The Knowledge Argument and Howard Robinson's 2016 case for Dualism and Mental Substance.

MRes Dissertation September 2020 (with minor corrections April 2021)

Daniel Maille UWTSD danielmaille9@gmail.com

DECLARATION

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed .....Daniel Maille..... (candidate)

Date .....29.9.2020.....

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used the extent and nature of the correction is clearly marked in a footnote(s). Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed .....Daniel Maille..... (candidate)

Date .....29.9.2020.....

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed .....Daniel Maille..... (candidate)

Date .....29.9.2020.....

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available for deposit in the University's digital repository.

Signed .....Daniel Maille..... (candidate)

Date .....29.9.2020.....

ABSTRACT.

Two thirds of this dissertation will consider standard physicalist responses to Frank Jackson's Knowledge Argument (1982) and argues they are unsatisfactory. The remaining third objects to arguments for dualism by Howard Robinson in his 2016 book: "*From the Knowledge Argument to Mental Substance: Resurrecting the Mind",* covering topics of scientific reduction, supervenience, and Robinson's thesis of 'conceptualism' regarding the mind dependence of various entities.

CONTENTS.

INTRODUCTION.

This dissertation is about the mind-body problem, specifically the *knowledge argument* which challenges the thesis of *physicalism,* which is a mainstream metaphysical thesis in western philosophy that the universe and everything in it conforms to the condition of being physical and that all truths are ultimately made true by physical facts. My analysis will follow the structure of a recent book-length treatment of the knowledge argument which also argues in support of *substance dualism*, which is the metaphysical thesis that the universe also contains mental items which do not ultimately conform to that condition of being physical: Howard Robinson, 2016, "*From the Knowledge Argument to Mental Substance: Resurrecting the Mind".*

My primary motivation is not to evaluate the book but rather to mirror the scope and relevance of Robinson's thesis for a program of study. I slant the focus more to issues most alive for philosophers today, and less on historical approaches like behaviourism and analytic functionalism, except where that consideration is more broadly enlightening such as Davidson's anomalous monism. Two very short chapters are effectively editorial references for completeness so the reader may reference those topics I judge least important given limits on word count. The case for substance dualism which I consider does focus on the thesis presented in the book as a contemporary case for dualism from Robinson who is a leading proponent in the field.

THE KNOWLEDGE ARGUMENT.

There are a group of epistemic arguments against physicalism which purport to show that consciousness is not entirely physical, i.e. that consciousness is constituted by something more than the brain and its nervous system. They aim to show that there is an *ontological* gap exhibited by facts of experience and physical facts, founded on an epistemic gap between physical truths and experiential truths.

The Knowledge Argument is a prominent member of that group, of which there are several versions, but Frank Jackson (1982) is credited with the strongest (Nagasawa 2003), though Robinson himself independently

around the same time presents a very close equivalent (Crane 2019). The argument takes the form of a thought experiment and here I will present my own version which is all but equivalent while forestalling unhelpful behaviourist style objections: Mary is a brilliant neuroscientist who has lived her whole life with monochromatic lenses implanted in her eyes at birth so that she can only see in black, white, and grey shade. She has learnt everything there is to know about the physics of colour and the physiology of how the brain perceives colour, indeed *everything* there is to know in physical terms about the processes of colour vision. She knows all the physical information there is to know ranging from spectral wavelengths through to the final neural processes that operate when we see colour, including the precise mechanism that may cause the vocal cords to utter something like, "There is a red tomato". For her fortieth birthday they substitute the chromatic lenses for normal lenses. When the anaesthetic wears off and Mary wakes, she sees a red tomato and experiences colour for the first time. Does she learn anything new? Yes, she learns what it is like to experience red. If this experience presents her with a new fact about colour, then the claim is that *physicalism is false.* The intuition is supposed to be that Mary obtains new factual knowledge about seeing red that she could not have deduced from all the physical information she knew previously. Here is a more rigorous formulation of the intuition by Chalmers (2002):

> "Let P be the complete microphysical truth about the world, and let Q be a truth stating that phenomenal redness is instantiated… Then the initial moral of the knowledge argument is that Q cannot be deduced from P by a priori reasoning. That is, the material conditional 'P ⊃ Q' is not knowable a priori." (p281).

The crux of the argument is the inference from the epistemic claim about colour blind Mary's knowledge to the ontological claim refuting physicalism.

There are various types of objections to the knowledge argument which I address:

- Outright denial that she learns anything new.
- Dispute regarding the a priori link between physical and phenomenal knowledge.
- The phenomenal knowledge from seeing colour is not a new fact 'that' – it is not propositional. Rather it is knowledge that provides Mary a new ability or know *how*.

- The experience provides her with new understanding or representation of facts that she already knew. This can be understood as her coming to learn an a posteriori necessity like in the scientific mode where water is discovered to consist of H_2O.
- The phenomenal-concept-strategy aims to show that her lack of a priori insight is generated by her ignorance of a certain sort of phenomenal concept, which she only acquires upon seeing colour.

Where it makes no substantial difference I will use the usual parlance from Jackson's thought experiment of "Mary before she leaves the room" for her monochrome state, and "outside the room" or suchlike to signify her with no lenses and seeing normally. The balance of attention amongst these responses will reflect my judging the weight of allegiance among working philosophers, for example giving more attention to the phenomenal concept strategy and less to Dennett's proposal about Mary anticipating what colour looks like.

Part One occupies two thirds of the dissertation researching standard physicalist responses to the knowledge argument mainly independent of Robinson, though prompted by his treatment and mirroring his chapter numbers to aid the reader who cares to assess that book in parallel. None of those responses are found to be satisfactory. Then follows a treatment of scientific reduction and supervenience. Part two takes up the remaining third and engages directly with Robinson's arguments for 'conceptualism' which he advances in support of certain categories of mind dependent entities intended to motivate dualism, to which I mainly object. The final chapter occupies just ten percent of the dissertation, reflecting a similar portion Robinson allocates for what I take to be no more than a preliminary case for substance dualism, and again to which I object. For reasons of space it was necessary to bracket and miss out portions of Robinson's treatise which I have indicated, with weight of attention determined by importance and overarching relevance.

CHAPTER ONE. Setting the Scene and Topic Neutrality.

In this opening chapter Robinson provides a logical map for the overarching aspect of his negative thesis against physicalist responses to the knowledge argument. I will mainly lay out those elements of most relevance for my treatment with little argumentation or objection at this point because his entry points are broad, but understanding how Robinson sets up the debate will help with the big picture before I take up an analytic focus in the rest of the dissertation.

Identity theory is presented as the beginning of modern physicalist attempts to relate first-person experience to the physical, and he argues that all attempts since then effectively amount to the same proposal, that the mind and brain are identical. The identity theory, most famously of Smart (1959) and Armstrong (1961) emphasised the notion of *topic neutrality* and this is a key notion running through Robinson's responses to physicalism. The identity is of brain processes and experience, so instances of each are not merely correlated, but are the *same thing.* I was confused about the meaning of topic neutrality when I first encountered identity theory years ago, I guess because it is very counter intuitive; when I dream a vivid image in my mind's eye of an apple, so that in some sense what is going on *in my head* is no different from what happens when I experience a real apple, then my subjective report about the apple can be said to be topic neutral regarding the neural activity and the seeing; neutral as referencing either the neural event or the phenomenal image, because they are identical. The idea is that the correct reference of topic neutral identities could be picked out without ascribing intrinsic properties and could be established empirically. In this way topic neutrality is a conceptual device to bracket the counter intuitiveness that when we refer to a phenomenal image, we are picking out nothing but a physical event or state with only physical properties. Nagel (1974) coined an expression referring to the essential first-person subjective aspect of an experience as there being a, 'what it is like' to a perceiver. We may refer intuitively to the subjective first person impression by 'what it's like', which acts like a broad catch all phrase for whatever appears to the subject, and the topic neutral conception will exclude those *essential properties* of that what-its-like for the observer. Properties such as redness and

brightness and perhaps beauty, which are radically unlike the physical properties constituted by those activated neurons also involved in the experience. I will take advantage of the convenient catch-all nature of the 'what-its-like' expression, and will use it frequently throughout the essay where appropriate (so that it's use does not blur important distinctions or beg the question). However, it does not roll off the tongue so I ask the reader to commit the abbreviation WIL to memory.

Robinson argues that while topic neutrality might seem a tempting answer to the knowledge argument, it doesn't touch the argument and it is important to see why, because he judges that all subsequent physicalist accounts which jettisoned topic neutrality are actually forced back to it and are open to the same or very similar objections. Before we get specifically to the knowledge argument in relation to topic neutrality, Robinson specifies three essential features of every identity theory and how the topic neutral concept is a persistent incoherence in physicalist positions which has been overlooked and neglected:

(i) "Conscious states are identical to neural states.

(ii) Neural states possess only those properties recognised by physical science. There are no 'emergent properties'.

(iii) We are not explicitly aware of those physical neural properties in being aware of our own conscious states." (p.5)

The intrinsic mental state consists only of physical properties, but our first-person mental awareness does not reveal those properties, so first-person knowledge of mental states is said to be topic neutral. Some physicalist theories do not appeal to topic neutrality by arguing for *token-identity* whereby particular instances of mental states are identified with particular physical states, instead of the *type-identity* invoked by Smart and Armstrong whereby experiential mental types are types of brain event. Robinson argues that this difference does not change their commitment to all three features. Robinson characterises other next generation physicalist theories as attempts to ground the experiential aspect of the mental state in some other relation than identity: those citing supervenience, realisation, or functional role. Robinson concedes they might hold some chance of explaining *propositional states* such as standing beliefs or facts in memory

which could be represented by abstractions without phenomenal properties, but not they cannot capture the raw feel component of mental states which are certainly not abstract, and so he will argue they fail to throw off the essential features of topic neutrality.

Robinson narrates the historical interplay of debate between Ryle, Chomsky, Quine, Putnam, and Dennett, which he interprets as attempts to skirt around the identity theory of Smith and Armstrong (and then David Lewis) before the real crux of the matter was noticed by Nagel and then Jackson. He tells us in sweeping terms that scientific method and how we perceptively infer meaning from the world conflicts with the intuition of our being *immediately* aware of sensations and not any supposed physical essence constituting that sensation. I will wait until later chapters for those objections in more detail and argue that his notion of 'conceptualism' which motivates those arguments is unsound.

Robinson spells out a different sort of objection which he believes is conclusive. Those next generation of token identity theories offer an improvement over Armstrong because they specify the identity between the functional role of a brain state and the WIL ('what-its-like'). He takes issue with a common analogy that functionalists use which substitutes brain events with states of computer hardware, because the causal significance of the neuronal state which realises or subvenes the mental event is determined by context, with immediate downstream neurons carrying the behavioural significance of the supervening experience, which they achieve because of their connecting setup. That neural-causal identity cannot be secured because, "behavioural dispositions are, so to speak, long-term dispositions, and not immediate ones, and cannot be identified with a specific central state or its causal output." (p.9) I take his diagnosis to mean that causal role and neural identity cannot configure a synchronic state, because the causal role of a brain process depends on its context, and different downstream neural states lose (or 'forget') the contextual significance of earlier states, so equating causal role with neural identity will not do the work required. I will push back on Robinson's objection because the issue is far from clear cut, as he provides is no *principled* reason why the level of computing complexity could not parallel the complexity of causal relations that is happening at the mental or behavioural level. The subject of computing hardware performing brain functions is a field of research in its

own right, which I will just note as a point of interest for further research, taking Robinson's objection as suggestive but inconclusive.

Robinson targets topic neutrality from a different angle. If Mary knows all the facts, then coming to know something of a general sort in a topic neutral fashion will not add anything knew. A subject having enjoyed an experience knows exactly what that experience is like, or we might say if the experience just is the knowing WIL, then learning something general about that experience will not bring new information about the experience in-itself, and this rules out topic neutrality under writing further information about that event. Robinson illustrates what he means; if you know Fred is in the garden, then being told that *someone* is in the garden will lack any new information for you. Knowing all the physical facts, if physicalism is true, then learning what seeing red is like is an epistemic parallel to be being told that someone is in the garden after you already know who is in the garden. I suspect a faulty a notion of physicalism is allowing the scope for this charge, because it seems to imply that Mary in the room should know *everything* in the room related to the experience of red and physicalism need not entail *all* experiential facts or it isn't clear why knowing the physical realizer identity must be an item of more general knowledge than facts about the WIL. Supported by Van Gulick (2004), I will argue for different ways of knowing an experiential fact which is compatible with physicalism.

Robinson contends that at least one of the three essential criteria for an identity theory which entail topic neutrality must be dropped by the physicalist, and there is only room to drop the first criteria of identifying conscious states with neural states. He thinks any standard form of physicalism which involves sensations as occurrent events will entail topic neutrality.

The rest of the Chapter more fully introduces the knowledge argument with some clarification: Jumping off from my introduction chapter, Frank Jackson includes in the facts which Mary learns before leaving the room "…all the causal, relational,...and functional roles consequent upon all this...otherwise we can suppose there is more to know than every physical fact...and that is what physicalism denies." (1986, p291). Robinson attaches two preliminary points that may help us later: The character of what Mary learns from WIL is the phenomenal feel of an object in the world. She is not aware of only an internal state because that state is a

transparent view of an object in the world; the knowledge argument involves what is termed the *transparency of phenomenal consciousness* (Moore 1922), by which our experience is not of an internal mental object but is of what causes that experience. Secondly, the knowledge argument may lean either on the thought that Mary learns something new on her release or that she lacked knowledge before release. Robinson claims nothing important depends on this distinction, but it will help us avoid potential complications with Dennett and aspects of the ability hypothesis if the formal specification of the argument stands on what she *lacks* while still in the room.

Robinson discusses a couple of arguments related to the knowledge argument (which I will not address: Chalmer's conceivable zombie and Levine's explanatory gap, pp.13-18), which he sees as distracting philosophers from the core of the knowledge argument which is independently sufficient to refute physicalism. All three related arguments together cast an overarching explanatory gap best uncovered by the knowledge argument between physical explanation and mental explanation: *that gap may be interpreted either de re or de dicto; de re is how conscious properties relate to physical properties, and de dicto is how psychological explanations relate to physical explanations.*

Before narrowing on the specific type of objections to the knowledge argument in coming chapters, the core issue they target is what way, if any, coming to experience colour can be framed or characterised as gaining knowledge she had previously lacked. The four types of responses which I mentioned in the introduction and where I will address them:

1. An outright rejection that Mary lacks nothing cognitively before leaving the room. She could work out and anticipate what it is like to see colour. Robinson sees this as the boldest objection, which is most famously advanced by Daniel Dennett, that I will address in Chapter 2.

2. Mary acquires a new ability on her release, which is a gain of know *how,* rather than factual or propositional knowledge *that.* This approach is mainly associated with David Lewis and functionalist theory of Mind. Chapter 3.

3. Mary gains a new way or mode of knowing something. She comes to know phenomenally what she already knew theoretically. This is known in the literature as the phenomenal concept strategy and attracts allot of attention so I will consider different variations of that strategy in chapter 5. The phenomenal concept strategy is aligned with the abilities objection by Frank Jackson himself who changed his mind about the knowledge argument and came to think of the acquired ability outside the room as a capacity to represent a neural state in a peculiar sort of experiential way (chapter 4).

4. Mary lacks but then acquires factual knowledge about the nature of phenomenal colour, which David Lewis calls *phenomenal information*. The new knowledge is not propositional but is advanced as knowledge by *acquaintance*, which is argued to be factual information about the nature of colour and colour experience. Robinson contends that the knowledge by acquaintance interpretation is required to support a thesis of property dualism which he sees as the only plausible response by a physicalist to the knowledge argument and claims that part one of his book will show this.

Property dualism might support the thesis of a non-reductive physicalism, which involves mental properties supervening or depending in some way on a physical base. It is controversial whether property dualism is consistent with a thorough going physicalism, which the chapters of part two will analyse with ideas around scientific reduction, before we then consider the positive case for substance dualism in part three. Like I said in the introduction, I will be mirroring his chapters for ready-made structure and ease of referencing the positions.

Dennett denies that Mary is ignorant of WIL before leaving the room, in a straight denial of the knowledge argument intuition. Robinson accuses Dennett of the "Jericho method": aiming to dissolve the knowledge argument with nothing more than dialectic and no arguments. This position deserves space in the dissertation, but not much, so I have attempted to locate the most pertinent and instructive aspects. Over the years Dennett develops three characterisations of Mary aiming to pump-prime different intuitions. Starting with Mary's blue banana (Dennett 1991) as just an illustration (it doesn't reach the level of intuition) that Mary could work out WIL from the science, then adding substance to how she would achieve that awareness as 'Swamp Mary' and 'Robo Mary' (Dennett 2007).

Dennett wants to make out that if Mary pre-release knew *everything* about the physical processes, then she will know WIL. He tells a story of Mary on her release being shown a blue banana and recognises that she was being a tricked because she already knew that it would normally be yellow. Her observers are bemused about how she could possibly recognise the difference between yellow and blue to know from sight that the banana should not look like that, and that they were fooling her. Also, Mary understands their puzzlement, because she knows everything about colour, yet they do not. Robinson interprets that Dennett could have one of two possible rationales for how this could happen. Either a standing disposition of verbally responding to certain wave lengths of light, "that is yellow", grounded in his functionalist understanding. Or otherwise Dennett is assuming a sort of omniscient awareness of matter, whereby Mary could correlate *particular* brain states with seeing blue, but for that Mary would need to observe states of her own brain when exposed to yellow and blue, which by hypothesis has never happened before.

Dennett effectively accepts that judgement about the blue banana story (2007, fn1), which renders his original article just a blunt denial of the knowledge argument intuition, then attempts to develop his position with two further thought experiments: Swamp Mary and Robo Mary. A Lightening strike hits Mary while she is in a swamp and spontaneously alters her physical constitution into the same state as if she were normally

experiencing red. Robo Mary is a robot in the form of a replicated human with the ability to program herself into that same state which the lightening caused in Swamp Mary. Dennett doesn't put it like this but in effect we are invited to transition from the "…logically possible cosmic accident path…" (2007 p.25) of Swamp Mary seeing red, to the parallel plausibility of Robo Mary deliberately manipulating herself into that same state. To satisfy the knowledge argument we need to know *how* Robo Mary might achieve her 'seeing-red configuration'. Swamp Mary achieves that state by fiat, but the task for Robo Mary is programming herself from the science, which would require her to deduce the WIL state from the facts and that raises afresh the problem targeted by the knowledge argument. Or perhaps we might see Robo Mary as introducing the ill-posed problem of imagining WIL for a robot to see red. Dennett muddies the water (2007 pp25-29) by drawing parallels with original Mary who might acquire the seeing-red state by rubbing her eyes or spontaneously daydreaming or such like and Robo Mary who can engineer herself into that state with interventions all described in pseudo-computer jargon which only beg the question about how. He eventually faces the crux of his task with Robo, "What matters is whether Mary (or Robo Mary) can *deduce* what it's like to see red from her complete physical knowledge, not whether one could use one's physical knowledge in some way or other to acquire knowledge of what it's like to see in colour." (Ibid p.29) He claims that there is no clear separation between deduction and other forms of "knowledgeable self-enlightenment" (Ibid p.29), such that we should baulk at the thought of Robo adjusting herself *and* programming herself deductively from the science. He suggests the adjustment is only comparable to experiencing in the shoes of others through empathy and imagination.

Robinson contends that distinction is irrelevant, and that the crucial difference between RM and colour-blind Mary is the former *putting* herself into a state, in contrast to the task of working out from 3rd person scientific information what that state is like. Dennett's blurring of that gap seems connected to his behaviourist stance, claiming that the newly acquired post-experiential property is *dispositional*. But the question is, disposed for what? A system being disposed to experience red is different from experiencing red. Let's say I am now physically disposed to experience whatever I may encounter, that is the physical aspects of my nervous system

are configured such that they only have to receive a sensation to be activated, the mental state I acquire upon that encounter is something added to the dispositional physical base. I might be accused of begging the question against dispositionalism by simply pleading that the WIL is an extra component to the disposition, but it seems the onus is not on me. Dennett seems to think that it is, “"...the richness we appreciate [WIL], the richness that we rely on to anchor our acts of inner ostension and recognition is composed of and explained by the complex set of dispositional properties..." (2007, p19-20, my parenthesis). Robinson presents nuanced arguments about behaviourism and functionalism missing out qualitative subjective properties (here and more extensively in the next chapter). I will only engage with those arguments where it is particularly helpful because they have been rehearsed ad nauseum over the years and their conclusions are widely accepted.

Robinson aims to show that Dennett’s behavioural cum functionalist stance renders Mary’s scientific omniscience irrelevant by boiling down the problem to MBB and RM supposedly being able to work out mental states from scientific knowledge under certain conditions: but why he asks should that be necessary if there was no difference between knowing how Mary would react verbally or otherwise and knowing WIL.

Before discussing recent philosophy which may bare on Dennett’s notion of a person’s ability to evoke WIL with empathy and imagination, I will just mention Robinson positing of another Mary variation to illustrate Dennett's predicament (2016, pp32-33): 'Extremely Observant Mary' (EOM) has no special scientific knowledge but finely tuned social skills to observe reactions. I will offer a much-enhanced EOM in favour of Dennett which better drives home the point. My EOM enjoys normal life unimpeded apart from her monochrome lenses *and* has the scientific knowledge of original Mary, while those fine social skills are supported by perfect memory for cataloguing all incidents of people reacting to colours. Even my OEM will not make more plausible Dennett's behaviourist intuition that she could bring on the mental disposition to imagine red.

A different way into this could be Mary dreaming the red experience (Dennett 2007 p23), which potentially lends plausibility to the idea of her evoking WIL at will, by voluntarily imagining that same WIL while awake.

Even if we grant for arguments sake that Mary could dream in red, her mental state would not have come about from scientific inference or otherwise from her learning and so the knowledge argument is untouched.

Churchland (1985, p25) connects daydreaming with an ability of Mary to correlate the image with a certain spiking frequency in her visual cortex allowing her to identify the image with a neuroscientific concept. This move runs into the same problem mentioned earlier that she would need to observe her own brain while observing colour which is off limits, while also raising further questions (that I will not consider) about Mary's experience of red in relation to a 'normal' experience of red and colour predicates in public language.

I will offer an argument from my intuition which involves speculation about psychology. Our ability to imagine unfamiliar experiences may depend on what I will call a 'similarity-spectrum': for example, comparing a medium pitch noise we have never heard to be 'in-between' low and high pitch noises on a similarity-spectrum of pitch. It is plausible that the ability to imagine operates with a similarity-spectrum framed by previous experience, and in this case enabling us to imagine a pitch between the low and medium or the high and medium. The philosopher Amy Kind (2019) sets up a similar notion and suggests an imaginative scaffolding that Mary might employ, with which she charitably reads Churchland having in mind as regards Mary's imaginative ability (2019, p174). Kind accepts that such considerations will not be decisive, but that they may help inform the reasonableness of the idea of Mary imagining colour. Kind describes a fictional example from an award-winning novel in which the writer apparently empathises with an autistic boy, and she credits the author with successful imaginative empathy even though it involves "relatively distant imagining" (Ibid p.175). I take 'relatively distant' to mean in my terms that the novelist is leaning on a similarity-spectrum of empathy that can provide a quite limited imagining of the autistic boy's experience of empathy.

Kind presents a real-life example which she suggests might somewhat help imagining WIL to be a bat: Animal scientist Temple Grandin (2006) claims the ability to adopt a perspective of seeing like a cow, which has enabled her to successfully design equipment to handle livestock: "I place myself inside its body and imagine what it experiences. It is the ultimate virtual reality system." (taken from Kind 2019, p.176). I am inclined to

resist Kind's charitable reading of Churchland because there is a total lack of similarity-spectrum in the case of colour, as I aim to illustrate, and which helps resist Dennett's argument that Mary could imagine colour. If the autism case is relatively distant on the empathic similarity-spectrum (for one thing empathy is not quantifiable) and seeing like a cow is relatively close because the altered perspective is simply lower to the ground with a different shaped body, then I suggest colour is altogether different because there is no imaginative-access-scaffolding to make the transition from monochrome to colour, as there is for the difference in height and shape between humans and cows or lowered empathy in the autism case. For this sort of reason Robo Mary cannot bare the intuitive weight that Dennett claims for her in terms of imaginative abilities; whereas one might imagine grey shades in between black and white, I do not begin to fathom how one could imagine, say seeing orange (having never seen it), without the imaginative scaffolding provided from previously experiencing red and yellow. This scheme is somewhat vague and a fuller specification might involve closer attention to different modes of experience, for instance grading empathy would involve altogether different considerations to the more simple comparison of height, and colour relations will be different again, but I trust I have provided sufficient gist for my point.

Given the affinity between *dispositions* and *abilities*, these considerations may more directly concern the 'ability hypothesis' which Robinson claims is Dennett's fallback position and is the topic of the next chapter.

Chapter 3. The Ability Hypothesis.

An established tradition inaugurated by Gilbert Ryle (1949) finds a distinction between two different sorts of knowledge. Propositional facts which science targets as knowing *that*, in contrast to learning an ability which involves non-factual knowledge and learning of *how*. The Abilities Hypothesis is an objection to the knowledge argument which frames Mary first colour experience as learning a new *skill or ability* by way of practical knowledge, rather than a new propositional *fact* as premised in the knowledge argument. The hypothesis is that Mary's first experience of colour enables her to recognise, imagine, and judge resemblances of colour. (Though this distinction is commonly observed in modern analytical philosophy, it is not universally accepted, for example Stanley & Williamson 2001 argue that the former subsumes that latter as a species). After outlining Robinson's position, I will present David Lewis and Nemirow's ability hypothesis interspersed with arguments from others, and reject the hypothesis while gaining valuable insight about Mary.

Robinson detects the ability hypothesis in some way conflating Mary's having the *capacity* to see red, with Mary having the *experience* of seeing red. He would say Mary pre-release was always disposed with the capacity which enables her to see red, and in that sense already had the ability simply was not exercised before leaving the room. Robinson provides some historical context, explaining that Ryle's behaviourism was quite obscure regarding the analysis of sensations (p.37), but he never aimed to establish dispositionalism so generally as to eliminate the aspect of inner subjective experience that I have bracketed as the WIL (final reminder, Nagel's 'what-its-like'). The ability hypothesis compliments a functionalist account of mind that equates the inner experience event with a disposition, and Robinson argues that it fails. I will first lay out Robinson's understanding before presenting my take on Lewis.

Robinson takes Lewis (2004) as attempting to explain how Mary is relieved of her ignorance of what it is like to see colour, and Robinson charges him of mistaking what an ability hypothesis is doing, essentially by conflating the *ability or capacity to recognise* red with her *knowing* before she leaves the room what red looks like. Robinson rehearses age-old claims (pp 38-43) to show that Lewis' functionalism involves a failed attempt

to explain away the qualitative aspect of our experience, while an ability account of Mary's experience *must* feature her having experienced what red looks like, or else her brain would only provide the unexercised causal capacity allowing her to discriminate colour – exercised when she first sees it. Robinson reads Lewis (1983) as failing to explain why the phenomenal aspect in itself contributes nothing to the operation of the functional state, while at the same time relying on the 'feel' for the ability to imagine or "reactivate that same state" (p.40). This ambiguity invites Robinson to charge Lewis with those same objections against topic neutrality; because apparently Mary's functioning results not from properties that she discriminates in qualia, but from those physical properties of her brain and environment. A purely functional account which renders WIL as causally impotent renders Mary's new experience as just an activation of potential mental pathways. If Lewis' account is correct, then a person would not learn any new facts by way of the qualitative aspect of a new experience, and that premise in the knowledge argument could be rejected. Robinson quotes several examples of Lewis referring explicitly to first-person sensations within his argument, which are inconsistent with his theory and uncovers the incoherence of identifying sensations with functional brain states.

Before I take up Lewis, I want to mention an interesting attempt by Phillip Pettit (2004) to merge a functionalist analysis of qualia with abilities. Pettit presents 'motion-blind' Mary as not experiencing motion while in captivity, achieved with a clever computer-generated strobe light set up to compensate for actual movement so that everything appears as though static. Pettit wants to compare why the visual experiencing of motion may not evoke the same intuition as the knowledge argument does with colour, that may help explain why our colour-Mary intuition is so strong and mistaken, and show that both Mary's can know before release what their outside experience will be like.

Pettit tells us the following: Motion-blind Mary can calculate every position of a moving ball while inside the room, and her beliefs based on those calculations remain unchanged when she leaves the room to actually see that movement. Then those beliefs come to serve her functionally disposed brain states which enable her to catch the ball. These same facts about the ball are satisfied by a different belief profile in virtue of different causal bases and the only change in her knowledge is a "shift in her perceptual and inferential

abilities" (p.116) which we are told is just a new way of knowing those same facts that she knew in the static room. In this way she is just exercising an ability in relation to those facts; using practical know-how from what she previously knew only intellectually.

Another way Pettit frames the thought experiment has Mary leaving the room to virtually position herself within the network of possibilities all grounded in the facts she previously knew about motion (p118). Pettit then presents as a parallel, colour-Mary resituating herself amongst the colour facts like motion Mary did with the facts of velocity, angular momentum, and such like. Pettit's essay is dense, but his argument seems undercut by a difference in the perceived objects which will not support the parallel he draws which depends on some equivalent element in each case. We have no intuitive problem thinking of motion in entirely physical terms, or completely described using only physical properties. The same is not the case for colour which is not fully described by physical properties (at least that is our strong intuition); whereas motion can be exhaustively specified by location relations against a background datum; colour by hypothesis involves some datum bound up in the perceiver. A different way of illustrating this thought is that while motion Mary's correct beliefs about position and motion inside the room might deductively generate more belief about motion for her outside the room, the beliefs which colour-Mary acquires when she exits share no logical equivalence with her previous monochrome beliefs. I might detect how Pettit would resist my undercutting, where he claims that if the knowledge argument is to have, "…any point, it has to be assumed that colour is a physically analysable property that is detectable by human beings in a physically analysable way; otherwise physicalism would be undermined before the argument ever gets going." (2004, p126). However, I understand the knowledge argument as *challenging* those physicalist assumptions contained in this sentence, not that it need assume them to get off the ground. Perhaps he means that the knowledge argument intuition rides on the disputable idea that colour is *essentially* an experience which would rule out physicalism by definition.

The last point might be in play where I detect Pettit and I maybe talking past each other by entering the knowledge argument from a different intuitive spring board: He says something very curious: "…Mary knows

all the objectual properties of colour, including the colour of this or that object, since she knows all the physical facts,…and *all corresponding intentional facts to do with which colour experiences are experiences of which colours*." (p.126, my emphasis). Two related things to say here. Pettit seems to be making the same move that Dennett made with the blue banana: conflating scientific knowledge with something approaching physical omniscience, because by hypothesis, monochrome-lense-Mary has no way of correlating intentional facts with colour experiences. Even if we allow her a sort of maximal scientific access so that she could reference exhaustive mapping of brain states with intentional states in limitless human specimens, she still will not know what red looks like. Secondly, Pettit illicitly includes intentional belief states as physical brain states by assuming they can be specified with scientific facts (or identified by exhaustive experiment in monochrome), because there is good reason to think that intentional mental events are partly constituted by phenomenal states (see, generally, Mendelovici 2018, ch3), and so the knowledge argument intuition arises again and around we go.

Lewis (1988 in 2004) is clear that having an experience is the best way of learning what that experience is like in a way we cannot be told about, but that this tells us nothing about the metaphysics of the mind or how science is limited. He posits the logical possibility of an internal change made to our brains by surgery or magic (ala Dennett's Swamp Mary), which brings about the same brain state that an experience would. Lewis imagines the possibility of a science lesson providing the same result as the surgical intervention, albeit with a method we cannot fathom yet. He mentions the imaginative access we have to new experiences made up from elements of old ones, and uses an example of a musician who imagines 'hearing' a musical piece from reading its score. His thought introduces what I called 'imaginative-distance' in chapter two, because the musician is in some way piecing together separate memories involving resemblances previously experienced. I said there how I could not begin to comprehend the first step of imagining WIL for Nagel's bat, and Lewis agrees that we cannot imagine potential experiences if they bear no resemblance to past episodes even in conjunction with science lessons, but he wonders, "…how new is "new enough?"." (p265) to render them in principle off limits.

He offers a further clarification to help us with Mary by introducing another sort of knowledge distinct from knowing *that* or knowing *how*: Knowledge *de se.* This kind of knowing essentially involves the bearer of that knowledge, with self-involving facts only true for that bearer, which may include information only relevant from the bearer's location in time and space. I take it that he draws this distinction as a way we might miss the point about what Mary learns the first time she sees red; in case we are thinking of it in the sense of a de se proposition because logically that state of hers is only achieved *when* she first experiences red. Lewis is warning us not to be "bewitched" by the first-person perspective (p.269) in conflating Marys de se proposition 'I am now seeing green', with her learning an objective fact about what green looks like. Perhaps I am because bewitched because I am sure not to be conflating the experience referred to as 'that is what green looks like' with a purported objective fact about green, or with the proposition "I am now seeing green".

Paul (2017) might help clarify what Lewis takes from the de se distinction; Mary discovers a new perceptual *truth* about seeing green with the experience of seeing green, and that truth could only be learnt with experience. It is the physical fact that she already knew about green seeing brain states, but presented in a distinctive experiential way. This means any number of truths can be grounded in the way green looks, but those truths need not be primitive facts about green and hence Mary doesn't learn a new fact about green, but just a new de se truth grounded in facts she already knew. (This understanding might imply that Robinson is thinking of facts as items which can be irreducibly perspectival, and on the face of it we would not expect physicalism to be refuted because it does not range over an ontology containing such perspectival facts because it offends an inclination to parsimony and simple ontology. Thinking out loud for now, a dualist ontology might be simpler overall if it conceptually packaged that super abundance of perspectival facts in an elegant way, but I get ahead of myself. We will see in future chapters how Robinson comes at parsimony from a different perspective).

Paul talks about the brain physically *realizing* the experience that Mary is having, which is different from Lewis who identifies the brain state as *constituting* the experience. I will investigate throughout chapters six

to eight how a physicalist account might explain a realizing relationship and whether it collapses into a Lewis style identity, as Robinson argues that it does.

Lewis formulates the Hypothesis of Phenomenal Information to which he says the knowledge arguments is committed:

> "That is the hypothesis that besides physical information there is an irreducibly different kind of information to be had: *phenomenal information*. The two are independent. Two possible cases might be exactly alike physically, yet differ phenomenally. When we get physical information we narrow down the physical possibilities, and perhaps we narrow them down all the way to one, but we leave open a range of phenomenal possibilities. When we have an experience, on the other hand, we acquire phenomenal information; possibilities previously open are eliminated; and that is what it is to learn what the experience is like." (1988, p271).

Phenomenal information involves the information an observer receives from the qualia or WIL of an experience and is irreducible and independent from physical information, being intrinsic to the experience itself, or an aspect or facet of an experience. Lewis argues that the knowledge argument rides on this notion of phenomenal information which is only received in an experience. If the hypothesis is false and there is no phenomenal information, there would still exist WIL to have the experience but no information about it would be phenomenal and all information about the experience would be physical, and could be communicated by science without undergoing the experience. Lewis says If we grant the hypothesis then the knowledge argument does refute materialism, and while he cannot disprove the hypothesis, he will argue that it is unappealing in conjunction with the ability hypothesis which "…does justice to the way experience best teaches us what it's like." (p275). I directly quote these few simple words because they succinctly range over the crux of the matter, and in a sense they encode for his position because Robinson would say experience *is* what it's like, and not that experience teaches us what it's like. Am I noticing here that either Lewis is tipping the scales at the outset or perhaps indicating a different project? I suppose we need not worry so long as he can explain away the projected intuition from the knowledge argument.

Lewis continues to interrogate the nature of phenomenal information by raising three different analogies (p.278-283) designed to illustrate it's strange and implausible character, such that a single state of affairs like

Mary seeing a red apple, can potentially generate an unlimited amount of phenomenal information, because we might think of Mary possibly occupying unlimited new perspectives by seeing the apple under different light from different angles and distance. But Mary would not learn of any new possibility regarding the nature of the apple beyond a limited number of observations. He is driving home the idea that phenomenal information is only constrained by endless possible perspectives to be had, yet there is only a single state-of-affairs informing that endless quantity of information. Lewis' intuition is offended by the peculiarity of thinking of a potentially endless array of information providing any new knowledge about the apple. This connects to my thought earlier when I wondered if Robinson was committed to a super abundance of perspectival facts and how they might fit within either a physical or dualist ontology.

Lewis tries a different approach, arguing that the hypothesis of phenomenal information works against the dualist intuition as well, because if we posited a science of parapsychology describing all the nonphysical causes, laws and properties, Mary could not learn about them in the room. She must also *experience* what they are like, because phenomenal information is independent of any informative information about the world, and not just physical information about the world. This is a curious challenge, in support of which the dualist should expect Lewis to prime the same intuition against parapsychology by saying something about the properties it studies and propositions that it generates, which we could then judge against the intuition we have about neuroscience and colour vision. We have a handle on physics enough to prime the intuition against physicalism, but no handle at all on the science of parapsychology to do the same. Lewis' position so far amounts to saying that phenomenal information, if it existed, would potentially ground unlimited unactualized experiences, which he finds altogether peculiar enough for him to be sure something is wrong with the notion.

A different kind of objection-from-peculiarity that Lewis raises in the penultimate section of the paper is titled, "From phenomenal to epiphenomenal". (p.280): phenomenal information is isolated from its supposed effects on the world because it is by hypothesis independent of physical information. His idea is that Mary might behave exactly like we expect her to when she leaves the room while seeing green instead of red. In

terms of the phenomenal information at play, the difference between seeing green and seeing red would make no physical difference to the effects on her brain. This is effectively the old famous objection of interactionism founded by Princess Elizabeth against Descartes. The difference with the knowledge argument however is that no separate mental stuff is mentioned in the premises, which only concern Mary's inability to access what Lewis is calling phenomenal information from inside the room, so Lewis is somewhat begging the question against the knowledge argument, or at least over relying on the hypothesis which he is arguing for. He connects this to the related issue of physical causal closure, which is an assumption made in physics that every physical event has a complete physical cause, leaving nothing for a non-physical stuff like phenomenal information to do. Betting against physics he says is a bad move, and a much safer bet on phenomenal information should consider it epiphenomenal, that is, causally impotent on physical stuff at least. The causal closure argument is a strong one for physicalism, after all physics is very successful. Robinson will offer independent argument for dualism later, but here he simply says that Lewis has no right to object to interactionism because he has not offered a physical explanation of Mary being 'surprised' by the phenomenal information that portrays colour.

Lewis presents another argument for epiphenomenalism without leaning on the authority of physics, involving apparatus of dependence and independence (pp285-287), which he takes to show the essential nature of qualia causing the same physical effect regardless of the colour presented to Mary. Robinson argues (p46) that this objection works equally against Lewis who would accept that numerically different neurons in the brain could achieve exactly the same mental event, because an item still has an effect just in case another item may have performed the same role. However it is not clear that Robinson can make this move because he requires the character of redness to be mentally distinct from greenness to the extent that it makes a different impression on Mary, while Lewis needn't implicate any distinctive physical properties to the individually identical neurons which realise the brain state. As far as Lewis is committed, any set of neurons with the right configuration can do the job, but Robinson requires specific sorts of phenomenal information to entertain Mary with different colours.

Lewis finally comes to the actual ability hypothesis by quoting a summary from Laurence Nemirov (1980, p.475-6). Although Lewis and Nemirov certainly unsettle the notion that there is a subjective fact denoted by 'what it's like to see red', they do not refute the idea of phenomenal information. After next rounding out our reading of Lewis, I will jump to a much-updated defence of the ability hypothesis from Nemirov (2007), which is not mentioned by Robinson.

Lewis ends his paper with further mention of the de se distinction, which I understand to mean Mary only comes to learn perspectival truths related to the truth that *she has experienced seeing red,* such as she can remember seeing red and she can visualise or imagine red, and recognise an experience of seeing red. I do not see this helping us describe Mary's situation; there is something special about the experience of colour phenomena, that while enabling imagination, memory, and judgements of resemblance, is not captured by de se truth. It will be surprising for the man spilling salt all around the shop when after following the trail he comes to realise it is *himself* who is spilling the salt, and *that* new item of knowledge was all that was missing about a drama he could know *everything else* about. In attempt to reduce any inclination to being fooled by the ambiguity of the term *knowledge* as Lewis warns against (p.289), I will frame the comparison with the looser term of 'awareness': seeing colour for the first time is not much similar to the messy shopper, because the item that enters awareness *is the raw feel of the experience,* not just the knowledge of self-involvement in an otherwise fully understood situation.

I will interject my resistance to Lewis with a paragraph now defending him from an objection by Rosenthal (2019, p.40), who points out that Mary's first experience might not provide any *ability* for recognising or imagining red because she may not afterwards remember anything about its character, and so she cannot be said to have acquired any of Lewis' cited abilities from the experience. Rosenthal may be setting the bar too high, because it is not unusual to forget an ability and cease to enjoy a skill while retaining the *disposition* acquired to support that ability or part of the disposition which is why riding a bike is much easier after a five years break than when first getting on as a child. Lewis could think of the disposition in terms of a functional brain structure that would activate for an occurrence of the ability when it was exercised in the future when

seeing red, while not being available from voluntary memory. If we consider a similar weakness for remembering propositional knowledge, for example knowing the capital of Australia but struggling to remember with 'Canberra' on the tip of the tongue, and then once prompted, 'It was there all along in the back of my mind'. This would plausibly be counted as *knowing* Sydney as the capital.

Michael Tye (2000) maintains that physicalism has a satisfactory response to the knowledge argument, but that the ability hypothesis is not part of it. His argument involves the specificity that operates with perception, in relation to the much cruder specificity for remembering and naming experiences. Apparently, humans can differentiate 10 million colours by sight, but we can only store representations and names for a small proportion, because our brains are simply not equipped to remember that amount. As apparatus for Tye's setup, let us index and label red1 all the way through to red10 000 000 for a computer to generate any shade on demand. While we can perceive a difference between red19 and red21, we cannot represent the difference in memory and tag those different representations with different concepts. We must bring colour samples to the shop to match up different furnishings rather than judging from memory, because we understand the concept red, but cannot determine specific shades because we lack the recognitional concepts. In this way we cannot judge colours with the specificity that we can experience colours. He points out that this disability to conceptualise from experience to memory is also relevant for sounds and shapes, whereby we can experience fine discriminations that we have no concept for, such as an inkblot whose shape we can see but for which we have no concept.

Tye agrees with Lewis and Nemirow that Mary acquires the ability to recognise and remember red once she has her first experience, but that she knows more than that while she is looking at red which cannot be characterised as an acquired ability. She knows what it is like to see that shade of red17, without knowing that it is named "red17". Then it is true to say that she knows what red17 looks like even if she only knows it indexically from her perspective as *that* red I am now looking at. Tye asks, exactly what ability is Mary learning? Because she will not reliably be able to identify other items outside he field of vision or on other occasions as red17 as distinct from red18 or 19, nor can she imagine red17 with correct specificity. So, while

she does not gain those abilities, it would be queer indeed to suggest Mary does not know what red17 looks like *while she is looking at it.* It is that aspect of knowing which Tye uses to conclude against the ability hypothesis as a general account of Mary's knowing WIL. Allot might hinge on Tye's use of 'reliability', regarding Mary's remembering red17, for it to qualify as an ability or not, because sometimes she correctly points out a match. I pointed out earlier with reference to Rosenthal (2019), that often-times we cannot remember simple propositional facts like the capital of Australia, but upon prompting, we 'knew it all along'. There may be scope to disagree with Tye on the finer points (for example we might credit Mary with an unreliable ability if she sometimes recognised red17 correctly), but I am persuaded that the ability hypothesis does not manage elements of WIL such as red17 or the shape of a cloud or an inkblot, because possible varieties of experience outrun our capacity to remember or imagine them, while also presenting Mary with something to know of what it's like.

Nemirow (2007) refers to Tye's challenge as, "Objection from Knowing With Particularity in The Moment." (p.35). Curiously, Nemirow spends a page emphasising that Mary must have the ability to recognise and imagine red17 in comparison to red19 *while* they are both in her field of vision (she may manipulate the colour in her mind's eye by altering its shape perhaps, and his argument isn't dependent on how good Mary's imagination is which I assume naturally varies amongst people). He tells us that Tye is mistaken in denying Mary those Lewis abilities while she is looking at red17, but Tye does not make that charge! Nemirow also accepts that Mary's ability to remember may quickly lapse as soon as she closes her eyes. Nemirow proposes that Mary's knowledge of what it's like to see red is a kind of second order perceptual awareness:

> "Arguably, at the first moment when Mary sees red17 she cannot remember having done so. However, the conscious awareness necessary for Mary to know what it is like to see red17 requires that Mary be able reflect on her experience, which she could do only in a moment after the first moment that she sees red17." (2007, p51, fn8).

This would mean that Mary doesn't know what it's like to see red until she has consciously reflected on the experience, which for Rosenthal coincides with her acquiring the ability to remember that first moment of being presented with red. This generally implies that we do not know what any experience is like until after

a moment's reflection, which is at best a questionable matter of psychology, and far too convenient for Nemirow. If I understand him correctly, there is no such concept as red17 to be had by a captivated Mary until she can reflect on the previous moment, and at that point she acquires the recognitional ability to represent red17. Even if we go along with Nemirow, I can reintroduce Tye's objection, and cite the qualia which is the specific hue of red17 as constituting what appears to Mary, and as the perceptual grounding of the concept red17. This reading of the hue itself as a conceptually empty item will likely inform how we might categorise that appearance in terms of knowledge, so that while it may not rise to the status of a fact in support of the knowledge argument, it will serve to defeat Nemirow's argument. Mary must see *that* hue and experience what it is like *before* it is cognitively taken up into a recognitional ability.

Nemirov presents another variation of this theme to argue further for the ability hypothesis, which coheres with my previous analysis: Mary is distracted while looking at red17, and while in that distracted state she does not know what it is like to see red17 because she is not presently looking at it. While distracted she can neither recognise (reliably I concede for argument sake) another sample of red17, or distinguish that sample from red16 or 18, and neither can she remember how to imagine red17. Granting Marys disability for the sake of Nemirov's argument, he concludes that while distracted she does not know what is like to see red17, and the ability hypothesis correctly predicts that Marys knowledge of what it is like to see red17 is only transitive *while* she is looking at it. Even if we grant all that, Mary can still be said to learn WIL to see the visual hue corresponding to red17 upon first sight, which is all the knowledge argument requires. He tells us that Tye uses this example, "…to show that knowledge of what it's like to see red17 cannot be identified with the ability to mentally point to an experience indexically, since Mary has this ability in her distracted state but lacks knowledge of what the experience is like." (2007, p51, fn9). Tye maybe taken to be pointing out that Mary's *imaginative* ability extends to mentally pointing indexically (from the inside as it were, at the hue which corresponds with red17), while at the same time she might not know that 'red17' looks like *that.* This way of parsing the situation also separates the recognitional concept red17 with the hue which appears to Mary, which seems enough to defeat the ability hypothesis while not establishing that Mary is learning a new fact.

I am not exorcised of the 'first-person bewitchment' which Lewis accuses me of. I do however gain some valuable insight from his distinctions for understanding the knowledge argument: Along with ability and information, he says *acquaintance* must be added to the pot (p.289). I believe Physicalism might better account for what Mary lacks inside the room with the notion of acquaintance, which I will broach in chapter five.

Chapter 4. (note on) Frank Jackson's revised position.

Frank Jackson changed his mind from the original position he held when he first presented the knowledge argument. This was one of the, "…more remarkable turnarounds in contemporary philosophy." (Stoljar & Nagasawa, 2004, p.35). Jackson bites the bullet that Lewis fires in the last chapter and takes qualia to be epiphenomenal and unable to cause knowledge (1995). Jackson argues (2003, 2007) that the best physicalist account of qualia is a representationalist theory that frames Mary leaving the room and perceiving red as her being only aware of representational content within her experience and not of any intrinsic quality of redness in the world. There is a common-sense requirement that Jackson must make good on the claim that redness does not really exist, and which prompts me to treat the position as low hanging fruit which I have chosen to delete from the dissertation to make way for prima facie more plausible and promising positions. There obviously must be some merit to Jackson's revised position but I must cull what I judge to be the least valuable or important elements of Robinsons treatment for the dissertation space I have. On the basis that Jackson's revised position is the least attractive approach to the knowledge argument in the literature, and mindful that 'Representationalism' would tie us up in broader complications without clear resources of its own to help with the knowledge argument (see Alter 2007, p72), I will take aim at other physicalist approaches. The 'Phenomenal Concept Strategy' is the subject of the next chapter.

Chapter 5. The Phenomenal Concept Strategy.

The 'phenomenal concept strategy' (PCS) was first named by Daniel Stoljar (2005) to indicate a distinctive sort of physicalist response to anti-physicalist arguments. I consider Robinson's rejection of the PCS a sort of meta-objection because the strategy is usually targeted at other anti-physicalist arguments outlined in the introduction, rather than the knowledge argument itself. Issues which the strategy involve are technical and esoteric and would require a full dissertation for anywhere near a sufficient treatment, but I will try and provide a flavour of the strategy in relation to Robinson's objections and the knowledge argument.

A phenomenal concept is a technical term for a special sort of concept that allow us to think about our phenomenal conscious states. These concepts concern phenomenal states from the direct or inner perspective as exemplified in the mind of the person deploying the phenomenal concept when introspecting or 'in conversation with herself' thinking about qualia or aspects of qualia to which she is paying attention. The PCS situates phenomenal concepts in an explanation of what Mary is missing before she exits the room, whereby she is ignorant of a phenomenal concept for the colour red, and not any physical fact about the red which she already knows. Different philosophers have proposed various accounts of phenomenal concepts all intended to frame qualia in this physicalist friendly *conceptual* way. Loar (1990, 1997) appears to be the forerunner of this approach, developed in various ways by amongst others Carruthers (2000), David Chalmers (2006), Levin (2007), David Pappineu (2002, 2007) and Katelin Balog (2012b).

Physicalists think that these phenomenal concepts are special or unique compared to other sorts of concepts. One way they are unique is by isolation from physical concepts which pick out the same referent, in that the phenomenal concept cannot be inferred from physical concepts, and this isolation supposedly is what partly makes the physical-phenomenal identity counter intuitive. They are also supposed to be unique by their process of acquisition which requires the *experience* of the referent that they are about, and they also inform us to some extent of what those conscious states are *in themselves* (compared to normal concepts which are

always *about* something else). The existence and nature of phenomenal concepts is disputed and getting a handle on exactly what they are and how they are supposed to work is difficult, as is understanding how they are supposed to help us understand the identity of phenomenal qualities with brain states.

It may help students of philosophy begin to get the notion of phenomenal concepts in view by hearing that Wittgenstein would not accept their existence, because there can be no intersubjective checks on their meaning or reference, and so his private language argument would render them nonsense. I will not be directly considering this behaviourist type position.

Brian Loar is usually considered the first to articulate the PCS. Loar focuses on the distinction between concepts and properties. While anti-physicalists he says often rely on an intuition about phenomenal concepts which "…are conceptually irreducible in this sense: they neither a priori imply, nor are implied by, physical-functional concepts." (1997, p.597). Loar says the anti-physicalist extends this intuition to conclude that phenomenal *qualities* cannot be identified with physical-functional *properties,* but that no sound philosophical argument has yet appeared to support that. Loar (like most everybody else) feels this intuition, and proposes that the PCS, "…may provide some relief, or at least some distance, from the illusory metaphysical intuition." (1997, p.598).

I will interrogate Robinsons rejection of the PCS and probe those theories he mentions to an extent for judging his rejection.[1] Robinson claims the PCS is wrong headed in targeting how we can know or *think* a concept rather than how there could *be* a phenomenal property different from a property exemplified in the brain. Balog says, "…epistemic and conceptual gaps can be explained by appeal to the nature of phenomenal *concepts* rather than the nature of non-physical phenomenal properties. Phenomenal concepts involve unique cognitive mechanisms, [which can be] physically implemented." (2012, p.1). Robinson's consideration of the PCS keeps returning to the charge that it confuses conceptual irreducibility with property irreducibility so that conceptualising from phenomenal properties is supposed to generate distinctive physical-functional properties. Robinson's core claim is that physicalists do not explain how this goes.

[1] There are numerous theories which fall under the PCS. I will only engage those mentioned by Robinson (which does seem to be a fair representation of those most cited in the literature).

Before Brian Loar, phenomenal concepts had only been used by dualists to point at private first-person aspects of experience, or what we now call qualia. Then physicalists began using them in the context of the PCS:

> "The strategy, as most generally expounded, consists in some way of claiming that phenomenal concepts contain the experience within themselves; generally this is conceived on the model of indexical's or of quotation. The result is meant to be that the 'explanatory gap' and all that goes with it, cease to be challenges to physicalism once one realises that these things are merely the by-products of the difference between physical and phenomenal concepts, and not a difference between public and private properties…" (Robinson 2016, p.75).

If phenomenal concepts do not acquire their content from the qualia which they designate, physicalists must explain from whence the content is derived. Robinson thinks that the quotational-indexical model of the PCS[2] specifically mimics dualism in a way which mirrors Russell's semantic of logically proper names whereby the qualia itself ('sense-datum' in Russell) features in a proposition. In this sense he judges the PCS as offering no innovation, but simply as claiming for physicalism, semantic features designed to refer to qualia. Physicalists he insists must account for this 'semantic transfer problem' and explain how qualia relate to physical properties, or at a minimum tell us how the PCS illuminates these two things. Robinson setting the stalls out like this may be tendentious, in that neither Russell nor anybody else have settled a theoretical status for qualia which would inform a thorough going semantics. That said, it is a fairly straight forward act to pick out qualia without committing to what they really are, so I will take the semantic transfer problem as a prompt to alert illegitimate uses of qualia terms by the physicalist.

Up to now in this dissertation I have resisted using the phrase 'hard problem', because I could avoid it with distinctions I have made and because it is sometimes abused in popular writing, but Robinson makes essential use of it here in his argument. The 'hard problem' was coined by David Chalmers:

> "…a mental state is conscious if it has a qualitative feel - an associated quality of experience. These qualitative feels are also known as phenomenal qualities, or qualia for short. The problem of explaining these phenomenal qualities is just the problem of explaining consciousness. This is the really hard part of the mind-body problem." (1996, p.4)

[2] There are numerous approaches within the PCS. The 'quotational' approach originated by Papineau seems to receive most attention in the literature.

The problem is that the properties science ascribes to the brain and the qualitative properties which the brain enables us to experience seem entirely different. Robinson objection centres around Leibniz's Law which says it is impossible for items with different properties to be identical, and he charges that the PCS must cope with Leibniz Law while accounting for the semantic transfer problem. Robinson objects that the PCS seems actually a retrograde step for physicalism because older theories like functionalism and topic neutral analysis which I discussed in earlier chapters were direct attempts to answer the hard problem because they take phenomenal data *not* to be the transparent presentation of phenomenal qualities like they intuitively seem to be, and in this way they claim not to violate Leibniz Law. Although those older theories fail, he says it can be seen how they are at least *attempts* to solve the hard problem, while in contrast it isn't clear how the PCS is even meant to address the hard problem, "…how does the fact that phenomenal concepts are in some sense discontinuous with physical concepts show how a physical state can manifest itself (or seem to do so) as a phenomenal quality?" (p.78). Robinson anticipates four options that the PCS could take (pp.79-91) and roughly aligns different physicalist attempts amongst them. Instead of assessing the four options he lays out for the physicalist, a more comprehensive way of considering Robinsons objection is prompted by a footnote concerning a paper by Balog (2012a), in which she lists eight desiderata of phenomenal concepts in the context of the traditional puzzles about consciousness. Robinson complains that none of her formulations confront Leibniz Law, and so they do not do, "…justice to the hard problem, neat, so to speak." (p92, fn1). Much will fall on how we interpret the hard problem. Balog (2012, p6) sees the puzzles as mostly epistemic in nature for physicalists to nominate qualia as *appearance* properties and not essential properties, and so not offending Leibniz's Law.

Robinson frames the PCS as confusing a de dicto problem with a de re problem and puts a question in Marys mouth, "how can that property that I am now experiencing be the same property as one of that set I already knew about from science?" (p.81-82). He insists that the real problem is only in focus from the first-person

perspective and presents the following argument along these lines (p.89), which I will claim over generalises. The first lines I paraphrase:

Call the set of physical facts P.

Call Mary's scientific mode of access to those facts S.

Call the experiential mode of access which she lacks in the room H.

> "If we regard S and H as external to P, then the addition of new mode of access will not alter P. But the physicalist hypothesis is that everything relevant is included in P: modes of access to physical facts are themselves simply physical processes and are included in physical facts. Therefore, if [Mary] knows all the relevant members of P, [she] should know all the facts about H, *including the fact of what that mode of access is phenomenally alike*." (p90. Emphasis is mine).

In effect this seems like a stronger version of Jacksons knowledge argument because it makes explicit a questionable connection between modes of access and physical facts. Making the connection explicit in this way prompts my objection. While I accept that a new mode of access to P will not alter P, it is a curious claim that modes of access to facts, being physical processes, are included in the physical facts. Leaving H open ended or without a rationale as to what is included in H, places an implausible requirement on P as follows. An echolocating bat enjoys a mode of access to P, and while Mary could be fully informed on the entire workings of echolocation, we wouldn't expect her access to P through S to underwrite her access to what it is like for the bat, not least because that would be impossible for Mary but relatedly because it would be a bizarre requirement on physicalism that S must include all facts pertaining to what all phenomenological perspectives are like. A weaker objection also works because he does not limit the experiential perspectives feasible for Mary. How are we to understand what he includes in H? We experience different phenomenology when viewing a picture from different angles, and it is unclear how that fact (if it were a fact) could be included in H. Robinson would push back and place the onus on the physicalist to specify why S does not range over P while including H (if physicalism is the thesis that all facts are physical facts). The phenomenal concept strategy involves proposals that Mary is ignorant of concepts and not facts.

My objection to Robinson is related to a weakness in other anti-physicalist arguments noticed by Sundstrum (pp272-74, 2011). PCS Physicalists accept that there is something right about the intuition that what appears to consciousness is how reality really is. Sundstrum quotes philosophers saying that we may use phenomenal concepts to pick out phenomenal properties by 'directly and essentially' referring to them (Loar), or by 'providing a grasp of phenomenal properties that reveals their essence' (Balog). Horgan and Tienson interpret those phrasings to license the following argument:

> "1. When a phenomenal property is conceived under a phenomenal concept, this property is conceived otherwise than as a physical-functional property.
> 2. When a phenomenal property is conceived under a phenomenal concept, this property is conceived directly, as it is in itself.
> 3. If (i) a property P is conceived, under a concept C, otherwise than as a physical functional property, and (ii) P is conceived, under C, as it is in itself, then P is not a physical-functional property.
> Hence,
> 4. Phenomenal properties are not physical-functional properties."
> (Sundstrum 2011, p272, quoting Horgan and Tienson 2001, sect. 3).

Sundstrum argues that the argument over extends what the likes of Loar and Balog are committing to by their phrasing about phenomenal concepts 'picking out the essences of properties', and that those PCS philosophers have in mind more nuanced understanding about how phenomenal concepts may reveal essences to the extent which blocks this argument. Sundstrum quotes Loar claiming that his model of direct reference involves, "directly rigidly designate" (Loar 1997, p603), and Balog explaining that her quotational approach employs phenomenal concepts in a way that, "will not afford any clue as to the fundamental nature of the referent. While they 'afford an insight into the essence of the referent', the sense in which they do so is 'by exemplification'; i.e. in the sense that phenomenal concepts use phenomenal properties to think about phenomenal properties" (Balog 2012, sect. 2).

Sundstrum shows by example how we can controversially refer directly or rigidly or by exemplification to phenomenal properties without meaning we have an insight into the essential nature of the referent. I will use my own example of a tambourine musical instrument for his explanation, which I hold in the air asking friends to observe while I look away. Sundstrum says this example will show the extent to which PCS theorists may mean that we can refer directly or rigidly or by exemplification without knowing essential properties of the thing. Me saying, 'look at the shape of this instrument', without necessarily understanding that an essential property of that shape is the ratio of its radius to its circumference, and so legitimately excluding that essential property from my concept, so Horgan and Tienson's argument is shown to demand too much and fail.

Before I move to consider the PCS theories directly, I will table another concern of Robinson which in one way or another we will see motivates his resistance to the PCS. He outlines what he calls the doctrine of 'weak transparency' which he claims must be true (2016, p.83-85): 'Weak transparency' is the doctrine that we must be correct in attributing some central features of our phenomenal experience of the world to our common sense concept of the world. Those phenomenal features are not required to be *actual* features of the world, "only that they *purport to be* and are features of our *naïve* or *manifest image* conception of the world." (p.83). Although this doctrine may be some-what vague, in some sense it must be true, although working out the details is fraught with difficulty as we find at least since John Locke presented the concept of primary and secondary qualities.

Robinson concedes that he does not understand the utility of phenomenal concepts for the physicalist. His rejection of the leading two contending styles of account as I said at the beginning amounts to a somewhat 'meta-level' rejection which does not engage the theories on their own terms.[3] Before I scrutinise those accounts in more detail than Robinson does, it will be helpful to mention the basic differences of approach. Loar (1997) and Levin (2007) propose *recognitional* phenomenal concepts. These concepts are possessed

[3] A variety of structural features are used to characterise and specify phenomenal concepts. I will only be concerned with the two varieties singled out by Robinson, which appear between them to receive the most attention in the literature. For a recent extensive bibliography of the different approaches see: https://www.oxfordbibliographies.com/view/document/obo-9780195396577/obo-9780195396577-0254.xml

partly by being able to recognise phenomenal experiences that are of a kind. The other approach employs '*quotational*' concepts as proposed by Papineau and Balog which are deployed when using something of the actual conscious state which is present to the mind. Loar's original idea in (1990/97) was that phenomenal concepts are *direct recognitional* concepts in the sense that a person can directly refer to an experience because the concepts way of presenting itself involves the actual experience, and that the experience does not fix a reference but puts the subject in causal contact with the referent. The recognition of a qualia triggers a recognitional concept such as '*that red*', which the subject can deploy upon future experiences of that red in the future. The subject thereby has acquired a recognitional concept by a demonstration occurring in the mind of 'that redness'. Such a recognitional concept is called a recognitional phenomenal concept.

Balog (2018) understands the field of research since Loar as generally composed along these two lines: the recognitional style account of Loar and Levin which emphasise direct reference and hence the conceptual role of phenomenal concepts, and the quotational style account of Papineau that we will come to, which develops that aspect of phenomenal concepts as somehow featuring the experience itself.

Robinson credits Brian Loar (1997) as the innovator of the PCS, and I have chosen to consider Loar's more developed account from his more recent 2003 article which focuses on anti-physicalist arguments prompted by Kripke's "Naming and Necessity" (1980). Loar thinks a certain reply to Kripke will allow the physicalist to accommodate the explanatory gap, because at bottom he diagnoses a conceptual issue grounding that gap, and not a property issue. I am mindful that Robinson considers the explanatory gap per se, to miss the crux of Mary's situation, however Loar is explicitly concerned with the status of properties "non-epistemically conceived" (2003, p.113), so let us try and glean any connection to the hard problem as Robinson has presented it.

Loar conveniently presents the question in conformance with Robinson's objection: does a phenomenal concept relate to a property, such that having the concept implies knowing the essence of the property? Kripke (1980) sets up the problem for Loar by showing that we directly refer to token sensations as we experience them, so that the way we designate those sensations means they could not be anything other than

that sensation. This seems to block a posteriori identification of that conscious state with a physical brain state, in contrast to run of the mill scientific identifications such as our coming to learn that water is H2O. The difference between referring to the sensation (which seems impossible that we discover a posteriori that it is a brain state, because we are so intimately acquainted with *that pain,* no explanation could override that identification) and water (which we in principle could accept an explanation that identified those transparent and fluid appearance properties with H2O because we might readily understand how H2O constituted those appearance properties) is called the 'explanatory gap'. That explanatory gap makes Mary's reference to redness seems necessarily to block a posteriori identification with a physical-functional brain state. Loar tells us that the physicalist who is a realist about properties should not suppose, "…that properties are individuated epi-stemically; …So the question is this: how might we explain the a posteriori status of a psychophysical property identity other than by supposing either that properties are epistemically constituted or that we directly grasp the essence of qualia?" (Loar 2003, p.5). Loar thinks the physicalist has to explain either why mind brain identity statements are special in the sense that references to sensations express contingent identity with brain states, or else explain how can the reference to sensation be connected a posteriori to some neural assembly without expressing different properties. His proposal of recognitional phenomenal concepts is an attempt to build such an explanation.

Loar conjectures that psychological items could be hard wired in a way that supports concept relations which only *appear* to identify different properties, because those physical neural connections underlying them do not provide a *cognitive* connection (pp.116-7). At first blush this might seem to threaten Robinson's doctrine of transparency (2016, Ch 5, sec 6), which I accepted as a rough datum for the dispute (keeping in mind the possible complication of primary and secondary qualities) because all sides accept that the qualities we perceive are in some sense qualities of the stuff in the world and are not just as it were 'in our heads'. Loar then characterises how phenomenal concepts are recognitional concepts (rather than descriptive or logical concepts), in featuring the phenomenal state itself.

I will try and give a flavour of Loar's technical proposal (2003) for our purposes. A phenomenal concept includes qualia aspects, from visual percepts 'taken-up' into psychological roles, perhaps in the form of phenomenal memories or vague 'patterns' enabling re-identification in the future. Loar specifies how phenomenal concepts are directly referenced, which I will quote because the general plausibility of the whole PCS is implicated:

> "…phenomenal concepts are not descriptive concepts; they do not have the form "the property I hereby ostend" or the like. It is tempting to say that a phenomenal concept picks out its reference, a certain quale, directly. This notion "directly" is not the weak one used in causal theories of reference. On that use a visual demonstrative concept of a tree refers directly if it does not do so by way of a classical Fregean sense or a descriptive condition. That visual demonstrative is mediated by a visual experience that can itself be reflectively attended to, even though the concept does not refer to it. And that visual experience is quite distinct from the tree. Now a phenomenal concept intuitively conceives of its reference without any such distinct intervening factor. It is moreover somewhat tempting to think that in exercising the concept one, as it were, grasps the essence of the property it picks out. But these are distinct thoughts. That phenomenal concepts are phenomenally direct does not entail such a grasp of essence. I see no reason why physicalists should deny the former, while they will deny the latter if grasping is meant to be, as it were, revelatory." (2003, p119).

Loar proposes that the recognitional act of referencing token experiences of a specific shade of blue must be grounded in a less specific disposition to recognise course-grained blueness (because though we can experience a thousand shades of blue, we cannot reidentify them). This disposition enables the deployment of a type-demonstrative phenomenal concept of course grained blue, but not token-demonstrative phenomenal concepts of each shade of blue. Loar claims an upshot from this is that because those proposed essential properties of each shade of blue cannot be mapped one for one with phenomenal concepts, then the dualist who thinks "acquaintance" (Ibid p.119) with the *essence* of a token experienced specific blue, cannot think she thereby is acquainted with the essence of generic blue, and therefore she ought to regard the disposition to re-identify course grained blue as crucial to deploying the phenomenal concept, and therefore the phenomenal concept cannot not be taken to be a direct presentation of essence.

Loar argues that if anti-physicalist arguments cannot be formulated without phenomenal concepts as connected in some sense one to one with essential properties, then there is nothing for the physicalist to

respond to. Loar says for all we can know, when we refer to a sensation, we are directly referring to that neural assembly which implements the phenomenal concept, however:

> "The appeal to recognitional concepts that pick out, say, neural properties is not intended to sidestep qualia. That would hardly count as a defence of psychophysical identities. Physicalism is the thesis that, however odd it may seem, that quale (which I am now conceiving phenomenally) might, for all we know, be a physical-functional property." (p.121).

This may license a description of Mary leaving the room to have a new experience that does not amount to her learning a new proposition of the sort "that is what it is like to see red17", where 'that' is functioning in the way it does for a proposition, because she can hardly know a property if she cannot determinately reference it.

Robinson's stance on the hard problem is that our acquaintance with properties in experience, or our epistemic relation to that experience is such an intimate connection that it cannot be explained by epistemic relations to anything else, whether directly referenced or not. This is to object that Loar cannot appeal to phenomenal memories or vague phenomenal patterns in the construction of physical phenomenal concepts because he is illegitimately using acquaintance properties (or properties not sufficiently different from acquaintance properties) where it is impossible that such properties could be in the brain and so cannot be used in constructive materialist accounts. Joseph Levine calls this the 'materialist constraint' on physicalist accounts of phenomenal properties, "that no appeal be made in the explanation to any mental property or relation that is basic." (2018). In the rest of this chapter I will firstly consider what Levin (2007) adds to Loar's recognitional account, before turning to Papineau's quotational account (2007), appended by Balog (2012), and then a different type of potentially undercutting objection from Pitt (2019).

Janet Levin (2007) aims to bolster Loar's account by proposing the functional characteristics of those recognitional concepts and which help frame how they might fit into Mary's epistemology. Loar characterised phenomenal concepts as directly denoting referents without mediation by a mode of presentation, and Levin adds that phenomenal concepts denote the neural properties which are causally implicated in our use of those phenomenal concepts for introspective tasks, which can be either token or type phenomenal concepts.

Token concepts are used for first person pointing at a token instance of an aspect of qualia, and they denote like a token-demonstrative concept. So the particular neural configuration which, "causes me to make introspective note of some experience I'm now having counts as the denotation of the token demonstrative "that (experience I'm having now)"."(2007, p.88). Although these token-demonstrative concepts can feature in our introspective knowledge, we may not remember them well enough to pick out the same aspect of experience in future. They only pick out, "instances of experiences with phenomenal properties, and not those properties themselves." (p.89). Levin says *type* phenomenal concepts are more useful against anti-physicalist arguments because they may pick out:

> "("that *kind* [of experience]"), which purport to pick out *kinds* or *properties* of experiences from an introspective perspective. The denotation of a phenomenal type-demonstrative will be the property—presumably physical—that's causally responsible for the application of that concept in the introspective *recognition* or *reidentification* of an experience as "that (kind) again" (p.89)

> "…If we manage, fairly consistently, to pick out the *same* physical property when noting, "Oh, that (twinge) again" or writing "S" in our diary, then our phenomenal type-demonstrative denotes that property. If there *aren't* such physical properties, of course, we're stuck with either dualism or eliminativism." (fn5).

The reference of type-demonstrative concepts are achieved through the, "causal and dispositional relations an individual has to her internal states that are effected by an introspective "pointing in"; that is, by the *fact* that she's in causal contact with a certain property and is disposed to reidentify it on subsequent occasions." (p.89). So, the difference between a type and token concept is that the former express repeatably recognizable phenomenal properties, while the token expresses particular experiences that have those phenomenal properties. But why couldn't the subject attend to that phenomenal property within the token experience? Because, Levin tells us, the subject cannot determinately reference that property if she lacks the ability to recognise it, and the only 'physicalistically acceptable' way to confirm that ability is to test her disposition to identify other instances of that property.

The difference between *non*-phenomenal type-demonstratives like "that (kind of) dog over there" and phenomenal type demonstratives is that the latter can only be acquired by introspective attention. Levin tells

us that the phenomenal and non-phenomenal type-demonstrative concepts share the following features: their referent essentially involves the perspective of the demonstrator, and it is *direct*, that is without identifying a mode of presentation. These features of phenomenal concepts are taken by physicalists to make them non-equivalent to physical or functional properties and so Mary's new experience is not of irreducible phenomenal properties, but rather a new way of conceptualizing what she already knew under physical-functional description.

I understand how Levin's account might be seen to model how phenomenal experience is processed by the brain, which may be the start of intuiting how Mary's experience of seeing red is identical with the activation of a physical-functional configuration that she already knew about under a different description. Levin is effectively modelling causal nodes involved by the brain in representing the world whereby, "phenomenal concepts are not supposed to characterise phenomenal properties at all"(2007, p105), which attracts Robinson's charge that the Levin's account effectively collapses back to Armstrong's topic neutral account: "if normal experience gives us topic neutral knowledge of our inner workings and Mary already has topic specific knowledge, then she already knew more before she had colour experience than can be acquired by having it." (2016, p89). Robinsons diagnosis here might be challenged. His idea seems to imply that the physical description under which Mary already knew about the neural property is *objectively exhaustive*, and it will therefore be more specific than the knowledge she can acquire with the phenomenal concept, and consequently we are left with no explanation of how the further (topic neutral) knowledge from the phenomenal concept can be new. It seems Robinson is implying again that scientific facts and description must exhaust all perspectival knowledge. Physicalism only requires that phenomenal truths supervene on physical truths, not that Mary will know all phenomenal truths because she knows all the physical facts. This will be explored in the coming chapters.

Something like Robinson's doctrine of weak transparency (2016, p.83-84) must be true or else we lose connection to the world in a way which positions the functional aspect of Levin's account as a 'just-so-story', which could be substituted for any causal model in the background without featuring our phenomenal sense

of the world at all. 'Weak transparency' is the doctrine that we must be correct in attributing some central features of our phenomenal experience of the world to our common sense concept of the world. Those phenomenal features are not required to be *actual* features of the world, "only that they *purport to be* and are features of our *naïve* or *manifest image* conception of the world." (p.83). Although this doctrine is left some-what vague (primary and secondary qualities?), it is clear enough to render Levin's account opaque as an explanatory account of Mary's experience in relation to her neural causal processes in terms Chalmers hard problem and with reference to Levine's materialist constraint cited earlier. For these reasons I can empathise with why Robinson might feel justified in his meta rejection of the PCS without engaging the intricacies of those theories to even the basic level which I attempt.

Now to Papineau (2002, 2007), who calls his approach a 'quotational' phenomenal concept strategy. He attempts to locate the phenomenal aspect more explicitly in the physical functional workings of the brain. A quotational concept in some way employs the very conscious state that one is thinking about. Papineau (2007) is perhaps easier to understand in regards how those phenomenal concepts are referenced. He tells us a phenomenal concept may feature in the physical-functional brain operation somehow similarly to a how a word or phrase features in an explanation: *by quotation*, in that a phenomenal concept will feature phenomenal aspects of experience as like encapsulated in quotation marks as "*that experience*". Papineau gives a plausible explanation of how facets of cognition may causally relate to such a phenomenal concept, including a simple model of how the brain could use those cognitive aspects or information packets for other processes.

Papineau spends pages specifying his understanding of phenomenal concepts as a special case of the species of perceptual concept (pp114-20), as stored sensory templates which function to enable our visual sense and visual imagination. These templates connect to various cognitive functions, partly serving as 'file holders' for associated information. He posits a 'structural hierarchy' which allows the phenomenal concept to signal either tokens or types depending on whichever sort the information is appropriate for, and which vary in how present they are to consciousness. Papineau aims to improve his 2002 theory which ran into problems

because featuring "*that experience*" tended to illegitimately run together both the phenomenal aspect of the experience and the experience applied in a theory of experience (presumably flirting with semantic transfer), when it should have explained those aspects of phenomenal concepts instead of presupposing them (p.121). His new thinking has phenomenal concepts as a special case subset of *perceptual* concepts, which will underwrite more sophisticated conceptual abilities for remembering specific colour shades for example than seemed possible with the type demonstrative concepts from Levin and Loar. Most of the article concerns the modelling of a neural 'filing system' and as far as I can tell is uninformative regarding the hard problem. Indeed, it largely seems to concern the easy problem of consciousness by describing the layering of cognitive mechanisms which could perform functions in relation to phenomenal concepts.[4]

It is most curious how the quotational element of phenomenal concepts is supposed to work in relation to percepts. Papineau's construal of 'using' and 'mentioning' phenomenal concepts (p.125-6) conditions his response to the knowledge argument. Being careful not to put words in Papineau's mouth, "…phenomenal references to an experience will deploy an instance of that experience, and in this sense will *use* that experience in order to mention it." (2007, p.123). He explains that when a phenomenal concept is exercised, a sensory template that is shared between real perception and visual imagination is activated. Phenomenal thought can then use that template to think about the perceptual experience itself, as distinct from the object of the experience. Any exercise of a phenomenal concept is like this, it will use either the experience of seeing a tree or the seeing the true just in the mind's eye. Then counting that imaginative experience as a "version" of the actual experience allows us, "to say that phenomenal thinking about a given experience will always *use* a version of that experience in order to *mention* that experience." (p.123). Papineau then evokes what is often called (since Moore) the transparency of perception: when we try to focus on the visual sensation or the qualia of the tree, we effectively just look more intently at the tree. (Personally, I seem able to parse the two things when I am looking at the tree. I say 'seem' because I am unsure how to articulate what I have in mind when I make that effort. Papineau does nod to this sort of possibility by saying he will bypass potential

[4] the 'easy problem' named in comparison with the hard problem because explanations of cognitive functions that do not include qualia are more obviously amenable to regular scientific causal explanation (Chalmers 1995).

debate about this (because presumably it will not affect his argument)). Papineau claims it is phenomenologically equivalent to think "phenomenally *about* an experience," and to think "perceptually *with* that experience". He claims there is no difference in what it is like to concentrate on the visual phenomenon of the tree, and what it is like to see the tree. He argues that because the same sensory template serves both the perceptual experience and various phenomenal thoughts about that experience, it follows that they both share the same phenomenology. I understand how my imagining of the tree after the experience could use a perceptual template that I acquired when I saw the tree, but not why we should say that my imagining and the experience both share the same phenomenology. For all his talk of imagining the phenomenon, it seems that my phenomenal thoughts about the tree and my percept from the tree might only feel the same (for arguments sake) *when* I am looking at the tree.

Papineau discusses a surprising implication of his view, which helps clarify my dissatisfaction with it. His phenomenal concepts derive semantic capacity from the cognitive functions they perform and not from their phenomenal aspect. The concepts 'gather' information about experiences which grounds how they refer to those experiences, and they are in part constituted by versions of the experiences with whom they share what-it's-likeness (WIL). Papineau concedes that the WIL aspect seems not involved in any essential part of the semantic operations of the phenomenal concept. To bring this out he evokes a thought experiment whereby humans evolved to attach information about experiences to words in some language of thought instead of sensory templates. Those brain states would equally refer to experiences, even though their activation does not involve any sharing of phenomenality with their referents. If this is right Papineau says, then it will challenge his idea that phenomenal concepts, "involve some distinctive mode of phenomenal self-reference to experiences." (2007, p.125). And if the phenomenal element of a concept achieves nothing toward that concepts capacity to refer, then how he asks, should we understand the notion that they are distinctively phenomenal? The language of thought model cannot 'quote' any version of the experience in the way a phenomenal concept is supposed to, and likewise he tells us that any functionalist account that has

causal stand-ins for the phenomenal aspect that are supposed to refer to experience for the same reason as phenomenal concepts do, is open to the same worry.

After describing this worry, Papineau curiously tells us that there is not really a problem here: other brain states which satisfy the same cognitive function by referring to experiences for the same reason as a phenomenal concept can carry out the same role because with his account it is only the cognitive function carried out by the phenomenal concept and not its phenomenology that refers to the experience. He sees it as a further question whether we wish to categorise those nonphenomenological (language of thought) states as 'phenomenal' given their lack of WIL, and in any case that would provide, "no ground for denying that they would refer to experiences for just the same reason that phenomenal concepts do" (p.125). I judged this move curious because it inclines me to wonder what help the PCS can be to understanding Mary's epistemic situation if it can just substitute the WIL aspect completely. It makes Mary's impression of red inconsequent to proceedings. Papineau gestures an explanation about the brain featuring the phenomenal aspect of phenomenal concepts just because they are automatically present along with their referents so brain evolution would not over burden itself when it can with least redundancy activate that ideal thought vehicle of the perceptual sensory template for thinking about those "selfsame experiences" (p.126). This approach disowns the spirit announced by the originator of the PCS who empathised with the intuition that phenomenal qualities cannot reduce to brain activity or any physical-functional type, and aimed the PCS so to, "…provide some relief, or at least some distance, from the illusory metaphysical intuition." (Loar 1997, p.598). Other than some plausible sounding out of the easy problem of consciousness, Papineau has provided very little relief from any metaphysical illusion. The WIL element of experience which he claims can in principle be replaced with brain architecture not involving any phenomenal aspect (with no difference in behaviour), indicates that all his discussion about conceptual operations is all on the functional-physical side of the equation and so does nothing to dissolve the intuition of non-identity which motivates the hard problem. His phenomenal concepts are consistent with zombies or robots experiencing no phenomenology (the lights are on but there is nobody home). I suppose he would view that favourably because it sidesteps problems with

qualia, but it also makes his solution incommensurate with the hard problem. I will next consider what Papineau carries over for explicitly engaging the knowledge argument.

We might wonder what essentially qualifies a phenomenal concept *as* phenomenal in Papineau's view if a non-phenomenal language of thought architecture can be causally equivalent. He parries this question away as one of definition which he is content to leave open, and offers a response to the knowledge argument which relies on other features of phenomenal concepts; Mary lacks the phenomenal concept of the experience of seeing red, which she can only gain upon seeing red. Then she will have acquired the sensory template which allows her to know about the experience of red in a phenomenal way which she can then *correlate* with her knowledge of brain waves and light frequency and what have you (I hesitate to say 'correlates' the phenomenal with what she already knows about seeing red, because she is supposed to be learning of an identity. This hesitation goes to the issue of whether having closed the explanatory gap, will she no longer intuit the non-identity?). Papineau leaves the phenomenal aspect in that black box handed out in Chapter two by Dennett, when he tells us the WIL is a contingent feature of humans for which we are constituted to configure sensory templates, but if we were differently constituted then we wouldn't need the experience to know what red looks like.

Papineau ends his brief discussion of the knowledge argument (p.128) by claiming one way his account is more powerful than other PCS accounts (those which require the phenomenal aspect to be activated when thinking related thoughts). He thinks we can use phenomenally *derived* information (as attached information files), to think about a particular experience without the cognition needing to 'mention' that experience and that this is a way that a phenomenal concept can qualify as "phenomenal" without a WIL aspect; it only needs to be phenomenally *derived* to qualify. Firstly, it is not obvious that I can infer or remember something from an experience of red without that phenomenal aspect appearing in some however minimal way to my mind's eye. Secondly, even if this is psychologically possible, his account only seems an improvement on Loar and

Levin in terms of offering more substantive profiling of the cognitive architecture that might support phenomenal concepts, which is just the easy problem again.[5]

Balog (2012) effectively offers further detail for a quotational account in terms of phenomenal conceptual structure explained with reference to levels of 'mentalese' as items in a language of thought. Balog elaborates a unique way in which we reference phenomenal concepts by experiential *acquaintance*, which I understand as another way of implementing Papineau's notion of 'mention'. Balog's aim is to put the nature of acquaintance to work more explicitly in the semantic-causal workings of phenomenal concepts so to obviate the inclination to posit nonphysical mental states. I will resist engaging more with Balog because her theory is effectively an elaboration of Papineau's, and clashes even more with Levine's materialistic constraint. I murmured that Papineau's and Balog's speculations about cognitive architecture only amount to a 'just-so causal story' as far as concerns the metaphysical nature of phenomenal qualities. I mean that while their causal profiling of cognition may be coherent, the plausibility of their accounts to that extent does not reach out to satisfactorily feature the experiential properties that feature in the hard problem. This is along the same lines as Levine's complaint about the quotational strategy, "trying to explain the substantivity of acquaintance by appeal to the *cognitive presence* of phenomenal properties in our phenomenal concepts, which, in turn, is explained by *physical presence*." (2018, p.15)

I have chosen to briefly consider a theory by David Pitt (2019) (which is unmentioned by Robinson) because it targets the plausibility of the PCS in relation to my 'big-picture' objection that the whole approach only amounts to a 'just-so-story', and will also back up my earlier two worries about Papineau's 'shared phenomenology' between perceptual and phenomenal concepts, and my psychological failure to attend to phenomenally derived information without activating sensory imagination, which Papineau appealed to.

[5] Robinson reports that Papineau told him in conversation that the PCS is not concerned with the hard problem, but with the knowledge argument, the conceivability of zombies, and the explanatory gap. Robinson claims that understanding misconstrues the knowledge argument as a de dicto problem. Correctly understood he claims, the knowledge argument is a de re problem, "to explain physicalistically the phenomenal quality that figures in the experience, not just to explain the role of the concept that characterises it."(2016, p81).

Pitt (2019) defends the view that for Mary to know what red is like *just is* to experience it. He claims that upon seeing red, Mary does not acquire propositional knowledge or concepts, or an ability: in experience she simply becomes acquainted or familiar with a phenomenal property, and to know what a kind of experience is like is to be acquainted with those phenomenal qualities which characterise it. He calls this epistemological category "acquaintance-knowledge".

Pitt emphasises the distinction between acquaintance knowledge which Mary acquires and knowledge *by* acquaintance. He thinks Earl Conee (1994) first proposed this account of Marys new knowledge calling it 'phenomenal knowledge', and that it may have been what Bertrand Russell meant when he said "there is no difference between the experience and knowing that you have it." (1940, p49). It is distinct from propositional knowledge which can be derived *from* the acquaintance. Pitt specifies that when one has knowledge that a phenomenal property Q is *like this,* then the demonstrative 'like this' refers to an instance of Q. Pitt will argue that this approach is the only way to substantiate the claim that Mary gains new knowledge, and it will involve the rejection of phenomenal concepts.

Pitt understands phenomenology to be constituted by discrete items which cannot be 'broken up' in any way needed to support a cognitive architecture like that proposed by the PCS. This crucial aspect of Pitt's proposal undercuts the PCS: "There is nothing one can *think* once one has experienced red that one could not *think* before experiencing it" (2019, p.89).

Pitt thinks that none of the ways the PCS individuates the content of a phenomenal concept by experiencing a quality can work; neither in virtue of *referring* to it, or by being a *constituent* in the manner of Papineau and Balog. Pitt tells us that intuitively, concepts cannot *be* either percepts or images because these are fundamentally distinct sorts of mental things, shown by the fact that we have concepts for things which cannot in principle be imagined or perceived. We might also have concepts of perceptible items which we cannot imagine like hens with 508 red spots. We can think about such things without being able to imagine them, so those concepts cannot be percepts or images. There is also the other the side of the coin; things we can perceive and imagine but cannot conceptualise, like when we stare at a scene or perhaps a non-thinking

animal looks at a field. Pitt insists that conceptual contents must be *thinkable* while images and percepts are in a different category of mental items which makes them necessarily not thinkable things. In that way, for one to report that they are thinking about the smell of a rose is a nonsensical category mistake. The sense in which we might think of such things he says is by possessing, "otherwise-content-individuated concepts that can *refer* to them." (2019, p.91).

Pitts overarching framework is that there exists a distinct experience of thinking; a *proprietary phenomenology of cognition,* which he calls 'cognitive phenomenology' (2004), and which is individuated and distinct from those other proprietary modes of experience like seeing and hearing. While this is a controversial thesis employing a small cottage industry of philosophers arguing for and against either the *existence* of cognitive phenomenology as a whole or sub-categories such as the phenomenology of different sorts of thinking (see Bayne et al, 2011). I take the plausibility of Pitts theory as support to hold up my categorising the PCS as a 'just-so-story' of causal modelling and almost irrelevant to the hard problem.

The metaphysical case for cognitive phenomenology (Pitt 2011) starts from the purported fact that conscious states must be phenomenally individuated if they are conscious at all. Sights and sounds are of a different kind than conceptual thoughts and are phenomenally constituted by radically distinct kinds of phenomenal properties, which undercuts how phenomenal concepts are supposed to work. Pitt insists there is no such thing as cross-modal experiences like smelling colours or hearing smells. According to Pitt, conceptual thinking is ontologically distinct in the way scent and sound are distinct; one can only *cognitively experience* thoughts. The crucial inference Pitt wants to make from this is that phenomenal concepts cannot be individuated as the PCS proposes. The conscious process itself is what individuates *cognitive* phenomenal concepts which are ontologically independent of the sense modalities (2019, p93).

Phenomenal concepts proposed as cognitive-phenomenal experiences, if Pitt is right they cannot have colours as constituents, so there is no such thing as the red-percept-or-image-containing concept for Mary to acquire when she leaves the room. Those percepts or image like phenomenal items can, according to Pitt be

thought about, by applying a concept to them, but the content of the phenomenal concepts cannot involve non-cognitive phenomenology.

Perhaps my simple idea of juxtaposing the plausibility of Pitts theory with the PCS doesn't require the laying out I have provided but my point comes further into view when we consider that the epistemology and ontology posited by Pitt is neutral in regards the ontological status of the properties she becomes acquainted with upon first seeing red. Acquaintance knowledge is a way of knowing a phenomenal property which requires direct experience of that property. All sides agree on this. He tells us that acquaintance knowledge is a fundamental type of knowledge for phenomenal properties, and that while we may become acquainted with theoretical entities like a 20-sided shape by seeing a picture of one, it is not the *only* way of knowing it such is the case with a phenomenal property. Just because Mary is not acquainted with red inside the room need not entail that red is not a physical property. She just gains a new way of knowing certain things, whether they are physical or not, and this way is only achieved *while* partaking of the acquaintance.

What all the accounts of the PCS have in common is their attempt to articulate how the notion of a concept may effect intimate relations with phenomenal qualities in closing the explanatory gap, while making no progress on the metaphysical nature of those properties such is required if progress is to be made on the hard problem, but this notion of acquaintance knowledge which is compatible with physicalism may point the most reasonable way of settling the knowledge argument.

The intuitively attractive category of 'knowledge by acquaintance' described by Russell and Pitt is a sort that can only be gained by having an experience. It seems a cogent thought that Mary was born blind and learnt all the physical facts about colour and vision through braille, yet just because she will never know what it is like to see anything, it would not follow that physicalism must be false. This need not be knock down against the knowledge argument because it skirts over a lot that has been considered in previous and coming chapters but it does alert us to what seems to be an obvious form of belief, even in perhaps the weakest case of an animal looking on the world. This sense of acquaintance seems in line with how Wittgenstein understood the universe as the totality of facts in the sense of being part of what constitutes reality and not just what can be

written down in books (Crane 2019). These sorts of facts cannot be learned or conveyed but only experienced and it would seem obvious that physicalism need not be required to specify all such facts. This approach makes room for the thought that we can learn about things we might already know about in a new way, by seeing them in a 'different light'. Terrence Horgan (1984) responded to the knowledge argument in this sort of way to resist the idea that Mary's new experience cannot be about something physical which she already knew about as is usually thought to be shown by the classic Frege example of the morning and the evening star.

More will be said in later chapters about why physicalism need not capture all that can be specified about physical stuff and how this might relate to science. For now, I will leave it as an intuitively appealing way that physicalism might not be threatened by the knowledge argument. Galen Strawson (2018) takes essentially the same line by driving a distinction between physicalism and "*physics-alism*" (2018, p.125), and argues that the knowledge argument only defeats the latter which would portend to capture all knowledge as grounded or made true by physics.

Chapter 6. Donald Davidson and Non-Reduction.

Non-reductive materialism is a category of position on the mind body problem which in various ways aims to establish a view that mental properties although constituted by physical matter are causally efficacious and not reducible to physical properties. It might be characterised as an attempt to find a way between reductive materialism and dualism. Donald Davidson's 'anomalous monism' is a famous early attempt of this. I must attempt a slashing economy in specifying basic aspects of Davidsons otherwise very sophisticated ideas, to the extent Robinson features them in his broad negative thesis against physicalism. Robinson argues that Anomalous Monism is reductive and Davidson's use or understanding of the notion of 'non-reductive' is unclear and that ambiguity has encouraged regular misunderstandings of the notion. Chapter nine will involve fuller discussion of reductionism.

Robinson distinguishes the two main senses of reduction which have not plausibly accommodated mental properties into a physicalist ontology. The first sense which is specific to philosophy of mind is the attempt to account for the mind in behavioural or functional terms, where psychological predicates for properties of first person sensations are analysed as dispositions to behave or functional states defined by their contribution to behaviour. The second sense is by scientific reduction of psychological properties to properties of physics. This involves a unified understanding of science to the extent that there exist bridging laws through the physical sciences of biology, chemistry and neuroscience, all reducible to physics so that the constitution of psychological states can be made out in physical terms.

Robinson describes the (British) philosophical culture as resisting what it viewed as the counterintuitive theories from Ryle, Smart and Armstrong (2016, p95), which represented those two reductive paradigms, and which understood Davidson's 'Mental Events' (1970) to be an attempt at an alternative *non-reductive* theory. Davidson used an ontology of events, whereby each physical event is also a mental event, and both events are constituted by the same causal relations, which seemed to avoid the requirement of reducing mental

properties to physical properties (more on this below). Robinson argues that Davidson does not avoid law like reduction in a way which requires an analytic reduction like functionalism. I will not argue with that reading of Davidson, but I will dispute some of Robinson's analysis that could affect the conclusions he draws.

Robinson offers a picture of Davidsons ontology of events as a 'flagpole' theory of events, whereby a single event (the flagpole) has a physical aspect and a mental aspect (two flags flying). A single event can instantiate physical brain properties and the mental properties of experiencing a quale, just like the flagpole can hold two flags of different colours. The single event is identified by its cause and consequences, and not by the noncausal properties of the event in the way the flagpole identity is not dependent on the colour of the flags. Davidson's theory is widely held (Yalowitz 2019, sec 6.1) to render the mental properties of the event as epiphenomenal because the physical properties of the event would do all the causal work for physical effects, leaving no causal involvement for mental properties (Honderich, 1982). Howsoever the event might be described, it is only those physical properties that are sufficient to make physical things happen, leaving mental properties unnecessary for causal processes (we should note that Davidson (1993) himself was alert to this scientific principle of physical causal closure, but that he understood the causal relation to fall under different metaphysical principles). This principle of 'no overdetermination' is also widely supported by philosophers (Robb & Hale 2019, sec 2.4), leaving livelier dispute around related notions of explanation, to which Robinson attends as we will now see.

Robinson cites this charge of epiphenomenalism in two ways, the first of which is "…the irrelevance of mental states or properties to causal explanation is itself constitutive of epiphenomenalism, for interaction involves causal explanatory relevance." (2016 p.97). Robinson argues for the causal relevance of mental properties based on their explanatory value, which I will argue is unsuccessful. The cogency of my claims need not challenge his conclusion that Davidson's theory is effectively a reductive functionalism, but they may inform our view of Robinson's positive thesis later (which as far as I can tell, is why Robinson is motivated to present a limited defence of anomalous monism).

Robinson claims the "accusation" (p.98) which I have quoted in the last paragraph is rendered inconclusive. He appeals to an aspect of Davidson's later developed thesis (1993) which will save it from the sort of objection which I referenced above about physical properties making mental properties causally redundant. Davidson says 'strict' laws may only operate at the level of fundamental physics, and while it is only with such laws at the level of physics or those sciences nomically reducible to physics that we can generate strict causal explanations, now Robinson:

> "There are looser forms of causal explanation at higher levels, including the psychological, and so it becomes just as proper to explain causally human action by reference to human thought as it is, for example, to explain the destruction of a village by reference to the force of a hurricane, given that meteorology is not nomically reducible to physics. In neither the psychological nor meteorological cases are there strict causal explanations…but there is what we intuitively recognise as causal explanation. If Davidson's theory allows our mental states to be casually on a par with hurricanes, it would not seem just to accuse him of epiphenomenalism." (2016, p.99).

I do not offer interpretation of Davidson's theory, my claim is that Robinson is not entitled to cite equivalent 'properness' of causal explanation across the meteorological and psychological domains (regardless of nomic reducibility), to the extent needed to justify his inference. It is not the case that mental properties can properly feature in a causal explanation of human action in the same way that hurricanes feature in a causal explanation of the village gone. We can take it that a 'hurricane' is a convenient reference to a collection or sum of physical properties which constitute that hurricane. It follows there is nothing counter intuitive by claiming that the casual consequences of the hurricane be identified with the physical properties of which it consists. In the case of the mental event, Robinson requires the reference to the human action to be other than an appropriate reference to the collection of physical properties which constitute that mental event, and he has given no reason why it need be anything more than a convenient reference and so the original objection is still in play because the mental property explanation is *causally* grounded in the efficacy of the physical properties.[6] The psychological explanation will not support an inference for reifying ultimately

[6] My objection is prompted by Jaegwon Kim's problem of explanatory exclusion (1989).

nonphysical causes, in contrast to the meteorological explanation which is consistent with purely physical causes and so the accusation of epiphenomenalism appears just.

Robinson discusses another way the epiphenomenalism accusation might arise because of Davidsons unclear presentation of how exactly mental properties are supposed to supervene on the physical. Robinson is of course alert to the distinction I make above between causation and explanation and says that psychological properties must be nomically reducible to the physical in the same sort of way as special sciences or else the monism of Davidson's theory is threatened: "mental properties are just high-level descriptions of the array of physically present signals,…Davidson has shown that there can be different descriptions of the same subject matter that are not nomically related to each other, without this rendering any of them wholly otiose in causal explanations…" (2016, p.101). Davidson's theory in this way is rendered ontologically reductive which is the sense we are interested in and so fails as a non-reductive physicalist theory.

Robinson (2008) claims we find clarity on Davidson's position from his 1987, 'Problems On The Explanation of Action'. Here we find the thesis that normativity is essential to psychological explanations but not to the other special physical sciences, but even if that were true (which Robinson argues is unlikely), we find no explanation of how psychology is reconcilable to monism (2008, p.141). Davidson's position collapses into a form of non-realism about mental properties, in a way which supports my earlier point that Davidson's mentalistic causal relations are only language dependent, "…the mental is not an ontological but a conceptual category…To say of an event…that it is mental, is simply to say that we can describe it in a certain vocabulary – and the mark of that vocabulary is semantic intentionality." (Davidson 1987, p.46).

Robinson goes on to make the case that this sort of irreducibility stemming from the essential normativity of psychology inspired an approach to the mind body problem which he labels 'Naturalism without physicalism'. These theories are enough off target not to detain us, so I will just mention them for completeness due to their historical significance: John McDowell (1994) '*Mind and Word*', H. Price (2011) '*Naturalism Without Mirrors',* and Rorty (2010) 'Naturalism and Quietism'.

Chapter 7. Mysterianism, Neutral Monism and Panpsychism.

We may view the topics of Chapter seven as minority positions on the mind body problem. For that reason and the space available I must proceed to more central concerns about arguments for physicalism entailing epiphenomenalism and then Robinsons positive case for substance dualism. With a previous draft I attempted elaboration of the issues in this chapter, but there is not the space for worthwhile philosophical insight, which is anyway least relevant to Robinson's thesis compared to the areas yet to be covered. I will summarise what Robinson argues from chapter seven with simple mention of theories for some level of completeness.

Standard mainstream physicalist accounts that we have considered before this chapter cannot account for the qualitative aspect of the conscious mind. McGinn (1989) argues that the constitution of our minds means we are cut off from understanding the mind body problem which will remain a mystery. Stoljar (2006) argues that science cannot exhaustively probe the nature of matter which would enable us to understand how conscious qualities are constituted. Bertrand Russell (1927) invented the theory called 'neutral monism', around the idea that physics can only investigate *relational* properties of matter, but not those *intrinsic* properties that may somehow include or constitute phenomenal properties. Galen Strawson (2006) and Philip Goff (2017, 2019) defend the thesis of 'panpsychism' which build on Russell to make conscious properties intrinsic to all matter. Robinson judge's panpsychism to suffer numerous problems which he deems insurmountable.[7]

[7] Bibliographical note. Considering the lack of progress made by traditional approaches, I register Goff (unread by Robinson at time of his writing) presenting a plausible research program.

Chapter 8. Conclusion of Part One.

This chapter is somewhat pivotal for Robinson's overall treatise. He presents his concluding treatment of the nature of qualia with arguments effectively linking his rejection of physicalism to his license for substance dualism. The standard physicalist accounts I have reviewed in this chapter have fallen short in their accounting for Marys new experience. Robinson concludes with an attempt to 'strengthen' the knowledge argument and achieve a more radical conclusion that physicalism is incoherent without this accounting. I object to his arguments.

Standard forms of physicalism considered so far have not presented a satisfactory response to the knowledge argument and Robinson will argue that physicalists have not appreciated the comprehensive force of the argument. He presents the dialectical situation to be one where the knowledge argument may support property dualism to the extent that physicalism adequately accounts for non-conscious reality which constitutes "almost 100% of the universe" (2016, p.133). But that the physicalist account struggles with that qualitative aspect of consciousness as Chalmers calls 'the hard problem', from which it would follow that qualia have "nothing to do with our robust conception of the physical as it applies to the vast mindless tracts of reality". (p.133). As I understand Robinson, He reads this situation as providing physicalists strong prima facie reason to expect some sort of physicalist account to theoretically absorb that extensionally miniscule aspect of the universe, which is a strange way of book keeping ontological categories, but I think he is speaking as to the inclination of the physicalist mind set. Jaegwon Kim (2007) is a notable physicalist who somewhat concurs with this big picture rendering of the situation, naming his influential book "*Physicalism or Something Near Enough"*. Robinson claims this dialectic radically misrepresents the actual situation because we *should* attend to the conception of physical matter rather than the mind.

Robinson explains that because science is mainly concerned with measurement and expressing its findings in mathematics, the result is an abstraction which cannot wholly capture our common-sensical conception of

the physical world and to this extent our abstract scientific concept of the physical is insufficiently concrete.[8] What is needed to concretise that abstraction is the addition of qualities – essentially sensible qualities – that figure so importantly in our naïve or common-sensical conception of the world. These are essential to our ability to 'cash' or 'model' or 'interpret' the abstract, mathematical conception." (p134). Robinson will argue with this idea that we can generate a much stronger understanding of the hard problem in terms of the knowledge argument. Whereas the normal takeaway from the knowledge argument might be that physicalism cannot explain qualia, Robinson wants to argue in addition that any physicalist conception beyond what can be abstractly and mathematically expressed must essentially feature qualia, and he takes this to show that physicalism is not, "merely incomplete, failing to cope with consciousness, but something more like incoherent, because it cannot give a coherent account of the physical itself." (p.135). Robinson claims that his arguments show that not only if the knowledge argument is sound then physicalism cannot capture qualia, but that all known attempts to refute the knowledge argument are undercut, because they all rest on a problematic conception of the physical.

His argument is that the knowledge argument generalises to include those primary qualities of objects which characterise the world aside from our perception, and not just the secondary qualities which are essentially subjective. He takes 'squareness' as an example of a primary quality which unlike the secondary quality 'red' is not defined in terms of what it is like to perceive it, but neither is the definition *independent* of our perception of squareness, for otherwise our conception would be "wholly axiomatic and mathematical." (p.135). He presses that while secondary qualities attach to a particular mode of experience and primary qualities do not, they must however be some way experienced by sight or touch in our case, or else he insists, we would have no conception of spatial properties beyond the abstract. He claims that it follows from this that physicalism without sensible qualities would lack any empirical content.

My objection is that physicalism as a doctrine need not include any specification of how humans come to learn about physical stuff or how they conduct science. Physicalism does not require a rationale about how

[8] Goff (2019) emphasises this backdrop to science and argues it has plagued our understanding of physics since Galileo.

we acquire knowledge. It will be interesting to locate Mary in relation to Robinson's idea (although it is unnecessary for my basic objection). It is one thing for Mary to be cut off from colour and so ignorant of what it is like to see colour, but we would have to imagine Mary born limbless in a pitch black sensory deprivation tank with only auditory stimulation for Robinsons idea to gain any purchase because we would need to deprive her of perceptual connection to spatial properties, and then hypothesise that scenario Mary has learnt all the physical scientific facts about spatial properties which seems a contradiction. Even if we imagine it could happen and Mary became fitted with a bionic arm, she would of course be surprised at her first experience of touch. This would not support any intuition against physicalism like Jacksons Mary might. Robinsons idea is driven by notions of sense *capacity* rather than any idea that everything is ultimately physical. There is no reason why physicalists need deny that knowledge can be gained only through experience *of some sort*.

Robinson is not fazed by my suggestion that it may be impossible to imagine experience like ours without spatial features, because some sort of particular perception is necessary for an empirically contentful conception to go beyond a purely axiomatic concept (2016, p.136). This amounts to psychological speculation which will not defeat my simple objection that physicalism is secure even if he is right. (His claim in foot note 1, that the potential impossibility of a developed mind like sensory-deprived-Mary is not to the point, is simply an attempt to instil the intuition that such a situation is possible, which seems redundant if such a case is beside the point!). There is some indication that Robinson is illegitimately equivocating physicalism with a descriptive-account-of-reality, which might drive the fault in his argument here against physicalism: "If, in general, the acquisition of experience did not teach something new, then a purely descriptive account of reality ought not to lack anything essential…physicalism that depends on a notion of the physical that is somehow independent of the qualitative nature of experience can only present us with a world that is so formal as to be empirically contentless." (p.137). To repeat my objection: physicalism does not require a rationale about how we acquire knowledge.

I take the rest of chapter eight to be trading on this equivocation or something close enough which confuses empiricism as a doctrine about how we only acquire knowledge through the senses, with physicalism as a

doctrine to the effect that everything conforms to the condition of *being* physical. Relatedly, a parallel knowledge-argument-for-primary-qualities does not get off the ground because there is no sense to the starting hypothesis to the effect that Mary could know all the physical properties of space before leaving her sensory deprivation tank. We can partially accept Robinson's conclusion to part one that the standard physicalist responses to the knowledge argument which I have reviewed up to Chapter Five are unsatisfactory and that qualia are essential cement for our empiricist understanding of the world, but he has not shown that physicalism is incoherent because of that.

This is the end of part one. I am persuaded that these standard physicalist responses to the knowledge argument which I have surveyed are not wholly adequate to answering Robinson's take on the hard problem as conveyed by the knowledge argument, while the notion of knowledge by acquaintance might hold the resources to fend of the threat to physicalism. In part two I will shift from researching other philosopher's positions to engage Robinson's arguments for 'conceptualism', which he proposes as a step on the way to dualism.

I next analyse Part Two of Robinson's treatise starting with chapter nine and consisting of four quite independent arguments for why physicalism entails epiphenomenalism for mental properties. I allocate weighted space to each chapter according to the extent I judge they illuminate the mind body problem and within the space this dissertation allows.

PART TWO.

Chapter 9. Reductionism and the status of the special sciences.

In this part of his treatise Robinson argues from several approaches that phenomena which are not metaphysically basic, are dependent on the mind. In this chapter Robinson presents a thesis about reduction which entails either that the properties of the special sciences including psychology are epiphenomenal, or they are formed by an ineliminable interpretation which an observer conceives from the objective physical base which places the mind outside of the physical, amounting to dualism. Chapter ten and eleven develop an account of vagueness and composite objects respectively, upon which his argument for the mind dependence of non-basic entities will also depend. I judge that his argument in chapter nine is clearest and most germane and relevant for now and the last sixty years of research on the mind body problem (see McLaughlin et al, 2011), which I will apply as a sort of meta-justification given word constraints for relatively less focus on vagueness and composite objects. I will begin by presenting a basic groundwork for the notion of reduction, to frame Robinson's explanation of the problem and to the extent that it is important for a physicalist solution to the mind body problem.

One may picture the physical sciences as consisting of different levels of description for a physical object. Let us take an oak tree: physics, chemistry, and biology will each cite properties of the oak tree using proprietary vocabularies, while referring to that one chunk of physical reality which is the tree. A core issue For Robinson is how we understand the relationship between these different scientific descriptions and associated ontology at each level, given that they all concern the same tree. A usual claim is that physics cites the fundamental description to which the special sciences are reducible.

Reduction between the sciences is understood as nomological (lawful) reduction; physics and chemistry both are taken to specify physical properties which constitute the physical matter of the tree, with each level of description connected by bridging laws that connect the laws of each science. Nagel 1961 is the classic exposition of this model, from which I take bridging laws to be like algorithms which can relate properties between the domains. This is a swift description of scientific reduction which Robinson explicates over many pages in chapter nine, where he also elaborates links with other historical notions of reduction and purported compatibility between them.[9] The crucial question for us is how psychological properties might supervene on physical properties or whether they are reducible to the physical base ontology. We will best concentrate on that discussion which is most up to date with contemporary theory where he describes, " 'A priori sufficiency of the base' reductionism" (2016, p.153).

David Chalmers (1996) calls this type of reductionism, 'logical supervenience', which is intended to capture a weak version of reduction for mental properties like qualia supervening on the brain where nomological reduction cannot be seen to be feasible. This sort of minimal supervenience between the properties cited at each level of description can be seen to characterise a relation of non-reduction, but a weaker relation of *dependence* of qualia on the brain. For illustration:

> "There is no logically possible world which, at the level of physics, is just like one in which a hurricane is destroying a village, but in which a there is not a hurricane destroying: the physics base is a priori sufficient. There is no need to invoke some elusive conception of supervenience here: in the broadest sense of 'logically possible', there is no possible world with the same physical base as the given one and no hurricane; the relation is one of entailment in the strongest sense…[in the same sort of way]…If some version of functionalism were correct about the mind, having atoms arranged just as they are on earth would be logically sufficient - though not necessary, for there might be other ways of making minds – for the existence of conscious beings." (2016, p153).

Logical supervenience only requires a dependence of *facts* between the levels where there can be no difference in dependent level facts without a difference in base level facts and so constitutes a logical entailment, which contrasts with nomological supervenience involving a causal dependence between levels.

[9] I will resist going over all that which in any case is technical and controversial and instead attempt to later bring out those aspects crucial to his argument.

A place holder term which covers all the ways in which properties may be reducible whereby any non-fundamental properties depend on the fundamental properties of physics, so that any difference in base properties means a difference in chemical and biological and psychological properties. Supervenience by itself only indicates a dependence or covariance relation (McLaughlin & Bennet 2018, 3.7). It will only indicate the *direction* of dependence by saying biological properties supervene on chemical properties or the colour of a painting supervenes on the paint, but the way supervenience occurs is presumed to be radically different in each case and so it may be taken as placeholder for deeper but unexplained relations.

Robinson proposes 'explanatory irreducibility' as a concomitant of logical supervenience; in the case of the hurricane which is constituted by nothing other than physical particles, it is the case that explanation at the level of hurricanes is not nomically reducible to explanation at the level of physics, which is the same situation regarding all the special sciences in relation to physics. Robinson pairs this explanatory notion to the ontological notion about the supervenience of properties. Properties of the special sciences are said to logically supervene while exhibiting independent explanatory value, which we may see as "…just different, higher order *ways of describing* the base subject-matter." (p.155). He claims the plausibility of this way of viewing the different explanations as motivating a *conceptualist* approach (which will become central to his thesis), and which supports a position of 'weak property emergence'. He explains contrasts the weak form with 'real property emergence', where a property has "no conceptually sufficient conditions in the base for its exemplification." (Ibid). It is this position we are told which characterises a realist dual aspect theory in the philosophy of mind, which rejects the logical supervenience of psychological properties and accepts the explanatory gap between the mental and the physical. He positions the 'non-reductive physicalist' as leaning on the attempt to create a sort of mid-way position between weak and real property emergence, "Real but supervenient property emergence…a property with no conceptually sufficient conditions in the base for its occurrence, but with some stronger dependence on that base than merely causal…"(p.155), which he says widely confuses what non-reductive physicalism amounts to through a misunderstanding of logical supervenience and he implies that this confusion is related to the issues I developed from Davidson in Chapter

six. Robinson claims to have made a plausible case that logical supervenience is the only credible notion of reduction, and he will derive some consequences from the limitations of logical supervenience. (I will not go over his earlier positioning of different notions of reduction, which in any case is insufficiently technical to be convincing. Instead I will try to rely on narrow aspects of his case that suffice for his argument). I will now consider and partially object to these purported consequences from which Robinson derives an opening for dualism.

Fodor (1974) denies that logical supervenience which characterises the dependence of mental on physical properties is *non*-reductive yet need not undermine a physicalist ontology, "because each instance of a higher-order concept will be identical with some structure describable in terms of basic physics, and nothing more. This token reductionism is all that physicalism and the unity of the sciences require." (2016, p.156). Token reductionism simply means that the sum of fundamental physical items in the base is identical with some set of physical items which physically constitute those mental or chemical properties. Robinson claims that "contrary to appearances, this may wrong" (Ibid), and physicalism may require more than token reductionism and something more like nomological reduction.

Robinson accepts that Fodor is right that the same physical stuff might be irreducibly described in different ways, but that Fodor does not explain how this is possible. Unlike nomological reduction which might explain how this can work, with logical supervenience it seems there must be something more over and above anything that can be identified with the base. There are two options he tells us, either there are new properties which do no causal work because (in parallel to what we found with Davidsons event ontology) the physical properties exhaust causal operations, and this would render psychological properties epiphenomenal just like properties of the special sciences, "and hence the fear that physicalism about the mind does not avoid epiphenomenalism."(p.157). Before we move to the second option, it seems that this just states the mind body problem which everybody takes to be mystery, which is not precisely an option as such for the physicalist (at least by the large majority who reject epiphenomenalism) as much as it is a problem to solved. My point here would be trivial if it were not that Robinson reads the situation as somehow saddling the physicalist with

a second option instead of working on the status of the problem as so far stated which as far as I can see is how the research tradition carries on. My point is perhaps dialectically accommodated by taking Robinson's point more like an attitude whereby he sees the old research program at a complete dead end.

The second option is worse for the physicalist and which the rest of his treatise will argue for in favour of the first option ['anything but a resigned epiphenomenalism!']: "The special sciences look like a 'top-down' conceptual interpretation of the base. This seems to suggest an interpreter or conceptualiser who views the world from his own perspective: "They are more like a perspective from outside on the same subject matter." (p.156). I will extract the contours and analyse this position which Robinson calls 'perspectivalist'.

Robinson characterises Physics as cutting reality at its ultimate joints along with any special science which is nomically reducible to physics cutting reality at larger joints such as cosmology. Those sciences such as propositional attitude psychology which are nomically *non-reductive* and only supervene logically he says involve a necessary interpreter because they only emerge because of an interaction between an observer and real patterns in the physical world. Robinson claims that this dualism of mind and body may only be metaphorical if the mind can be treated as part of the physical domain which is being interpreted, otherwise the mental acts must be grounds for ontological dualism. Robinson proposes three features of the special sciences that support his perspectivalist justification for ontological dualism. All three seem to involve in some way the dubious notion uncovered in the last chapter which illegitimately implied physicalism must account for *how science is performed.* This is indicated in the general framing of the three features: "If scientific realism is true, a completed physics will tell one how the world is, independently of any special interest or concern: it is just *how the world is."* (p.157). It seems that perspectivalism is generally motivated because objective facts must be specified in a language, but this is only a necessary element of *communication* and not of the elements of the world (however specified) which make those facts true.

The three features of the special sciences are:

(i) They are selective: their subject matter concerns only parts of the world such as organic chemistry being only about living things, in contrast to the entities of physics which constitute everything

physical. I am not sure where to start with this item; it is a methodological tautology of science that it is performed by scientists and if there were no practical limitations or perhaps we had 'all-seeing' technology, then our scientific categories-or-selection might be very different, yet they would target the same objective truth-makers in the world (whatever makes our facts true). This item just makes the scientist essential to doing science and is no threat to physicalism.

(ii) The special sciences are teleological, or interest driven. That scientific study is interest driven means nothing more than scientists aim to explain phenomena which is interesting, or which provides some practical facility. Other than science inevitably falling short of explaining *everything,* I fail to any significance for ontology here. That scientific explanations use time as a background metric is enough to qualify them as teleological and is simply an essential feature of our interaction with the world, and anyway an arbitrarily large portion of special science is concerned with synchronic phenomena. A potentially more interesting observation is that science individuates processes based on targets of interest rather than essential start and end points in nature. Robinson marks this to be true of the whole of biology and medical sciences which consist of *non-atomic* entities. We might strengthen this claim in noting that atomistic physics reduces to quantum level physics consisting of events which science individuates for abstract explanatory convenience rather than any conception of objective start and end points. In any case this second feature seems to be an important feature of explanation rather than ontology, perhaps related to Davidson's mental events being explanatorily irreducible but causally impotent over and above the physical properties which constitute them.

(iii) Many of the entities of special science are *Gestalt*-like phenomena. Robinson means that we as perceivers determine parameters of physical similarity between entities, which might not reflect similarity in the way fundamental particles in physics exhibit exact sameness. An example to illustrate this point could be two electrical storms only similar to the extent the weatherman catalogues them. Robinson is not claiming that the storms being similar is mind dependent. There are he says using Dennett's phrase 'real patterns' in the world, "but, like *Gestalten*, they are reified as being of a certain

kind by an interpretive act." (p.158). We may object to his crucial use of '*reify*'. The meteorologist will not claim that a storm is anything more than a sum of basic physical properties which for explanatory convenience we label as a storm, and which is objectively continuous with elements that 'surround' it or with which it borders.

Robinson believes that these three features of the special sciences suggest that logical supervenience,

> "presupposes a perspective on the subject matter, which is the viewpoint from which these interpretive acts take place. [nomological] reductionism enable one to understand the special sciences in the light of physics without the addition of such interpretive perspectives ('bottom-up' theories). But the perspectival approach, involving a view from outside the subject matter, seems to give the interpreting mind an *irreducible* role in the creation of these sciences." [emphasis is mine] (2016, p.158).

I emphasised only the word '*irreducible*' to highlight that part of his claim which is either not justified if he means to posit an ontological dualism of mind, or isn't clear if he means by 'role', '*what the mind is doing*' is irreducible, because no part of his argument targets the notion that the diachronic process partaken by the conscious mind cannot logically supervene on the physical-realizer-brain-process compatible with physicalism. Indeed the whole claim would be made compatible with my earlier comments if we substitute 'irreducible' for '*ineliminable*', because all that we might rescue from Robinsons argument is that the *scientist* is an essential causal node in the formulation of those sciences which logically supervene on physics, and this no way offends physicalism.

Robinson presents what he takes to be, "Perhaps the clearest case from within science of the role of interpretation and imposition by our interests is the importance of function in the biological sciences." (Ibid). Robinson accepts the conformity of biological explanations (that are essentially functional by only making sense in relation to teleological concepts like survival), with the logical supervenience of the mechanistic base in some way accounting for the operation of a biologically understood feature. So far so good. "But to understand it within biology you need to know what it is for. You need to know both what *role* the liver plays and *how* it does it. The latter question would not arise if the former matter were not an issue." (Ibid). The functional explanation will speak to things we are interested in, without in anyway revealing the fundamental

physical process that is ultimately responsible for the higher-level description. Robinson claims that this appears to make Fodor's theory of non-reductive physicalism in its equivalence with logical supervenience essentially dualistic, because it allows the same piece of physical reality to be viewed from outside in different ways for different explanatory requirements. I cannot see how this example needs be judged differently than I covered with my previous comments, so Robinson has *not* shown in any *ontologically* meaningful way that (because science cannot eliminate what he calls the 'external perspective') "the interpreter must transcend the physical world that he is interpreting" (p.159).

In the next chapter Robinson mounts another argument for dualism from the nature of ordinary language.

Chapter Ten. Vagueness, realism, language and thought.

Robinson takes the previous chapter to have undermined physicalism either through rendering entities of the special sciences epiphenomenal, or by placing the mind outside the physical system. I judged his arguments for the latter to be unclear by lacking the ontological import he seems to invest in them. He will present further arguments that 'place the mind outside the physical' in this chapter from considerations about language but I first want to clarify what we must take from chapter nine in regards the first claim about special science epiphenomenalism.

There is an important distinction to be made between the entities of psychology and those of the other special sciences in terms of their potentially undermining physicalism. The entities of the special sciences do not undermine physicalism to the extent that it is unproblematic if the causal activity for example of hurricanes or genes or the chemical compound stability is completely constituted by their physical base properties. That is, any lack of understanding about exactly *how* the special science properties are constituted by the base properties, will not put pressure on the acceptable notion that a sum of base properties is doing the causal work that the higher level sciences refer to as a collection within their explanatory domain. In other words, those explanations which logically supervene on the physics are expected to be compatible with the exhaustive causal operations of the base. This is not the case for psychology which requires that mental properties are causally potent over above the causal operations of the base and so they must supervene differently to the other sciences if psychological explanations are true because they require mental properties to be casually autonomous unlike special science properties. So, Robinson cannot claim that physicalism is undermined by epiphenomenalism in *all* the special sciences, but only by epiphenomenalism in the psychological sciences. The project then for non-reductive physicalists is to explain *how* mental properties can be causally potent *as* mental properties, because they cannot appeal to token identity with physical realizers as is unproblematic for the special sciences.

In this chapter Robinson further presents his thesis for 'conceptualism': that all non-basic entities are dependent on a perspective which entails that those entities are mind dependent. This thesis was clear enough in relation to how we might view entities in the special sciences and the utility of scientific perspectives but failed to show any ontological upshot. His argument is somewhat parallel in this chapter, going from the logic of language and the use of vague predicates for non-basic entities he argues there cannot be comprehensive logical equivalence between ordinary language and discrete non-vague predicates in use for basic properties. He aims to show that natural language is necessary to engage with the world and therefore non-basic concepts are ineliminable from a proper understanding of the world and this justifies ontologically reifying the perspective which generates language.

Like I did for Jackson's representationalism in chapter four, I cannot fully engage all aspects of Robinson's treatise for lack of room and so I must cull those elements which may be undercut without much treatment because they are least illuminating. Further, it seems impossible that we can generate ontological conclusions from the nature of language (as we could not from scientific language) and so I will pass over the topics of chapter nine.[10]

[10] Some further 'practical' justification for passing over chapter 10 comes from a broad stroke objection to ideas contained within it from Barry Smith and Michael Martin in a long footnote on page 176. Robinson retreats to a summary view that, "linguistic representation of the world is usefully messy."

Chapter 11. Composite objects, the special sciences, conceptualism, and realism.

Robinson provides a conceptualist treatment of mereological composite objects which obviously fit his category of non-basic entities. Before briefly surveying his ideas about composite objects which are most intriguing or perhaps most suggestive of what Robinson is after, I will reaffirm and clarify how I understand conceptualism in relation to physicalism.

Physicalism does not conflict with a realist view about non-basic entities conceived by the special sciences (except perhaps first-person conscious properties which Robinson does not treat 'conceptually'). Those entities or properties which are conceived by science are by *physicalist* lights an amalgamation of physical entities constituting any entity useful to science and therefore compatible with any explanatory convention or theory of science as discovered and defined via subjective identity conditions depending on the interest of a scientist tracking elements of physical stuff (this long winded paragraph captures all conceptualist candidates that Robinson offers by undercutting all argumentation (in chapter eleven up to section seven) for the claim that "the composite does not exist in a realist sense, but only in a conceptualist one…the best that a physical realist can do." (p.178)). Ignorance of a lawful relation between physics and the special sciences or an actual *lack* of lawful relation (Kim, 2012) will not offend physicalism, which only need posit token identities related by logical supervenience. A supporting idea which cleanly removes any potential complication about the nature of mind instigating non-basic entities and forestalls his claims about non-realist meaning, is that truths about those entities will be made true by a set of fundamental physical truths.[11] Ahead in Chapter thirteen Robinson considers Daniel Dennett's physicalist methodological position called the 'intentional

[11] The plausibility of my all-encompassing undercutting of the arguments and logical mapping in chapter eleven (sections 11.1 to 11.3 inclusive) is strengthened by:

a) Robinson's bracketing off that modern portion of contemporary metaphysics (p.180) associated with a classic anthology *Metametaphysics,* edited by Chalmers (2009).
b) Robinson's concession that the conceptualist/realist distinction is ambiguous compared to (even the rough) difference which Hume characterises by worrying inside and outside of his study. (p180).
c) Robinson believes it unhelpful to bring in formal semantics for questions of reductionism, the operations of vague predicates or whether predicates should be treated conceptually or realistically (p181).

I take these simplifications combined to help clear the way for the 'undercutting' I have proposed.

stance', which exhibits some affinity with conceptualism and so will be interesting to see how Robinson saddles up to Dennett's view.

Robinson proposes two different and intriguing supports for conceptualism. The first explicitly challenges my previous proposal by arguing for the causal inefficacy of non-fundamental levels (p.181).

My line above holds to Jaegwon Kim's position (2007), which Robinson labels the 'common sense' position on the causal efficacy of special science entities: that the causal powers of a house are constituted by the bricks that make the house. Robinson claims that even if this is correct in terms of causal powers, the truth that the house *exists* is neutral between conceptualism and realism and conceptualism about the existence of the house wins support from *identity conditions* for the house:

> "If there really is a table there and not just atoms arranged table wise, it could, at some time have been constituted by at least a few different atoms from the one's that actually constituted it. This shows that there is a real difference…something grounds the identity condition difference. The conceptualist has a non-mysterious account of this: it is grounded in the nature of the concept 'table'. For the realist to respond 'well, it is a table, isn't it' does not seem to be explanatory." (2016, p183).

I suspect many controversial issues around identity could be relevant here for which I am unqualified and to which Robinson only nods with mention of Armstrong on universals, and so with another appeal to lack of space I will attempt adequate response as follows: physicalism does not care what conceptual convention we use to determine the correctness of whether one table is the same or different from another and by definition atoms are qualitatively indiscernible one from the other and so substituting one or any number of atoms in a table will render the altered table qualitatively indiscernible from the unaltered table and absolutely the same in terms of its potential and actual casual powers. Further, if the changed table is only by convention considered a different individual, then we are brought back to the notion that this will only show that a thinker applying a concept is a necessary component of that process, and that does not threaten an ontologically physicalist view of artefacts.

The second *sort* of argument that Robinson presents for a conceptualist treatment of composite objects (I take his arguments up to section 11.7 to be of a sort captured by my comments above) comes in the form of a "knock down argument" (p.190) based on the strict falsehood of Newtonian science:[12]

(1) Either or both of relativity and quantum theory are broadly correct.

(2) The Newtonian picture of the world is strictly false, though workable because it approximates to the quantum/relativistic facts on all but the very small or very large.

(3) Properties ascribed in the special (physical) sciences (involving space or time) are an essential part of the Newtonian, as opposed to quantum or relativistic pictures.

Therefore

(4) None of the properties ascribed to objects in the special sciences are strictly true of the world.

(5) Concepts that are workable but do not correspond to properties that exist in the full realist sense (that is without conceptualisers) fit the definition of the merely conceptual.

Therefore

(6) The concepts of the special sciences are to be understood conceptually, not realistically.

Robinson takes it that all special science properties are hence trivially over simplified to the extent that Newtonian properties are a simplification, or else we must say that special science properties are wholly real features of macro reality, not of micro reality, but we cannot say that because they are fundamentally false. I laid the simple argument out in full so the reader can judge whether I am missing something. Even if it is the case that special science concepts do not coincide with amalgamations of fundamental aspects or properties of space time or quantum fields, why can't we just live with physicalist notion that they are not actually real in some ultimate sense but they are useful approximations rather than Robinson's insistence that they are made real by our conceptualising? This need not invite anti-realism about the features of the world picked

[12] Author remark: As mentioned back in the introduction I chose Robinson's treatise as a 'ready-made' program of study for its breadth and clarity. I was also attracted by its boldness which encourages non-expert engagement with the arguments (even while on a steep learning curve). Whether Robinson's thesis is sound or only suggestive, I admire him for sticking his neck out.

out by special science or by our use of artifacts, instead we may remain untroubled that we do not have a god's eye view of the comprehensive physical specification of everything.

Robinson does not argue against Kim's common sense position on causal powers, and so it seems conceptualism might only amount to a verbal dispute about what to call things, if only Robinson didn't repeatedly insist on a substantial difference between real entities and those made real by our conceptualising which is at best unclear.

In the next chapter Robinson extends his conceptualist treatment to fundamental physical objects by proposing that there are no real physical individuals because they cannot sustain counterfactual truths of origin. I will provide only brief and summary treatment adequate for this dissertation.

Chapter 12. Why there are (probably) no physical individuals.

Robinson argues that any physical individual down to the atomic basic level will necessarily have vague counterfactual identity conditions, and so there will be no fact of the matter pertaining to their identity outside of our concepts.

Various scenarios are considered: such as questioning whether a five percent change in the material that makes up a table would render it a different individual, to questioning if a particular individual atom at the beginning of time when it was created had a slightly different location would be the same individual or not. A somewhat summary dismissal of contending theses from Locke, Wiggins and Salmon, help motivate the idea that such vague identity can only sit within a representational ontology which may only be understood within

a conceptualist framework, and that only proper non-physical individuals such as *minds* will sustain a range of counterfactuals.

Robinson quickly ranges far and wide explicitly leaving aside questions about the nature of the identity conditions on locations as per the birth of atoms, and claiming support from quantum theory which apparently views quantal entities as best treated as "something less than individuals" (p.208). A half proper treatment of counter factual identity would require a whole dissertation of its own and so I am pleased to leave it aside by reading that we can accept everything else in the treatise while objecting to this chapter (p.205). Lack of space for the topic will not constrain me reapplying my earlier judgement against what Robinson wants to claim from it. Even if the identity of things is dependent on our concepts, it does not follow that the stuff which makes up those things cannot be physical.

The next chapter on 'Dennett and the human perspective' will be the last of Robinson's arguments for conceptualism which altogether help license his substance dualism which will be the topic of the remaining sections.

Chapter 13. Dennett and the human perspective.

Robinson's final application for conceptualism is Dennett's instrumentalism and 'moderate realism'. Dennett is after a physically reductive account of *intentional* mental content (1987), and like the earlier chapter on Davidson I will only be concerned with the support Robinson wants to take for his position and with Dennett's theory only that extent. It will be helpful again to frame my resistance to Robinson's conceptualism as he poses it in this chapter:

> "…the ontologies of the physical world, other than the most basic one, must be understood in a conceptualist, as opposed to a realist, sense, and that this leads to a dualism within the physical realist framework, for conceptualism of this sort presupposes a human perspective on to the physical world from outside of it." (2016, p.210).

My resistance can be summarised by pointing to the inaccuracy of citing a *presupposition* of the human perspective. I see conceptualism as tautologous to the effect that a perspective is simply *required* in the world (rather than presupposed) to generate 'non-basic' concepts, and it is far from clear why that required perspective need be from outside the physical world.

Where Robinson describes conceptualism, "as the *irreducibility of the cartesian perspective*: it forces us to see the thinking subject as something different from, and in addition to, the realm of physical objects." (Ibid), I only see conceptualism equating to the requirement for a perspective *for* a cartesian perspective, and ineliminable perspective does not amount to irreducibility. Robinson claims that conceptualism parallels Dennett's physicalist argument for the *ineliminability* of the intentional perspective (for understanding mental content) in a way that provides a concession to conceptualism by making the Cartesian perspective ineliminable. But as per my previous thought, ineliminability will not support Robinson's notion regarding a 'presupposing' of the human perspective 'from outside', and so I claim no concession is earned. Robinson is trading on the causal equivalency he illegitimately conceives for psychology with the other special sciences. The problems besetting Dennett's theory arise from his target of intentionality as an aspect of consciousness,

but those problems do not apply to a Dennett style treatment of non-basic properties as Robinson needs for the concession he claims (p.212).

Dennett's theory may suffer from an internal regress due to his analysis of intentionality essentially involving an intentional perspective and his strategy to avoid this regress is to somehow connect intentional features with real patterns in the world. Robinson objects to this move because he claims those patterns are also mind dependent, and in effect they would not be those patterns if it were not for the perceiving mind, and so he seems to saddle science with the same regress due to the intentionality of minds that perform science: "the foundations of these judgments of similarity -- the movements of bodies in space -- are entirely real, but their reification as patterns -- as seen as unities of a certain sort -- involves the action of mind" (p. 217).

Robinson's has interesting arguments for the mind dependence of patterns in general, to the extent that Dennett's patterns might escape the regress. Robinson proposes a conflict in Dennett's claim that patterns are objectively there in their own right "irrespective of anything dependent on the subject" (p.216), because Dennett also relies on the notion that those patterns are only *discriminable* from a perspective. Robinson may be correct about this tension when the pattern is united through an exercise of intention exhibited in a behaviour, but he is not correct about patterns in general, "…where there is genuine physical similarity." (p.217).

Robinson takes the case of a collection of dots on a piece of paper in the shape of a circle from our perspective which, "…a visually more sophisticated mind might see as a polygon with the appropriate number of sides. Patterns are reified by the action of the mind. Otherwise – in the case of dots – there are just dots in certain physical places." (p.217). As in a previous chapter the suspicious move is to suggest the pattern is *reified* by the mind. I claim nothing can be reified by the action of a mind which has no effect on those materials which partially effect the reification (for arguments sake in conjunction with the mind). Everything which constitutes the pattern is there all along and all that is added to the situation is a mind required to perceive a pattern already existing. A consequence of Robinson's view is a potential quantity of distinct patterns without limit *just in that collection of dots,* owing to innumerable potential vantage points from

different perspectives of distance and size from which that collection may be viewed. This implies a case of potentially cascading reification of patterns only dependent on perspectives that could be taken while those dots remain *just what they are*. These patterns require nothing more than perspectives, and nothing in those patterns need be 'presupposed'. Perhaps this bizarre consequence will not embarrass Robinson who finds the ultimate explanation in substance dualism which he takes to be part motivated by this thesis of conceptualism.

I am reminded to make use of the old thought experiment about the falling tree in the forest and the whether it makes a sound if nobody experiences the event. Absolutely all conditions are present in the objective 'pattern' that would cause a sound if an ear was there to be affected by those conditions, but without an ear there is no pattern experienced and therefor no sound. A sound depends for its existence on being heard, and the hearer brings nothing to those conditions which cause that sound other than satisfying the requirement of *being there*. The only sense in which a pattern is 'reified' by an observer is in being *seen* by an observer, such as when one's attention shifts between the rabbit and duck in the famous experiment or when we see a face in the clouds.

This concludes part two of the dissertation. Physicalism is not undermined by conceptualism per se. Part three will consider Robinson's case for substance dualism.

Part Three. Mental Substance.

Ninety percent of this dissertation is complete, in which I have attempted some level of rigor treating the knowledge argument and then Robinson's conceptualism. I risk less rigour with the remaining word count in taking up Robinson's core argument for mind as immaterial substance, so making room a couple of more intriguing ideas which are tabled. This positive part of Robinson's treatise takes up only ten percent of the book and must be viewed more as a preliminary for the position because substance dualism involves numerous and substantial issues which are not mentioned (see Loose et al, 2018).

Most or all of Robinson's arguments for conceptualism he claims are consistent with property dualism, but now he will press for the mind as immaterial substance. The notion of substance is loaded going back to Aristotle, but Robinson proposes a simple model for now; instead of the brain hosting bundles of mental properties which only exist because of the brain, the mind as an immaterial substance can be thought to some degree causally dependent on the brain but not conceptually dependent; In the way a particular roundness property of a ball only exists with that ball it belongs to, the ball thought of as substance could exist without *that* particular roundness property.[13]

Robinson's argument for substance dualism is different from the approach instigated by Descartes, and is intended to avoid two main controversies concerning the theory of modality and the 'Cartesian theatre' which separates the inner mental life from the external world. Robinson's argument contests that, "…bundle dualism [which seemingly for argument sake he equates with property dualism] - the theory that the mind is not a substance but only a collection of immaterial properties or states – cannot accommodate certain essential features of personal identity – what makes a person the particular person that he or she is." (2016, p.235). This is the core argument that I will be quick about. More fascinating than his argumentation for mind

[13] This bare-bones model of 'substance' is all I will fit, and accepting Robinson's authority (2020) to the extent that his claims may be treated separately from potential background controversy about substance per se.

as a substance is the further contention that the mind is "a *simple* substance – one of the world's atomic entities, though in a rather special sense." (Ibid).

Robinson's core argument concerns identity in terms of counterfactual circumstances of origin. He takes his previous arguments regarding the origin of component parts of a ship to show that there may be no fact of the matter whether a certain ship would be the same ship if a proportion of it as different by different origin. He claims we may accept an overlap of constitution for the ship, and likewise for a human body; in these cases natural language or intuition cannot decide if the body is the same or not, however the same is not true in the case of identity of a mind, which chimes with ideas of Madell: that one's present consciousness or any state of present consciousness we might imagine either belongs to you or it does not, with no indeterminacy or question of degree (1981, p.91). Robinson thinks an overlap of mental constitution must happen in the form of overlapping psychic constitution, but the notion of numerically identical *psychic* parts overlapping cannot be applied in the way it applies to bodily parts (p.237). He explains that, unlike the notion of a person's body being twenty five percent different at origin, there seems no sense in the parallel question for mental identity: if we pose a twenty five percent change in mental life, which mental events constitute the identical seventy five percent? Or how might we imagine a person with a "ghostly" (p.237) seventy five percent of mental presence?

Other dualist theories argue by doubting the notion of mental identity over time instead of origin by questioning similarity of identity under different experiential history, using a notion Robinson calls 'empathetic distance'; in effect, how much connection might one imagine under a possible alternative life history? How would my identity be affected if I received a telephone call yesterday with good or bad news? – presumably not much (Lund 2014). How would my identity be affected if I chose to support the Blues when I was young? – maybe quite allot by suffering more than I did supporting the Villa. We can at least imagine alternative life histories using empathetic distance, in contrast to questions with counterfactual circumstances of origin. Robinson suggests there is no empirical traction for imagination to get going for alternative origin questions, because we have no basis to ground feasible thought experiments say for a person born from

different ovum but with exactly the same life experience as me: would that person be me? In the end, this demonstrates a strong intuition according to Robinson that fundamental physical facts in the world do not determine psychic identity.

The concept of *haecceitas* is evoked by Robinson, which might entice an advance from cognitive-blankness if it is more than mystery-mongering as many apparently hold it to be: "*Haecceitas* translates as 'thisness' and is, according to certain philosophers, the features of an object which, additional to its ordinary properties, makes an individual thing the particular that it is." (2016, p.239). Robinson wants to use this concept to help derive his view that there is always a matter of fact governing whether a purported counterfactual person is identical to oneself because individual minds exhibit haecceitas in a sense that:

> "…we all understand, namely our identity as subjects. It is because we intuitively understand this that we feel we can give a clear sense to the suggestion that it would, or would not, have been ourselves to which something had happened, if it had happened; and that we feel we can understand very radical counterfactuals - e.g. that I might have been an ancient Greek or even a non-human - whereas such radical counterfactuals when applied to mere bodies - e.g. that this wooden table might have been the other table in the corner or even a pyramid, makes no intuitive sense." (p.239)

Robinson claims that there is sense to be made of the question whether one could exist in another body, "…it seems to have content – in a way that a similar suggestion for mere bodies does not." When I attempt to imagine myself born in another country for example, I just don't sense how to *begin* pondering that, other than by attempting a shallow-looking-out-in-my-current-mood-but-in-a-different-place kind of a way, which only seems carry more content or intuitive sense than the attempt to think about the table because some incoherence threatens my imagining *that table I am looking at,* as the other table. The essential property instanced with my minimal cogito looking out on the world *can* be imaginatively separated from my body, but the parallel exercise does not go through for the table. At best this just brings us back to those Cartesian problems about imagination and conceivability which this approach was meant to avoid. I am left with the sense that there is something to Robinson's argument and his evocation of haecceitas, but it does not seem to generate any more-hefty-idea than was provided by Descartes.

One might approach this notion by imagination or thought experiments about subjective identity 'from the inside' so to speak. Doing so inclines me to question the order of explanation in play *prior* to the notional relevance of empathetic distance. I cannot get a handle on indeterminacy when I think about a subject either being me or not being me. Robinson does not rely on Cartesian arguments for dualism partly because they involve deep controversy about imagination and possibility, but he seems to dig an even deeper hole with this approach – I cannot decide if the intrigue I sense is because he is onto something or because there is nothing to think about? I experience a kind of cognitive blankness with a poised resistance to judge. Is that just a failure of my imagination or an inability to reason about such things, or both? I will leave that for the reader to decide, because there may be mileage to be gained with David Lewis that I can only touch on.

There may be a way into the question of personal haecceitas 'from the inside' with the framework of Lewis's possible worlds along with analysis of persistence conditions on identity (Ninan, 2009). Robinson blatantly assumes the falsity of Lewis's counterpart analysis and possible worlds metaphysics, presumably because of the realist understanding of possible worlds. Without taking up Ninan's analysis I will just note that there may be mileage in thinking about de se possibilities which relate to haecceitas of personal identity in a way purportedly compatible with physicalism. However, Ninan concedes that there may not be much to say to the likes of Robinson ("Nagel, Blackburn, Madell." (p.460)) by way of intuition that would change his mind.

Robinson claims that a projectivist view about possibilities under which judgements about counterfactuals are neither true or false but instead are expressive of attitudes is compatible with his view that physical objects and minds must be treated fundamentally different under counterfactual judgments, because those projections underdetermine real differences (p.243).

Robinsons core proposal that counterfactual identity facts in the case of minds enjoy a reality in a way that such claims about physical objects do not, is not decisive – beyond the intuition that only-me-as-I-exist-right-now-could-be-me. But for what it is worth, as far as that intuition goes, it is as clear as any proposition one could imagine, perhaps because it is maximally primitive and obviously true, as Descartes famously noticed.

I come now to the final section of the chapter and my dissertation, which is more intriguing than counterfactuals of origin stuff and may provide a finish with fascinating flourish. It is metaphysically speculative and concerns what might be said regarding what immaterial consciousness *is*, or how we might further characterise it beyond saying that it is not physical? The issues involve the purported property of consciousness being a *simple unity*.[14]

One problem for the idea that the conscious mind is a substance is accounting for the intermittency of consciousness during periods of deep sleep or general anaesthetic. Strawson (2009) proposes that the self may be composed of innumerable selves each consisting in a span of attention at different times. Robinson offers a more elegant if less naturalistic solution to this problem via consideration of a different problem confronting the theory of the self being a simple substance lacking parts: How can the conscious human subject be understood as a simple entity when that subject enjoys such an array of capacities and faculties with unlimited beliefs, hopes, memories, etc? Robinson approaches this problem "…by considering the 'unity in diversity' that is an essential feature of thought." (2016, p245).

In attending to how a thought is delivered by a sentence, one is aware of the content of that thought as a whole, while the expression of that content as a sentence takes place over time. Robinson develops this idea in support of an argument by Peter Geach, that the "activity of thinking cannot be assigned a position in the physical time-series." (1969, p.34). That one must be conscious of the whole thought to deliver a sentence demonstrates: "Something that has an essential unity finds expression in something that is complex." (p.245). The claim is that the experience of thinking is spread out 'alongside' the flow of empirical time over which the sentence is expressed, and in some sense the thought is prior to the sentence (not necessarily before) and with a unity unpossessed by a string of sounds constituting a sentence.

[14] The notion of consciousness being simple and unitary in the sense to be discussed is quite intuitive to me, while being difficult to pin down more analytically. Twelve years ago, I sat passively as an outsider at the back of a Birmingham seminar room while Galen Strawson gave a talk on this topic. Near the start, A scientifically oriented post-doc philosopher impatiently pressed Strawson for a definition of 'simplicity' as it was being used to characterise consciousness. He was encouraged to attend the unitary nature of his current feeling of self, or something like that. He was unimpressed with the suggestion and soon left the room.

Robinson concedes that this is a quite mysterious doctrine, but that it does accord with the phenomenology of thinking and to that extent I agree with him. I wonder if we can make progress by considering the phenomenology of hearing a sentence. There is also the experience of the received thought 'emerging' only once the sentence is complete (if say an unanticipated verb or noun comes at the end). Having never carefully attended to the phenomenology of thinking in relation to hearing, I just now switched on the radio to do just that. There is a sense in which we consolidate the words down to their essential gist and one is left with an idea without the memory of the component words. This could just indicate that we do not think with words rather than it purportedly showing some incommensurate time scale for thought. It is not obviously implausible to explain the 'wholistic' phenomenology of thought in relation to vocalised or heard words, just by way of different cognitive faculties operating in tandem, and we are not able to parse the time metric for thinking as we can for the vocalised sentence. This would seem to be amenable to empirical research under the assumption that we lack sensitivity required to parse our thinking in a clocked fashion. Robinson's idea that thinking is in some sense timeless and because of that happens outside physical events is unclear and can be resisted along the lines I suggest.

He draws support for the idea (that I here put in a more tractable form which does not force either a physicalist or dualist context and perhaps holds off my previous simple style of objection), that 'thinking does not meaningfully coincide with clock time', from the greater counter-intuitiveness that comes with ways we might try to avoid the idea. A prominent example of which is by Jerry Fodor who treated thinking as a computational process.

Fodor (1975; 1979) apparently treated "consciousness as irrelevant to thought, which is a computational process carried out in the purely formal 'Language of Thought' (LOT) in the brain" (2016, p.245). To cut a long story short which could otherwise range over a large literature in the Philosophy of Language, Robinson sides with Chomsky of fifty years ago whose theory is offered to demonstrate a consequence of LOT or indeed any natural language that must be driven purely by syntax, that so much of our vocabulary would need to be innate. Robinson takes from this that a grasp on meaning and understanding of coherent life projects *must*

depend on something more than "neural/computational machinery" (Ibid), which implicates a single self (substance) existing over time, and impossible to be grounded in physical stuff. Robinson steers this sketch back to his idea about thought not operating on physical time, by remarking that Strawson's ephemeral selves who occupy spans of time cannot account for the unity we feel pursuing our projects as single selves. In thinking about how the essentially conscious self must exist during periods of purported inactivity, he again draws on Geach to locate a potential solution which casts doubt on our common-sense understanding of temporal relations and time. Geach would diagnose my earlier resistance to entertain the timelessness of thought as driven by a:

> "(perhaps unacknowledged) assumption of a Newtonian or Kantian view of time: time is taken to be logically prior to events, events, on the other hand, must occupy divisible stretches or else indivisible instants of time. If we reject this view and think instead in terms of time-relations, then what I am suggesting is that thoughts have not got all the kinds of time-relations that physical events, and I think also sensory processes, have." (1969, p.35).

Robinson's suggestion is that physical time might be derived from, or less fundamental than, the relation amongst events that relate objects and occurrences which do not take place in what we think of as physical time.

Robinson (2007) develops this understanding of time to explain how the simplicity and unity of the self could seem to express itself as a diversity with reference to ideas about the atemporality of God whose "…relation to the physical world could be cognitive and volitional, without any temporal component." (2007, p.60). I offer a rejoinder that this notion of time cannot help Robinson's proposal for substance dualism like it elsewhere might support or cohere with his idealism or the atemporality of God. I claim his notion about time actually *requires* idealism for his argument to go through, or at least requires a non-realism about physical time, which he has ruled out with his contention that immaterial substance is casually related to the physical. That is enough to implicate realism about physical time (whether or not non-basic in relation to mental events); because it means that the interaction between mental and physical substance *does* coincide with clock time to the extent determined by the physical aspect of the causal event. Robinson cannot readily table this mysterious doctrine about time consistently with his substance dualism.

My argument that Robinson's ideas about time might better fit within an idealist theory could also be said about his 'conceptualism' of previous chapters. This prompts me to wonder if Robinson's overarching impetus is an attempt to motivate his preferred metaphysic of idealism, with naturalistically friendly-looking ideas that tentatively support the half-way house of substance dualism:

> "I shall argue in the final chapter that the individuality of mind is what makes possible the understanding of other things in individualised terms. This picture paves the way for idealism, though that will be a topic for another work." (p.205).

While Robinson's case for substance dualism is only suggestive, I am pleased to have chosen his clear and wide-ranging treatise for the education it has provided me about the mind-body problem, and I am much the wiser for it.

GENERAL CONCLUSION.

I am persuaded that the standard physicalist responses to the knowledge argument which I considered in part one are not adequate to answering Chalmer's or Robinson's take on the hard problem of consciousness. The notion of knowledge-by-acquaintance which I broached in Chapter five, recorded at least since Russell, seems perhaps the best approach to defend physicalism from the knowledge argument.

Neither Robinsons ideas about reduction in the sciences, or his related 'conceptualist' thesis about the mind dependence of certain entities, show with the clarity he claims that "the interpreter must transcend the physical world that he is interpreting." (2016, p.159). Those ideas therefore do not license or clearly motivate the notion of mental substance.

To the extent that Robinson's case for substance dualism in the final chapter trades on his positive arguments (aside from his criticism of alternative theories), it is very intriguing but only a suggestive preliminary for a fuller account which he plans to offer in the future. I am pleased to have chosen his clear and well scoped treatise as a program for my dissertation, which leaves me much wiser about the mind-body problem.

BIBLIOGRAPHY

Alter, 2007, "Does Representationalism Undermine the Knowledge Argument?" in T. Alter & S. Walter (eds.) 2007: 65–76.

Alter, T. & S. Walter (eds.), 2007, *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, Oxford: Oxford University Press.

Armstrong, D.M., 1961, *Perception and the Physical World*, London: Routledge.

Balog, K., 2012a, "Acquaintance and the Mind-Body Problem", in C. Hill & S. Gozzano (eds.), *New Perspectives on Type Identity: The Mental and the Physical*, Cambridge: Cambridge University Press, 16–42.

2012b, "In Defence of the Phenomenal Concept Strategy", *Philosophy & Phenomenological Research*, 84: 1–23.

2009, "The Oxford Handbook of Philosophy of Mind." Edited by Ansgar Beckermann, Brian P. McLaughlin, and Sven Walter. Oxford University Press.

Bayne, T., and Montague, M., (eds.), 2011, *Cognitive Phenomenology*. Oxford and New York: Oxford University Press.

Carruthers, P., 2000. *Phenomenal Consciousness: A Naturalistic Theory*. Cambridge University Press.

Carruthers & Violett, 2011. "The Case Against Cognitive Phenomenology". In Montague & Bayne 2011

Chalmers, D., 1995. "Facing up to the problem of consciousness". Journal of Consciousness Studies. 2 (3): 200-219.

1996, *The Conscious Mind: In Search of a Fundamental Theory*, Oxford: Oxford University Press.

2002, "The Content and Epistemology of Phenomenal Belief", in Smith, Q.& A. Jokic (eds.), *Consciousness: New Philosophical Essays*, Oxford: Oxford University Press, 220–272.

2006, "Phenomenal Concepts and the Explanatory Gap", in Alter, T. & S. Walter (eds.), 2007.

Churchland, P., 1985, 'Reduction, Qualia, and the Direct Introspection of Brain States.' *The Journal of Philosophy,* Vol. 82, No. 1, pp. 8-28.

Coleman, Sam (Editor), 2019, "The Knowledge Argument (Classic Philosophical Arguments)". Cambridge University Press.

Conee, E. 1994. 'Phenomenal Knowledge.' *Australian Journal of Philosophy 72: 136-50.*

Crane, Tim., 2019, "The Knowledge Argument is an Argument about Knowledge", in Coleman 2019.

Davidson, Donald., 1993, "Thinking Causes", in Heil and Mele, 1993.

1980, *Essays on Actions and Events*, Oxford: Clarendon Press.

1970, "Mental Events", in Davidson 1980.

Dennett, D., 2007, "What RoboMary Knows", in T. Alter & S. Walter (eds.) 2007: 15–31.

1991, *Consciousness Explained*, Boston: Little, Brown, & Co.

1987, *The Intentional Stance,* Cambridge, MA: MIT Press.

Fodor, J.A., 1979. *Representations: Philosophical Essays on the Foundations of Cognitive Science.* MIT Press (US).

1975, *The Language of Thought.* Cambridge, MA: Harvard University Press.

1974, "Special Sciences: Or the Disunity of Science as a Working Hypothesis", *Synthese*, 28: 97–115.

Geach, Peter, 1969., 'What Do We Think With?' in his *God and the Soul.* Cambridge.

Grandin, Temple. 2006. *Thinking in Pictures: My Life with Autism,* 2nd edn. New York: Vintage Books.

Gillett and Loewer (Ed) 2008. *Physicalism and its discontents.* Cambridge University Press.

Goff, Phillip., 2017, "*Consciousness and Fundamental Reality".* OUP.

2019, "*Galileo's Error: Foundations for a New Science of Consciousness*" :Rider.

Heil, J. and Mele, A.(eds.), 1993, *Mental Causation*, Oxford: Clarendon Press.

Honderich, T., 1982, "The Argument for Anomalous Monism", *Analysis*, 42: 59–64.

Horgan, Terence., 1984. 'Jackson on Physical Information and Qualia.' *Philosophical Quarterly 34: 147-52.*

Peter van Inwagen & Dean Zimmerman (Ed) 2007. *Persons, Human and Divine.* Oxford.

Jackson, Frank. 2007, "The Knowledge Argument, Diaphanousness, Representationalism", in T. Alter & S. Walter (eds.), 2007: 52–64.

2003. "Mind and Illusion" . *Royal Institute of Philosophy Supplement, 53*, 251-271. doi:10.1017/S1358246100008365

1995, "Postscript on 'What Mary Didn't Know'", in P. Moser & J. Trout (eds.), *Contemporary Materialism*, London: Routledge, 184–189.

1982, "Epiphenomenal Qualia", *Philosophical Quarterly*, 32: 127–136.

Kim, Jaegwon. 1989., "The Myth of Nonreductive Materialism", *Proceedings and Addresses of the American Philosophical Association*, 63: 31–47.

2007., *Physicalism or Something Near Enough*. Princeton University Press.

2012., Against Laws in the Special Sciences. *Journal of Philosophical Research* 37.

Kind, Amy. 2019, "Mary's Powers of Imagination", In Coleman et al, 2019.

Kripke, Saul. 1980, *Naming and Necessity*, Cambridge, MA: Harvard University Press.

Lavazza & Robinson (Editors), 2014., *Contemporary Dualism – A Defense.* Routledge.

Levin, Janet., 2007, "What is a phenomenal concept." in T. Alter & S. Walter (eds.) 2007.

Levine, Joseph., 2018. *Quality and Content: Essays on Consciousness, Representation and Modality.* OUP.

2007, "Phenomenal Concepts and the Materialist Constraint." in T. Alter & S. Walter (eds.) 2007.

Lewis, David. 1983. "Mad pain and Martian pain", in *Philosophical Papers*, Vol. I. Oxford University Press.

1988, "What Experience Teaches", *Papers in Metaphysics and Epistemology* (Cambridge: Cambridge University Press, 1999), pp. 262–90. Also included in Ludlow et al, 2004.

Loar, Brian. 2003. "Qualia, Properties, Modality." Philosophical Issues, 13, Philosophy of Mind, 2003

1997. "Phenomenal States" (revised version), in The Nature of Consciousness, ed. N. Block, 0. Flanagan, and G. Guzeldere. MIT Press.

Loose, Menuge, Moreland (ed's), 2018., *The Blackwell Companion to Substance Dualism.*

Lund, David, 2014., "Materialism, Dualism, and the Conscious Self." In Lavazza & Robinson.

Ludlow, Nagasawa and Stoljar (eds.), 2004. *There's Something About Mary*, (Cambridge, MIT Press).

Madell, G. 1981., *The Identity of the Self.* Edinburgh University Press.

McGinn, Colin., 1989 “Can We Solve the Mind--Body Problem?” *Mind*, vol. 98, no. 391.

McLaughlin, Brian and Karen Bennett, 2018., "Supervenience", *The Stanford Encyclopedia of Philosophy* . <https://plato.stanford.edu/archives/win2018/entries/supervenience/>

McLaughlin, Beckerman, Walter (Eds)., 2011. *Oxford Handbook of Philosophy of Mind.* OUP.

Thomas Metzinger (Ed) 1999. *Conscious Experience*. Imprint Academic Press

Moore, G. E. 1922. “The refutation of idealism.” In G. E. Moore *Philosophical Studies*. London : Routledge.

Montague & Bayne (eds.) 2011. *Cognitive Phenomenology.* Oxford University Press.

Nagel, Ernest., 1961, *The Structure of Science. Problems in the Logic of Explanation*, New York: Harcourt, Brace & World, Inc.

Nagel, Thomas., 1974. "What Is It Like to Be a Bat?". The Philosophical Review. 83 (4): 435–450

Nemirow, Laurence, 1980. review of Nagel’s *Mortal Questions*, *Philosophical Review* 89

2007, “So This Is What It's Like. A Defense of the Ability Hypothesis”. In Alter and Waler.

Ninan, Dilip, 2009, “Persistence and the First-Person Perspective”, Philosophical Review, 118: 425–464.

Papineau, D., 2002, *Thinking about Consciousness*, Oxford: Oxford University Press.

2007, “Phenomenal and Perceptual Concepts,” in T. Alter & S. Walter (eds.) 2007.

Paul, L.A. 2017. *"DE SE* Preferences and empathy for future selves." *Philosophical Perspectives, 31, Philosophy of Mind, 2017 doi: 10.1111/phpe.12090*

Pettit, Philip., 2004. "Motion Blindness and The Knowledge Argument". In Ludlow et al.

Pitt, D. 2019. "Aquaintance and Phenomenal Concepts". In Coleman (Ed) 2019.

2011., "Introspection, Phenomenality, and the availability of intentional content." In Montague & Bayne (eds) *2011.*

Raffman, Dianne., 1995. "On the Persistance of Phenomenology," in Thomas Metzinger (ed.), *Conscious Experience.* Schoningh Verlag.

Robb, David and Heil, John, 2019., "Mental Causation", *The Stanford Encyclopedia of Philosophy* (Summer 2019 Edition) <https://plato.stanford.edu/archives/sum2019/entries/mental-causation/>.

Robinson, Howard., 2020, "Substance", *The Stanford Encyclopedia of Philosophy.* <https://plato.stanford.edu/archives/spr2020/entries/substance/>.

2016, *From the Knowledge Argument to Mental Substance: Resurrecting the Mind.* Cambridge University Press.

2008, "Davidson and Nonreductive Materialism: A Tale of Two Cultures." In Gillett and Loewer 2008

2007, "The Self and Time". In Peter van Inwagen & Dean Zimmerman (Ed) 2007.

1993, "Dennett on the Knowledge Argument." *Analysis*, vol. 53, no. 3, 1993, pp. 174–177. *JSTOR*, www.jstor.org/stable/3328467.

Robinson, William, 2011. "A Frugal View of Cognitive Phenomenality". In Montague & Bayne(ed) 2011

Russell, Bertrand., 1927. *The Analysis of Matter.* London: Allen and Unwin.

1940. *An Inquiry into Meaning and Truth.* London: George Allen & Unwin.

Ryle, Gilbert., 1949, *The Concept of Mind. Hutchinson & Co.*

Smart, J.J.C., 1959, "Sensations and Brain Processes", *Philosophical Review*, 68: 141–156.

Strawson, Galen., 2018, "The-Mary-Go-Round." In Coleman 2019.

2006, *Consciousness and Its Place in Nature.* Thoverton: Imprint Academic.

Stanley, Jason and Williamson, Timothy. 2001. “Knowing How.” The Journal of Philosophy, 98: 411-44.

Stoljar, Daniel, 2005. "Physicalism and Phenomenal Concepts". *Mind & Language*. 20 (5): 469–494.

2006, *Ignorance and Imagination: The Epistemic Origin of the Problem of Consciousness*, New York: Oxford University Press.

Subdstrum, 2011. “Phenomenal Concepts”. Philosophy Compass 6/4 (2011): 267–281.

Spener, Maja, 2010. “Disagreement About Cognitive Phenomenality”. In Montague & Bayne 2011.

Tye, Michael., 2000. “Knowing what it is like: The ability hypothesis and the knowledge argument”. In Ludlow et al, 2004.

Daniel Stoljar and Yujin Nagasawa., 2004, “Introduction”, in Ludlow et al, 2004.

Yalowitz, Steven, 2019., "Anomalous Monism", *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2019/entries/anomalous-monism/>.