreme ends of the scale.

No-issue theories, as philosophical doctrines, can usefully be distinguished from a certain other opinion in the vicinity, a more sociological than philosophical thesis. This view was well expressed by Chomsky (1968, pp. 83-84):

> It is an interesting question whether the functioning and evolution of human mentality can be accommodated within the framework of physical explanation, as presently conceived, or whether there are new principles, now unknown, that must be invoked, perhaps principles that emerge only at higher levels of organization than can now be submitted to physical investigation. We can, however, be fairly sure that there will be a physical explanation for the phenomena in question, if they can be explained at all, for an uninteresting terminological reason, namely that the concept of "physical explanation" will no doubt be extended to incorporate whatever is discovered in this domain, exactly as it was extended to accommodate gravitational and electromagnetic force, massless particles, and numerous other entities and processes that would have offended the common sense of earlier generations.

One of the points made in this quote was foreshadowed earlier: namely that aspects of theories that 'offend' us now – the lack of an explanation of how something physical can interact with something non-physical, for example – may appear entirely benign to later generations. The apparent 'no-issue point' expressed is that whatever turns out to be necessary to explain the behaviour of (already countenanced) physical phenomena will very likely itself be labelled 'physical'. This thesis, however, concerns a *sociological* point rather than a philosophical one. A prediction is made of future human behaviour, that of

future scientists. It is conjectured that the boundaries of a term is likely to be allowed to expand. The argument given in favour of that point is an induction on the history of science: similar adjustments to the boundaries of terms have taken place in the past.

A philosophical problem remains, however, whether or not the sociological point holds true. If we could give a non-empty definition of what *we* take a physical entity to be, then we should be able to say of the future scientists whether they had *changed the meaning* of our term 'physical' or not. By applying our own well-defined term we could look at their theory and determine whether or not it invoked any entities that *by our lights* are non-physical. If it did not, they would not be said to have changed the meaning of our term 'physical', if it did, they would. This is quite distinct from the question of what *words* they apply to these entities; if they have changed the meaning of our term 'physical', then the fact that they *call* the new entities "physical" does not make them physical entities in *our* sense. Chomsky's thesis concerning future changes in word-usage does not establish that the issue concerning dualism is vacuous, or (as he says) "uninteresting". So the comparatively uninteresting sociological point is independent of the philosophical issue that no-issue theorists raise.

The *philosophical* question is whether a non-empty definition of physicalism *can* be given, under which physicalism will be at least plausible. Hempel (1980, pp. 194-95) gave a succinct formulation of the problem, and in consequence the problem has sometimes come to be described as "Hempel's dilemma". What Hempel argues is that physicalism is either vacuous or highly improbable.

Monists, it is argued, claim that at the end of inquiry we will have on our inventory of the world (the names of) entities of only one kind: the physical ones. But what are we to understand by 'physical' here? It is reasonable to assume that any *a priori* characterisation of what counts as being physical will

fail. In support of this we may cite historical examples. For instance, as Crane and Mellor succinctly point out (1990, p.186):

> [i]n its seventeenth-century form of mechanism ... materialism ... attempted to limit physics *a priori* by requiring matter to be solid, inert, impenetrable and conserved, and to interact deterministically and only on contact,

all of which are requirements that matter, as it is currently conceived, fails to fulfil. Given such examples of disappointing historical performance, we should expect there to be no *a priori* characterisation of what a thing must be like in order for it to count as 'physical'.

Thus Hempel concludes that the term 'physical' must be given meaning *a posteriori*. To find out what all physical things have in common, we must turn to one of the sciences: physics. But which physics? The first horn of Hempel's dilemma criticises the selection of *current* physics for the task of giving content to the term 'physical', as few believe that current physics will survive indefinitely unamended. Physicalism anchored in present day physics is almost certainly false.

The second horn criticises *future* physics as an option for giving the term content. In the absence of a prior, determinate account of what 'physical' means, the claim is that there is no principled way of demarcating entities that are permitted to figure in this future science from entities that are not permitted. Then *any* entity can be admitted to future physics.

If any entity can be admitted to future physics, the claim that what counts as 'physical' is *defined* by featuring in future physics certainly renders the thesis that everything is 'physical' true, but it does so only because it renders it *vacuous*. Any entity needed in the explanation of any other thing will be a 'physical' entity, by these lights. In the absence of a determinate account of the subject-matter of 'physics', properly so-called under our *current* meanings, the

term 'physics' would then be defined in effectively sociological terms, as whatever people in the future *call* 'physics'. Hence the doctrine of 'physicalism' would become transformed into the sociological thesis described by Chomsky, the thesis that when we understand how the mind interacts with the body then the mind will be *called* 'physical'. That thesis may be true, but it cuts no ice against dualism.

The upshot of Hempel's dilemma would seem to be the no-issue theory: the doctrine of physicalism is either almost certainly false, or else vacuously true. There is no *substantive* question at issue between materialists and dualists. This no-issue theory has been vigorously defended by Crane and Mellor (1990).

The spectrum-view put forth in this paper shares important aspects of the no-issue view, but is also importantly different from it. Both the no-issue view and the present view pick up on the same point: physical science will change. Both agree that unless you say *how much* your science is allowed to change without being considered a *different* science, then there is no genuine content to the claim that some future version of *this very science* will account for the mind as well as the body. The no-issue theory argues that no principled measure of how much the science is allowed to change without change of subject-matter can be given in advance and that, since there is no other way to give content to the debate between monism and dualism, the question of whether monism or dualism is true is vacuous.

In contrast, we argue that there *is* a way of giving content to the claim that science can only change *so* much without change of subject-matter, according to our current conceptions: by reference to a historical example. It is not a foolproof way, to be sure, for there is still room for disagreement. However, to acknowledge that disagreement may persist is *not at all* to say that no progress has been made, for agreement may be reached, too. If agreement were reached we would agree both on the content of the claim being made and on who is more of a dualist than whom.

A.7 Present-Day Physics

We have contrasted our spectrum-theory, that dualism comes in degrees, with other real-issue theories and with the no-issue theory. We will now attempt to shed light on the spectrum-theory by considering an alternative real-issue theory that belongs toward one end of our proposed spectrum.

According to the theory that dualism comes in degrees we might think of the scale as limited at one extreme by the thesis that aspects of the mind are *permanently* out of reach of scientific study, and at the other extreme by the thesis that current physics already supplies *all* the necessary tools for the explanation of all aspects of the mind.

Thus, one way to contest the no-issue thesis and respond to Hempel's dilemma, is by defending a definition of the physical that ties it unashamedly to present-day physics. To illustrate this latter, quite extreme thesis, consider the following remarks from Smart, in response to the no-issue challenge:

> My own reply to this charge of emptiness is as follows: for the purposes of philosophy of biology and the philosophy of mind we can tie 'physicalism' to the principles of *present day* physics. ... I concede that there are sure to be revolutionary changes in physics, but I deny that such changes are likely to be relevant to the philosophical problem about mind and its relation to the physical world (1978, p. 339).
>
> There will be no important difference which is relevant to the physicalistic theory of mind, because neurons are 'ordinary matter', and the physics of ordinary matter is essentially complete and is unlikely to change in important respects (1978, p. 340).

The position here is that any explanation of the mind that would be available from a future physics is already available from the standpoint of present-day physics, because none of the things that will change in physics – theories of 'dark matter' or whatever – will be involved in the explanations of how the ordinary matter in neurons gives rise to (or constitutes, or *is*) the mind. Future

science might introduce exciting new chapters; but all these new chapters will only vindicate the already existing chapters on the sodium pump and the firing of neurons, and other such behaviour of ordinary matter.

Smart's position is *not* that science is *already explaining* the mind. Evidently that is not the case – there is much about the mind we do not yet understand – so that claim would not be interesting.

Moreover, the claim that all the resources *physics* needs to supply have been provided is, on its own, not yet the view that science *as a whole* possesses all the resources required to explain the mind, or that all that is left to do is to draw inferences and work out details. To reach *that* position a further claim is needed: that no additional resources will be required from the non-physical sciences. Such a view – which Smart espouses – belongings at one extreme end of our scale.

A benefit of the view defended in this paper is that it can accommodate the many views that exist in the close vicinity of the extreme view. It is possible to claim that having all the tools from physics *just is* having all the necessary tools *simpliciter*, but it is also possible to resist that conclusion. One might claim, for example, that in addition to the tools supplied by physics some higher-level conceptual tools will be needed as well. If some concepts from psychology cannot be restated in the language of physics without loss, and if explanations in distinctively psychological terms are necessary to fully explain the mind, then it follows that we will need further conceptual tools, tools, let us suppose, that only psychology can supply us with, and tools we may well not yet possess.

Suppose, then, that physics already supplies all the tools *it* needs for the explanation of physical phenomena, and that a theory of the mind needs to posit no new entities, but that a theory of the mind *does* require new conceptual tools of some kind. In that case the resulting theory clearly does not belong at

the *very* end of the spectrum stretching from extreme materialism to extreme dualism. Exactly where it does belong may be the object of some disagreement. And there are other complications in the vicinity. Consider, for example, a view which combines Smart's position on physics with the view that explanations from neuroscience are *in principle* re-stateable in the language of physics, but also with the view that any explanation of the mind *that humans could understand* would have to be carried out using conceptual resources that lie outside those required in physics itself. (As an illustration, suppose that physics never needs to deploy the concept of a *weed*, but that psychology does, when dealing with the psychology of gardeners.) Where such a view should be placed on the spectrum would depend on how different the 'extra-physical' conceptual resources were from the conceptual resources required by physics itself.

(Note that we do not claim that *all* conceptual changes one might think necessary to account for the mind would push a view toward the dualist end of the spectrum; only that *some* conceptual changes would do so. A change from an '*is*' of identity to an '*is*' of *constitution* is a good example of a conceptual change that in our opinion would transform a materialist 'identity theory' into something that is arguably relatively close to some historically identifiable theories that would normally be regarded as versions of dualism.)

How a view should be classified depends on the magnitude of changes of *all* kinds that are seen as necessary for the explanation of the mind: ontological and conceptual. The scale-view accommodates the many views that result from this complexity and recommends the framework of historical examples for discussion about how the views compare.

To illustrate further, we apply our spectrum-theory to three salient theories in the philosophy of mind. Compare Smart's position, at one end of the spectrum, with the views of two different dualist positions: the interactionist dualism of Descartes, and the epiphenomenalism of the early Jackson.

Descartes appeared to be relatively optimistic about the truth and completeness of the physics of his own day. He appeared to be confident that the physical explanations for the motions of material objects was virtually complete, and would not need to change very much in the future. In this respect, his view was a counterpart of Smart's view on the near-truth of current physics. Yet, unlike Smart, Descartes held that physics stood in need of very extensive *supplementation* by a theory about the mind. So on our spectrum-theory, Descartes would be judged to be situated toward the dualist end of the spectrum. That is as it should be.

Now consider the Jackson of "Epiphenomal qualia" (1982). Jackson appeared to endorse (at least for the sake of the argument) something close to the opinion of Smart concerning the near-truth of current physics. Yet he argued that full knowledge would require a significant *supplementation* of current physics. This supplementation, however, would consist in coming to know "what it is like" to have various experiences; and acquisition of this kind of knowledge would not involve anything like the articulation of the sorts of complicated theories Descartes envisaged. Thus, on our spectrum theory, the early Jackson would be judged to be quite a bit of a dualist, but not quite as much of a dualist as Descartes. That classification is, we submit, appropriate.

A.8 Avoiding Vacuity

The majority of responses to the challenge posed by Hempel's dilemma do admit, unlike Smart, that present day physics is likely to change a great deal. Furthermore, most of them do admit that some of those necessary changes or supplementations in physical theories are likely to be relevant to any future attempts to explain the place of the mind in the material world. Thus, most real-issue theorists who respond to Hempel's challenge deliberately avoid Smart's tactic of grappling with the first horn of the dilemma. They therefore face the second horn of the dilemma. In consequence, defenders of real-issue theories

must (and do) take what is effectively the strategy that is discussed in this paper: to admit that physics may change in some respects, but – in the case of self-identified materialists – to say that it will not change *very much*.

This, however, leaves the position vulnerable to the no-issue allegation of vacuity. How odd does a new posit have to appear to us, for it to no longer be countenanced as a newly-discovered but nevertheless 'physical' feature of the world? When physicalists defend their theory against dualism by, in effect, saying that current science will not have to change or be supplemented *very significantly* in order to explain the mind, their contention is indeed lacking content in the absence of any indication of what sorts of changes the would count as "very significant" ones.

This is where both dualist and materialist real-issue theorists can gain assistance from the strategy recommended in this paper. Consider what sort of theory would be required for the explanation of thoughts, or feelings, or rational deliberation, or whatever. Ask yourself whether your expectation is that the articulation of such a theory would require *very large* changes in, or supplementations of, current physical science. This is vague; so search the history of science for an example of a change such that your expectation is that we would need to change current science *that much* in order to explain the mental. Now your thesis has real content.

After following the procedure just recommended, it may turn out that one philosopher is, as you might say, a "before-and-after-action-at-a-distance" kind of dualist. Another might be closer to a "before-and-after-continental-drift" kind of dualist. And there are numerous other positions. There is still room for disagreement, and plenty of it, about how various "before-and-after" scenarios compare for magnitude. It will, however, very likely turn out that the majority of positions on the mind-body problem are dualist to *some* extent. We have a principled way of describing the differences between them.

And not only that: as we claimed above, our way of thinking about dualism leads to interesting consequences. The first was that it suggests new questions, like "How much of a dualist are you?" The second is that it may turn out that some views are closer to each other than we thought they were. For, after having gone through the process we suggest it may indeed turn out that some philosophers that previously self-identified as being on different sides of the divide between monism and dualism (conceived as a dichotomy) now point to *the very same transition* in scientific history as the exemplar of a change of the magnitude they think necessary for a science to emerge that can account for the mind. Of course, that may mean that they disagree about how big that past change actually was. Inasmuch as they do we should, perhaps, say that some disagreement persists between them. We think that there is an interesting sense, too, however, in which the disagreement on this issue would then have been resolved.

A.9 Concluding Remarks

Why has the traditional dichotomy between dualism and monism been presupposed for so long? One likely explanation is that monists have just assumed that the similarities between today's science and that which they think will be able to account for the mind will be much deeper and more significant than the differences. Perhaps they have assumed that the new science will be much more like current science than it is like ancient superstitions. Perhaps they think, as they might put it, that we are 'nearly there' – and take their opponents to hold the mistaken belief that we are not even close, or even that we are on completely the wrong track.

On the other side, perhaps the dualists assume that monists downplay the amount of work or change that will be required and the magnitude and significance of the developments that will be involved before we arrive at a theory that explains the mind.

Monists and dualists alike may be in for some surprises if they take up our suggestion for a framework for discussion. Perhaps some monists will be surprised to see how non-superstitious many dualists are; and perhaps some dualists will be surprised to see that many monists acknowledge that much change will be needed before science can account for the mind. Both would be welcome outcomes.

References

Alston, William. 1971. Varieties of Privileged Access. *American Philosophical Quarterly* 8 (3):223-41.

Armstrong, D. M. 1968. The Headless Woman Illusion and the Defence of Materialism. *Analysis* 29:48-49.

Averill, Edvard, and B.F. Keating. 1981. Does Interactionism Violate a Law of Classical Physics? *Mind* 90:102-07.

Ayer, A. J. 1956/1958. *The Problem of Knowledge*. London: Macmillian.

Benacerraf, Paul. 1973. Mathematical Truth. *The Journal of Philosophy* 70 (19):661-79.

Bigelow, John, and Ole Koksvik. Unpublished. Dualism by Degrees.

Bishop, Robert C. 2006. The hidden premiss in the causal argument for physicalism. *Analysis* 66 (1):44-52.

Broad, C. D. 1925. *The Mind and its Place in Nature*. Edited by C. K. Ogden, *International Library of Psychology, Philosophy and Scientific Method*. London: Routledge & Kegan Paul.

Burns, Linda. 1986. Vagueness and Coherence. *Synthese* 68 (3):487-513.

Campbell, Keith. 1970/1980. *Body and Mind*. Notre Dame: University of Notre Dame Press.

Chalmers, David J. 1996. *The Conscious Mind - In Search of a Fundamental Theory*. Oxford; New York: Oxford University Press.

———. 2002. Consciousness and its Place in Nature. In *Philosophy of Mind: Classical and Contemporary Readings*, edited by D. J. Chalmers. Oxford: Oxford University Press.

———. 2003a. The Content and Epistemology of Phenomenal Belief. In *Consciousness: New Philosophical Perspectives*, edited by Q. Smith and A. Jokic. Oxford; New York: Oxford University Press.

———. 2003b. The Matrix as Metaphysics. Available from http://whatisthematrix.warnerbros.com/rl_cmp/phi.html. Accessed 08.08.06.

———. 2006. Re: Conceptual Problem. (Personal Communication: Electronic Mail). Melbourne, 28.08.2006.

Chomsky, Noam. 1968. *Language and Mind*. New York: Harcourt, Brace & World.

Cornman, James W. 1978. A Nonreductive Identity Thesis About Mind and Body. In *Reason and Responsibility: Readings in Some Basic Problems of Philosophy*, edited by J. Feinberg. Encino, USA: Dickenson Publishing.

Crane, Tim, and D.H. Mellor. 1990. There is No Question of Physicalism. *Mind* 99:185-206.

Davidson, Donald. 1974. On the Very Idea of a Conceptual Scheme. *Proceedings and Addresses of the American Philosophical Association* 47:5-20.

Descartes. 1649/1985. The Passions of the Soul. In *The Philosophical Writings of Descartes*. vol. 1. Cambridge: Cambridge University Press.

Dieks, D. 1986. Physics and the Direction of Causation. *Erkenntnis* 25 (1):85-110.

Dowe, Phil. 2000. *Phyiscal Caustion*. Edited by B. Skyrms, *Cambridge Studies in Probability, Induction and Decision Theory*. Cambridge: Cambridge University Press.

Dunne, J.W. 1927/1958. *An Experiment with Time*. Third ed. London: Faber and Faber.

Eccles, John C. 1986. Do Mental Events Cause Neural Events Analogously to the Probability Fields of Quantum Mechanics? *Proceedings of the Royal Society of London* B227:411-28.

Elitzur, Avshalom. 1989. Consciousness and the incompleteness of the physical explanation of behaviour. *Journal of Mind and Behavior* 10:1-20.

Fodor, Jerry A. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Edited by M. A. Boden, *Explorations in Cognitive Science*. Cambridge, MA: MIT Press.

Goldstein, Herbert. 1980. *Classical mechanics*. Reading, USA: Addison-Wesley.

Halliday, David, Robert Resnick, and Jearl Walker. 1997. *Fundamentals of Physics: Extended*. Fifth ed. New York: John Wiley & Sons.

Hempel, Carl G. 1980. Comments on Goodman's Ways of Worldmaking. *Synthese* 45:193-99.

Jackson, Frank. 1980. Interactionism Revived? *Philosophy of the Social Sciences* 10:316-23.

———. 1982. Epiphenomenal Qualia. *Philosophical Quarterly* 32:127-36.

———. 1986. What Mary Didn't Know. *The Journal of Philosophy* 83:291-95.

———. 1998. *From metaphysics to ethics : a defence of conceptual analysis*. Oxford: Oxford University Press.

Jackson, Frank, and David Braddon-Mitchell. 1996. *Philosophy of Mind and Cognition*. Oxford; Malden, Massachusetts: Blackwell.

Johnston, Mark. 1992. How to Speak of the Colors. *Philosophical Studies* 68 (3):221-63.

Kim, Jaegwon. 2005. *Physicalism, or Something Near Enough*. Princeton: Princeton University Press.

Kneale, William C. 1959. Broad on Mental Events and Epiphenomenalism. In *The Philosophy of C. D. Broad*. The Library of Living Philosophers, edited by P. A. Schilpp. New York: Tudor.

Larmer, Robert. 1986. Mind-Body Interaction and the Conservation of Energy. *International Philosophical Quarterly* 26:277-86.

Lewis, David. 1966. An Argument for the Identity Theory. *Journal of Philosophy* (63):17-25.

———. 1983. New Work for a Theory of Universal. *Australasian Journal of Philosophy* 61 (4):343-77.

———. 1988/1999. What Experience Teaches. In *Papers In Metaphysics and Epistemology*. Cambridge Studies in Philosophy, edited by E. Sosa, vol. I. Cambridge; New York: Cambridge University Press.

———. 1995. Should a Materialist Believe in Qualia? *Australasian Journal of Philosophy* 73 (1):140-44.

———. 1997. Naming the Colours. *Australasian Journal of Philosophy* 75 (3):325-42.

Lewis, David, and Rae Langton. 1998. Defining 'Intrinsic'. *Philosophy and Phenomenological Research* 58 (2):333-45.

Melnyk, Andrew. 1996. Formulating Physicalism: Two Suggestions. *Synthese* 105:381-407.

Montero, Barbara. Forthcoming. What does the Conservation of Energy Have to Do with Physicalism? *Dialectica* ('Online early').

Montero, Barbara, and David Papineau. 2005. A defence of the *via negativa* argument for physicalism. *Analysis* 63 (5):233-37.

Oppenheim, Paul, and Hilary Putnam. 1958. Unity of Science as a Working Hypothesis. In *Concepts, Theories, and the Mind-Body Problem*, edited by H. Feigl, M. Scriven and G. Maxwell. Minnesota Studies in the Philosophy of Science, vol. II. Minneapolis: University of Minnesota Press.

Oppy, Graham. 2001. Physicalism. *PLI* 12: What is Materialism?:14-32.

Papineau, David. 2000. The Rise of Physicalism. In *The Proper Ambition of Science*, edited by M. W. F. Stone and J. Wolff. London; New York: Routledge.

———. 2002. *Thinking about Consciousness*. Oxford: Oxford University Press.

———. 2006. Re: The Rise of Physicalism and Appendix. (Personal Communication: Electronic Mail). Melbourne, 20.07.2006.

Penrose, R. 1987. Quantum Physics and Conscious Thought. In *Quantum Implications: Essays in Honour of David Bohm*, edited by B. J. Hiley and F. D. Peat. London; New York: Routledge & Kegan Paul.

———. 1989. *The Emperor's New Mind : Concerning Computers, Minds, and the Laws of Physics*. Oxford: Oxford University Press.

———. 2004. *The Road to Reality: A Complete Guide to the Laws of the Universe*. London: Random House.

Pettit, Philip. 1993. A Definition of Physicalism. *Analysis* 53 (4):213-23.

Popper, Karl R., and John C. Eccles. 1977. *The Self and Its Brain*. New York: Springer International.

Quine, W.V. 1969. Natural Kinds. In *Ontological Relativity and Other Essays*. New York, London: Columbia University Press.

Robinson, Denis. 1993. Epiphenomenalism, Laws & Properties. *Philosophical Studies* 69:1-34.

Robinson, William. 2003. Epiphenomenalism. In *The Stanford Encyclopedia of Philosophy (Spring 2003 Edition)*, edited by E. N. Zalta. Available from http://plato.stanford.edu/archives/spr2003/entries/epiphenomenalism/. Accessed 16.07.2006.

Russell, Bertrand. 1912/1967. *The Problems of Philosophy*. Edited by M. Abercrombie and A. D. Woozley, *Oxford Paperbacks University Series*. London: Oxford University Press.

Schrödinger, E. 1952. Are there quantum jumps? *British Journal for the Philosophy of Science* 3:109-23 and 233-42.

Shaffer, Jerome A. 1965. Recent Work on the Mind-Body Problem. *American Philosophical Quarterly* 2 (2):81-104.

Shoemaker, Sydney. 1975. Functionalism and Qualia. *Philosophical Studies* 27:291-315.

Smart, J.J.C. 1978. The Content of Physicalism. *The Philosophical Quarterly* 28:339-41.

Spurrett, David, and David Papineau. 1999. A Note on the Completeness of 'Physics'. *Analysis* 59 (1):25-29.

Strawson, Galen. 1989. Red and 'Red'. *Synthese* 78 (2):193-232.

Taylor, Richard. 1963. *Metaphysics*. Edited by E. Beardsley and M. Beardsley, *Foundations of Philosophy*. Englewood Cliffs: Prentice-Hall.

Warren, Virginia L. 1986. Guidelines for the Nonsexist Use of Language. *Proceedings and Addresses of the American Philosophical Association* 59 (3):471-84.