Mirroring and Indeterminacy:

Towards Indeterminate Mind-Brain Identity

Zhuo-Ran Deng

November 2016

A dissertation submitted in fulfillment of the requirement for the degree of Doctor of Philosophy in Philosophy

Department of Philosophy

University of Canterbury

Abstract

In this dissertation I offer my objections to three famous arguments concerning the mind-body problem.

The first argument is Saul Kripke's (1980) modal argument against psychophysical identity theory. Kripke argues that if pain is identical to C-fibre firing then this identity must be necessary. However he points out that the identity is, if true, also *a posteriori*, and he argues that this alleged a posteriori identity cannot be accounted for in the way that scientific a posteriori identities are accounted for. He concludes on this basis that pain cannot be identical to C-fibre firing, and, more generally, that alleged psychophysical identities are false.

The second argument is David Chalmers' (1996) 'zombie' argument against materialism. Chalmers argues that zombies are conceivable, that the conceivability of zombies entails the possibility of zombies, and that the possibility of zombies is inconsistent with the truth of materialism. He concludes that materialism is false.

I show that these arguments both share the same logical form—a form distinctive of what I call a 'conceivability argument'. I show that for any such conceivability argument, *C*, there is a corresponding 'mirror argument' that is deductively valid and has a conclusion contradicting *C*'s conclusion. I show that a proponent of *C* can challenge the premises of the mirror argument only at the cost of undermining *C*'s premises. I conclude on this basis that conceivability arguments are fallacious in general, and, more particularly, that both Kripke's modal argument and Chalmers' zombie argument are unsound. This critique of these two arguments constitutes the first part of the dissertation.

The second part is devoted to Hillary Putnam's (1967) multiple realisability argument against identity theory. Putnam argues that if human pain is a neural firing pattern in the brain, then octopus pain will likewise be identical to some physical state of the octopus—say, an excitation pattern in the jelly-ish tissue of the octopus brain. But while human pain and octopus pain feel alike, neural firing and jelly excitation are not alike. It follows from standard logic that human pain is not identical to neural firing patterns.

In reply, I attempt to reconcile identity theory with multiple realisability by advocating a semantics in which identity statements involving vague terms such as 'pain' are *indeterminate*. I develop a non-classical axiomatic theory of indeterminate identity relations, which implies that indeterminate identities are non-transitive. I also show that the principle of the transitivity of identity is a vital inference rule in Putnam's argument. If my analysis is correct then Putnam's argument is invalid.

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Deputy Vice-Chancellor's Office
Postgraduate Office

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

'Appendix' (page 77-97) is the following article:

Campbell, D., Copeland, B. J., & Deng, Z. (2017), 'The Inconceivable Popularity of Conceivability Arguments' in Philosophical Quarterly, 67 (267), 223-240.

Please detail the nature and extent (%) of contribution by the candidate:

33%

Certification by Co-authors:
If there is more than one co-author then a single co-author can sign on behalf of all
The undersigned certifys that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: *Doug Campbell* Signature: DCampbell Date: *21-11-2016*

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Deputy Vice-Chancellor's Office
Postgraduate Office

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

Discussions of conceivability arguments' 'self-sacrificial' problem (pp. 36-38 and p. 60) is extracted form the following talk:

Copeland, B.J., Campbell, D.I. & Deng, Z. (2017), 'Mirroring, Zombies, and Non-Reductive Consciousness', *Digitalization for a Sustainable Society*: The 2017 Summit of the International Society for Information Studies, Chalmers University of technology, Gothenburg, Sweden (presented by Copeland).

Please detail the nature and extent (%) of contribution by the candidate:

33%

Certification by Co-authors:

If there is more than one co-author then a single co-author can sign on behalf of all

The undersigned certifys that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: Jack Copeland Signature: *B J Copeland* Date: *26-09-2017*

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Deputy Vice-Chancellor's Office
Postgraduate Office

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

The indeterminacy rebuttal of Putnam's octopus argument (pp. 120-138, 142-147, 150-158) is elaborated upon the following paper:

Copeland, B.J., Deng, Z. & Campbell, D.I. (forthcoming), 'Fuzzy Identity Theory'.

Please detail the nature and extent (%) of contribution by the candidate:

33%

Certification by Co-authors:
If there is more than one co-author then a single co-author can sign on behalf of all
The undersigned certifys that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: Jack Copeland Signature: *B J Copeland* Date: *26-09-2017*

Acknowledgements

I am greatly in debt to my two supervisors, Distinguished Professor B. Jack Copeland, and Dr. Douglas Campbell. Throughout my candidacy, they have been kind, encouraging, and most importantly, full of useful philosophical insights. Portions of this dissertation have been turned into a published co-authored article by the three of us. I would like to express my sincere gratitude to Jack and Doug for sharing their invaluable wisdom with me. I wouldn't have worked out the two main themes of my thesis without them. Doug's knowledge of anti-physicalist arguments has helped me enormously in researching the mirroring strategy. Jack's own writings on vagueness and indeterminacy were a deciding factor for me in coming up with the idea of indeterminate mind-brain identity. Thus, indeterminately, this work is the brainchild of all three of us; determinately, they have made this work far better than it could have been.

I have also been helped by other philosophers. I would like to thank John Bigelow, David Chalmers, Frank Jackson, Cathy Legg, Alex Miller, Harold Noonan, Diane Proudfoot, and Michael-John Turp, who have all offered very useful comments and constructive criticisms.

I would also like to thank the following for their bibliographical suggestions: Dugald Murdoch, Charles Pigden, Mark Steiner, and Robert Stoothoff.

Special thanks to Zoe Reeves for her editing and proofreading work.

Last, but not least, thanks to my family and friends for their understanding, patience and support.

Contents

Part I. Mirroring and conceivability arguments

Part II. Indeterminacy and the multiple realisability argument

Chapter 1. Introduction

1. 0. Arguments against identity theory through Lewis' eyes

Identity theory was the brainchild of U. T. Place (1956),[1] Herbert Feigl (1958),[2] and J. J. C. Smart (1959).[3] An important and very influential version of identity theory was subsequently developed by David Lewis (1966).[4] It was Lewis' version of identity theory that attracted the attention of two philosophers, one of who, like Lewis, had spent their career at Princeton, namely, Saul Kripke, and the other one is Hillary Putnam. Kripke and Putnam were the authors of two of the major objections to identity theory that will be critiqued in this thesis. Moreover, Kripke's argument against identity theory was the forerunner of David Chalmers' 'zombie' argument against materialism,[5] which is the third major argument that I will critique. Because Lewis' version of the theory is the version that prompted the arguments I will critique, it is appropriate to start with Lewis.

In his 1966 article 'An Argument for the Identity Theory', Lewis outlines for the first time his position on the mind-body problem. Lewis labels himself a psychophysical identity theorist, since he argues for the view that every experience is identical with some neural state.[6] *A fortiori*, he holds the materialist view that every mental state is a physical state.

[1] Place, U. T. (1956). 'Is Consciousness a Brain Process?' in *British Journal of Psychology*, Vol. 47(1), pp. 44-50.

[2] Feigl, H. (1958). 'The "Mental" and the "Physical"' in *Minnesota Studies in the Philosophy of Science*, Vol. 2, pp. 370-497.

[3] Smart, J. J. C. (1959). 'Sensations and brain Processes' in *Philosophical Review*, Vol. 68. Pp. 141-56.

[4] Lewis, D. (1966) 'An Argument for the Identity theory', in *Journal of Philosophy*, Vol. 63(1), pp. 17-25. Reprinted in Lewis, D. (1983) *Philosophical Papers, Volume 1,* pp. 99-107.

[5] As many philosophers do, I will use 'materialism' and 'physicalism' interchangeably in this dissertation.

[6] Lewis's usage of the word 'experience' is inconsistent. In 'An Argument for the Identity Theory', 'experience' is to be taken as experience-universal, not experience-particular. The meaning of the term is capable of being instantiated by different agents at different times. The term herein denotes metal event type. Contrarily, in his 1995 'Should a Materialist Believe in

Lewis begins by refuting a classic criticism. This type of counter-argument against identity theory appeals to Leibniz's Law and can be formulated in various ways. In essence, the reasoning goes: if identity theory is true, then mental states and physical states should have no differentiating properties between them. What foes of identity theory have to do is to find a differentiating property which is possessed by mental states, yet lacked by physical states. It seems that physical location is one good example of such property.[7] Neural states have physical location, to be precise, they are inside an agent's skull. But mental states are abstract. Therefore, they are analytically unlocated. In reply, Lewis puts a question mark on the inference from the abstractness of X to the unlocatedness of X and calls it a 'metaphysical prejudice'. The transition between these two properties is, at best, subject to further metaphysical consideration, and thus enjoys no analytic necessity. Moreover, as Lewis claims, neural states are abstract too. So it turns out that abstractness cannot discriminate between the mental and the neural.

Another example of the differentiating properties that the opposition has in mind is the Fregean *senses* of mental state-ascriptions and neural state-ascriptions. It seems obvious that 'I am having a headache' and 'I am in such–and–such neural state' do not share the same intensional meaning. From this the critics conclude mental states are not identical with neural states. For this objection, Lewis concedes that the opposition is right about the differentiating property – that mental-ascriptions and neural state-ascriptions do not share the same sense. But this kind of discrimination inflicts no harm on the central tenet of identity theory, since 'we can explain those discrepancies without denying psychophysical identity and without admitting that it is somehow identity of a defective sort'.[8] Here, Lewis hints at an important point about identity theory in general, that the sense and reference distinction is a useful guide for understanding the theory. The critic often claims mental states and neural states are two kinds of thing due to their different meanings. But instead of this flawed understanding, we shall say, according to identity theory, there is only one thing, the neural state, which can be known under two modes of presentation. The sense is the differentiating property, and no identity is being claimed here by identity theory. The reference, on the other hand, always denotes one and the same thing, and therefore, is where the identity holds. One more point must be added

Qualia', 'experience' is to be taken as experience-particular. The term therein denotes mental event token.

[7] Lewis cites Shaffer (1961) as an example of this argument.

[8] Lewis ([1966] 1983), p. 101.

here. The fact that Lewis appeals to Frege's distinction of sense and reference may lead to an incorrect impression. For if the identity between mental and neural state is just like that between water and H_2O, one will inevitably treat the mental-neural identity statements as *a posteriori* truths.

By showing that this objection is unsound, Lewis also clarifies what identity theory really asserts. Then Lewis turns to present his reasons for believing the view that the mental is identical to the neural, and a fortiori the mental is physical:

> My argument is this: The definitive characteristic of any (sort of) experience as such is its causal role, its syndrome of most typical causes and effects. But we materialists believe that these causal roles which belong by analytic necessity to experiences belong in fact to certain physical states. Since those physical states possess the definitive characteristics of experience, they must be the experiences.[9]

The argument has a simple form with only two premises:

LM1. Analytically, the *"definitive character"* of any mental state is its causal role.

LM2. Physical causes are adequate to bring about physical phenomena.

Therefore,

LM3. Analytically, mental states are identical with physical states.

One feature of this argument immediately arises. Lewis's materialistic conclusion is not derived from a *substantive* claim, that mental states are such-and-such (and such-and-such are physical). Rather, the first premise makes a *conceptual* claim about mental states. It stipulates how we shall describe mental states, not what sort of things they are. Lewis is upfront about the fact that his reasoning for LM1 is derived from behaviorism, a theory his theory aims to replace. Like the behaviorists, Lewis also believes we pick out a mental state by its typical bodily cause and effect. Nonetheless, his theory improves on the behaviorists' approach in three aspects:

1. Mental states are just as real as their causes and effects.
2. Mental states can be interdefined.
3. Untypical causes and effects are allowed.

[9] Lewis ([1966] 1983), p. 100.

Lewis offers no evidence in support of LM1 other than appealing to the behaviorist account of mind-body causation. What's new and interesting is his claim that mental states, by analytic necessity, are causal roles. That is, we inevitably pick out a mental state M by its causal role if we understand the meaning of the term 'M'. The analytic status of Lewis's mental-physical identification is explained in full detail in his 1972 paper 'Psychophysical and Theoretical Identifications'[10] which I will discuss shortly.

Lewis notes that LM1 alone cannot establish the materialist doctrine, since it is possible for a causal role to be filled by non-physical phenomena. LM2 aims to rule out this possibility. We pick out mental states by their causal role in the bringing about of physical phenomena, and since we do not need causes other than physical causes to bring about these phenomena, it follows that mental states are physical. LM2, according to Lewis, is 'a traditional and definitive working hypothesis of natural science' and it assures that physics is, in his words,

> […] a true and exhaustive account of all physical phenomenon (i.e. all phenomena describable in physical terms) … the theory governing phenomena out of which that phenomenon is composed and by the way it is composed out of them. The same is true of the latter phenomena, and so on down to fundamental particles or fields governed by a few simple laws, more or less as conceived of in present-day theoretical physics.[11]

Unfortunately like LM1, Lewis does not offer a detailed explanation of why we should believe this. Due to this lack, what he really means by 'explanatory adequacy' is unclear to me. I suggest two interpretations. Lewis has in mind either just LM2, or *the causal closure of physics*. The latter is a doctrine that includes LM2, plus an extra component:

LM2. Physical causes are adequate to bring about physical phenomena.

+

CCP. Nothing else can bring about physical phenomena.

It is not hard to see that the causal closure of physics asserts a stronger doctrine of causation than the explanatory adequacy of physics. Hence, it is less defendable. It is possible for one to hold the explanatory adequacy and deny the causal closure, but not vice versa. So the question of whether or not Lewis's second premise claims the stronger

[10] Lewis, D. (1972). 'Psychophysical and Theoretical Identifications' in *Australasian Journal of Philosophy*, Vol. 50, pp. 249-58. Reprinted in Lewis, D. (1999). *Papers in Metaphysics and Epistemology,* pp. 248-61.

[11] Lewis ([1966] 1983), p. 105.

doctrine is a fruitful one. The answer to this question would dictate the truth of premise two and therefore the soundness of the argument, given the true theory of causation only preserves the explanatory adequacy of physics. On the other hand, if the stronger doctrine turns out to be the correct analysis of causation, then it would not matter which doctrine Lewis really means by saying 'explanatory adequacy'. The second premise could survive even if he in fact claims the weaker one while the stronger one turns out to be the real deal, because, again, causal closure of physics includes the explanatory adequacy of physics. It follows that it would be a safer bet to interpret Lewis' 'explanatory adequacy' as just explanatory adequacy, since he doesn't have to claim more. To do the opposite (i.e. to treat it as claiming the causal closure of physics) is to invite further considerations, as the truth or falsity of his second premise will be relying on the outcome of an entirely different metaphysical debate, namely, the debate on the nature of causation.

Either way, Lewis makes it clear that the second premise is not a denial of non-physical phenomena at all. Consequently, he accepts the possibility that non-physical phenomena may coexist with mental states. The mere coexistence does not ipso facto influence the physical nature of mental states. It is just that nonphysical phenomena, if there are any, do not causally interact with physical phenomena. Next, by LM1, do not causally interact with mental states. From here, Lewis' physicalist conclusion is deduced.

1. 1. Accommodating multiple realisability

Lewis defends a 'type-type' identity theory. Since the 'type-type' theory holds that mental states type M are identical with neural states type P for anyone at anytime, seemingly, it follows that two agents would be different in neural attributes only if they are different in mental attributes. Thus, neural discernibility entails mental discernibility. By *modus tollens*, mental indiscernibility entails neural indiscernibility.[12] To make it more transparent, suppose there are two agents Fred and Greg. Call the mental states of this duo M_f and M_g respectively; likewise, call their corresponding neural states P_f and P_g.

[12] Another way of putting this idea is to say that the mental supervenes on the neural or the physical. So it looks like we are presented here with a supervenience component of Lewis' theory of mind. However Lewis' is vitally different from the standard supervenience theory, namely that he thinks supervenience is a kind of reduction and thus the mental not only just supervenes on but also is ontologically reducible to the physical. Lewis' view on supervenience in given in his (1994) 'Reduction of Mind' in S. Guttenplan (ed.), *Companion to the Philosophy of Mind*, pp. 412-31.

It should be fairly visible that according to the type identity theory, $M_f = P_f$, $M_g = P_g$, and it also implies the following:

MR1. $(P_f \neq P_g) \rightarrow (M_f \neq M_g)$ – If Fred and Greg have different physical states, then they have different mental states.

MR2. $(M_f = M_g) \rightarrow (P_f = P_g)$ – If Fred and Greg have the same mental states, then they have the same physical states.

Due largely to this later implication, the 'type-type' theory in general is prone to a powerful challenge known as the multiple realisability objection. This attack, proposed by Putnam, will be examined in detail in Part II of my dissertation.

This objection emphasises a common belief held by most people: different things that enjoy little physical similarity can share the same type of mental state. Combined with MR2 it yields a *reductio* argument as follows:

1. If "type-type" identity is correct, then $(M_f = M_g) \rightarrow (P_f = P_g)$.
2. Mental states are multiply realisable, so it is possible for $(M_f = M_g)$ & $(P_f \neq P_g)$.

Therefore,

3. The "type-type" theory is incorrect.

Suppose Fred is a perfectly normal human being, and Greg is an octopus, if the type identity theory is correct, then $(M_f = M_g) \rightarrow (P_f = P_g)$. In other words, if Fred and Greg have the same mental states, then they have the same physical states. Now, suppose both Fred and Greg are capable of having mental states, say being in pain. Then by $(M_f = M_g) \rightarrow (P_f = P_g)$, we should get $P_f = P_{g,}$ But isn't it empirically evident that $P_f \neq P_g$? P_f is purportedly C-fibre firing. How can Greg the octopus have this physical state?

Thus, any credible theories that aim to identify mental attributes (or properties) with physical attributes (or properties) need to bridge this gap between a commonsense (i.e. multiple realisability) and a logical implication of the theory (i.e. MR2). Lewis' 1980 article 'Mad Pain and Martian Pain' is one of his attempts to solve this riddle.[13]

[13] Lewis, D. (1980a) 'Mad Pain and Martian Pain', in N. Block (ed.) *Readings in the Philosophy of Psychology*, pp. 216-22. Reprinted in Lewis (1999), pp. 122-30. A more direct Lewisian reply to Putnam is offered in Lewis' 'Review of Putnam', which I will examine in Chapter 4.

A ‘Martian’, according to Lewis, is a type of hypothetical creature that has an entirely different physical realisation. Their way of generating mental states has no resemblances to our neural system. Nonetheless, if Lewis’ causal analysis of pain is correct, a Martian is deemed to be in pain when they engage in typical pain behaviors like ours. It implies that something could be in pain without any neural state like ours. Meanwhile, MR2 says $(M_f = M_g) \rightarrow (P_f = P_g)$, that the two agents must have the same physical states if they share the same mental states. An identity theorist would have trouble with the presence of Martian pain if her theory identifies the property ‘P’ with ‘neural states’. Because she would be committed to the view that if pain is such-and-such neural state, then an agent without that such-and-such neural state is not in pain. Lewis departs himself from this kind of identity theory. According to Lewis psychophysical identity is essentially psycho-causal identity since the identity holds between pain (or any other mental states) and whatever occupies the causal role of pain (or any other mental states). In short, Lewis’s theory allows that mental states can be multiply realised, because causal roles can be multiply realised, and therefore, is immune to Putnam’s attack.

Despite causal roles’ multiple realisability—the idea that many different sorts of states can occupy them—Lewis believes that they can only be occupied by physical ones. Unfortunately, he offers much less insight on the materialist part of his theory than for the causal/functional part of his theory. There are no words about why the causal roles must be played by physical states in ‘Mad Pain and Martian Pain’. Regarding this issue, I think Lewis would point his finger at the explanatory adequacy of physics as the underlying premise for his claim that causal roles must be physical, as hinted at the abovementioned LM2.

Another challenge that Lewis deals with in this paper is that he addresses a problem involving atypical pain behaviors. This in turn shows that Lewisian materialism is superior to behaviorism.

A ‘Madman’ is someone who does not behave the way most people normally would given a specific set of mental states, say pain. For the Madman to be in pain, he may perform in some very strange ways such as ‘crossing legs’ and ‘snapping fingers’, and he would show no sign of discomfort like our usual pain behaviour such as groaning and screaming. Yet, for all these atypical pain behaviours the Madman may still have pain. ‘In short, he feels pain but his pain does not at all occupy the typical causal role of pain', as Lewis summarised.[14]

It is not hard to see why Mad pain troubles people who subscribe to behaviourism. A behavirorist's account of what mental states are is defined entirely in terms of their roles in behaviour. So it is very difficult for the behaviourist to agree that the Madman is in pain, due to his lack of relevant behaviours. One easy way out of this is to claim that mental states are disjunctive concepts. There is one theory of pain for the common men; regarding the Madman, there is another. Then the concept of pain becomes: to be in pain is to be either something or another. Lewis deems the disjunctive approach to be desperate. To make room for the Madman while maintaining that the concept is a single unified one, Lewis recalls a point he discussed earlier in 'An Argument for the Identity Theory', namely, that although we pick out a mental state by its typical bodily cause and effect, untypical causes and effects are allowed. What pain (or any other mental state) is, according to Lewis' account, is built into the folk[15] psychological understandings of pain (or any other mental state). Since people (i.e. the 'folks') rarely think that pain must cause behaviours like groaning and screaming *at all times*, the folk understanding of pain allows exceptions. On this account, the Madman's pain is deemed to be a state that 'comes near to realising commonsense psychology'. Pain, according to Lewis, turns out to be what *mostly* occupies the causal role of pain 'in an appropriate population'. Lewis then concludes the matter by outlining a schema that he claims would define 'appropriate population':

> Perhaps (1) it should be *us*; after all, it's our concept and our word. On the other hand, if its *X* we are talking about, perhaps (2) it should be a population that *X* himself belongs to, and (3) it should preferably be one in which *X* is not exceptional. Either way, (4) an appropriate population should be a natural kind – a species, perhaps.[16]

If two or more of the four criteria were filled in by X*,* then X is said to be in pain for X*'*s population. In the case of the Madman, he has a state we have, and he belongs to the group of human beings – a specie, criterion (1), (2), and (4) are granted. Thus, Mad pain is explained.

[14] Lewis (1980a), p. 123.

[15] Lewis uses 'commonsense' instead of 'folk'.

[16] Lewis (1980a), p. 126.

1. 2. Analyticity

I shall explain a unique feature of Lewisian physicalism as a whole, namely, the analytic status of his theory.

Lewis establishes a physicalist view that a mental state is defined by its causal roles. This view can be taken by two ways. One might think that one day when technology is sufficiently advanced we will find out through experiments that (i) mental states are causal roles, and (ii) the explanatory adequacy of physics is true. On this way of thinking, the Lewisian remark is just an empirical fact.[17] On the other hand, one could take it as a conceptual truth. To be in pain, is by analyticity, to have such-and-such causes and effects. Lewis makes the latter, stronger claim that folk-psychology defines what mental states are. His 'Psychological and Theoretical Identifications' explains why we know this analytically.

Lewis' persuasion is a simple *modus ponens*:

1. Mental-physical identifications are like theoretical identifications.
2. Theoretical identifications are implied by the theories that make them possible, and not posited independently.

Therefore,

3. Mental-physical identifications are implied by mental and physical theories, not posited otherwise.

The first premise needs little explanation. The large chunk of Lewis' paper is devoted to explain premise 2, which can be rephrased in other words: the terms in a theory are implicitly defined by the content of that theory. Lewis takes two steps in explaining this. First, he gives an example in which the meanings of the names in the story only unfold by the story itself. To elucidate the idea, Lewis introduces new terminology: T-terms are the terms a theory is going to implicitly imply; O-terms are terms we have already understood before the theory. According to Lewis, T-terms can be uniquely defined by sentences

[17] This view can be labeled as 'empirical functionalism', or as Ned Block calls it 'psychofunctionalism'. A thorough discussion on the difference between Lewisian/analytic functionalism and empirical functionalism is offered in Block, N. (1980a) 'What is Functionalism' and (1980b) 'Troubles with Functionalism' in his (ed.) *Readings in the Philosophy of Psychology, Vol. 1.*

involving only O-terms. Hence, T-terms can be identified with O-terms-descriptions. Then Lewis takes the second step, which essentially, is a technical elaboration of the first step. The technique Lewis employs here is the use of Ramsey sentences.[18] From these Lewis concludes, the meanings of mental state-terms are derived from common knowledge associated with these terms (i.e. folk-psychology). The T-terms (mental state-terms) denote nothing unless the statement of the relevant Ramsey sentence involving O-terms (folk-psychological analysis of mental states) are true. Since the O-terms in this case (folk-psychology) describe causal relations, the T-terms (mental state-terms) are causal roles. In short, if we ever know what mental state-terms mean, we must know that something occupies the relevant causal role, hence the analyticity.

One point that needs to be noted immediately is that the analyticity of identification is conditional. Lewis does not argue for the analytic truth of any theoretical identification given by a theory. Instead, it is a statement of the form that *if* there are in fact any *referents* for the T-terms, then they are identical to the things that satisfy the relevant Ramsey sentence that is framed using the O-terms. The corresponding theory supplies us with what *would be* analytically true of *the referents of* the T-terms, if they do have any referents. So, in the mental-physical case, it is not a Lewisian claim that mental states pains are, by analyticity, identical to such-and-such causal roles; rather, what is analytic is that *if there are mental states at all*, then they are defined by their causal roles. As Lewis sums up, '[h]ence it is analytic that *either* pain, etc., do not exist *or* most of our platitudes about them are true'.[19]

The argument for analyticity looks clear. However, it is worthwhile to note that the bulk of Lewis' materialist argument could still be sound even if we rejected the analyticity claim. Lewis could infer that the mental state is the brain state without worrying whether the claim from folk psychology about the mental state is analytic or synthetic. Jack Copeland points out a crucial historical fact regarding Lewis' motivation for claiming the conceptual link between mental and physical.[20] The claim was posed in an era where most philosophers endorsed Quine's 'Two Dogmas'.[21] Hence, most

[18] On a side note, Jack Copeland speculates that in this paper of Lewis' makes the first ever mention of the term 'Ramsey sentence'. Thereafter the term was widely used by philosophers. See Chapter 4 for a fuller discussion of Ramisification.

[19] Lewis (1972), p. 257.

[20] Verbal comment.

[21] Quine, W.V.O. (1951), "Two Dogmas of Empiricism" in *The Philosophical Review* 60: 20-43.

philosophers were leaning towards the eliminativist side on the analytic-synthetic distinction. Being a former student of Quine, Lewis' analyticity claim at that time seems very unusual, especially given that much of his materialist theory wouldn't be impaired by the analytic status. Copeland offers a possible explanation on Lewis' motive: it is possible that the argument for analytic status was a deliberate and quiet attempt to evade the Kripkean attack on identity theory. Although Kripke's famous modal argument came four years later than Lewis's 1966 'An Argument for the Identity Theory',[22] Copeland suspects that Lewis had already known it before 1970 since they were close at Harvard during the late 1960s and at Princeton during the 1970s.

Lewis' motive aside, this observation leads to the question of how Kripke's argument works against identity theory, which in turn leads to the first main theme of my thesis.

[22] Kripke's argument is stated in his 1980 book *Naming and Necessity*. Note that the three lectures that the book is based upon were given in 1970.

Part I:

Mirroring and conceivability arguments

Chapter 2. The modal argument against identity theory

2. 0. Outline of Part I

Saul Kripke’s modal argument against identity theory, and David Chalmers’ zombie argument against materialism are two influential arguments in the mind-body debate. In this part of my dissertation I will attempt to show that both these arguments are fallacious. I will do this by showing that these arguments, and indeed any arguments sharing their logical form, are *mirrorable*. That is, for any such argument, there is a corresponding *mirrored argument* that is deductively valid and has a conclusion that contradicts the original argument’s conclusion. I will show that Kripke and Chalmers can challenge the premises of the mirror argument only at the cost of undermining their original argument’s premises.

Part I will be divided into two Chapters. The current Chapter will start by synopsizing Kripke’s argument and will proceed to explain the reformulated version of it that paves the way for the mirroring objection against it. An exposition of the mirroring argument will be given in 2.2. before I assess putative and tentative criticisms of it in 2.3. Chapter 3 is similar, but it attacks Chalmers’ argument instead.

A general diagnosis of the fallacy which both Kripke’s and Chalmers’ arguments commit to has been summarised by Douglas Campbell, Jack Copeland and I in our (2017) ‘The Inconceivable Popularity of Conceivability Arguments’.[23] This article extends the arguments I present in Part II, by targeting not only Kripke and Chalmers, but also the Cartesian argument for mind-body dualism and Alvin Plantinga’s modal ontological argument for theism. This general diagnosis is given in the appendix.

[23] Campbell, D., Copeland, B. J., & Deng, Z. (2017) ‘The Inconceivable Popularity of Conceivability Arguments’ in *Philosophical Quarterly*, Vol. 67 (267), pp. 223-240.

2. 1. Kripke's argument

2. 1. 1. Argument D - the Cartesian root

In his Sixth Meditation, Descartes argues that mind and body are distinct:

> First, I know that everything which I clearly and distinctly understand is capable of being created by God so as to correspond exactly with my understanding of it. Hence the fact that I can clearly and distinctly understand one thing apart from another is enough to make me certain that the two things are distinct, since they are capable of being separated, at least by God... I have a clear and distinct idea of myself, in so far as I am simply a thinking, non-extended thing; and on the other hand I have a distinct idea of a body, in so far as this is simply an extended, non-thinking thing. And accordingly, it is certain that I am really distinct from my body, and can exist without it.[24]

Kripke's argument against identity theory has significant elements that resemble this argument of Descartes.[25] To be precise, they share a key premise. In this section I will consider a particular reconstruction of this Cartesian argument, with the aim of highlighting the premise in question before moving on to explore Kripke's argument.

It is noteworthy to stress that ever since its inception, this Cartesian argument has been subjected to many different interpretations. Among these, some see it as outright fallacious. As stated by Georges Rey, one might reconstruct the argument as follows:

1. Being conceivably unextended is a property of my mind.
2. Being conceivably unextended is not a property of my body.

----

3. My mind is not identical to my body.[26]

At first glance, this version of the argument might appear to have considerable strength by appealing to Leibniz's Law – if two things are identical then they ought to share all of their properties. By specifying a property that my mind possesses and my body doesn't, it seems mind-brain non-identity can be derived via *modus tollens*.

[24] Cottingham, J., Stoothoff, R., and Murdoch, D. (1984). *The Philosophical Writings of Descartes*, p. 54.

[25] Kripke is upfront about the similarity that his argument bears to Descartes'. In footnote 77 of *Naming and Necessity*, Kripke acknowledges this but fends off the suggestion that his argument implies Cartesian dualism. See my footnote xx below for more details.

[26] Rey, G. (1997). *Contemporary Philosophy of Mind*, pp. 56-7.

Unfortunately, the argument is invalid. We can run a parody argument to see why. It seems that one's epistemic status is a good example of discerning properties:

1. Socrates knows what water is.
2. Socrates does not know what H_2O is.

----

3. Water is not identical with H_2O.

Water has the property of being known by Socrates, and H_2O does not have this property. By Leibniz's Law, water is not identical to H_2O. But surely water is H_2O! The message is therefore clear: to restore validity to Descartes' argument, we should avoid the temptation to reconstruct it as a reasoning from Leibniz's Law. Below is one possible step by step reconstruction.

How might Descartes' argument be more charitably interpreted? First, by saying 'I know that everything which I clearly and distinctly understand is capable of being created by God so as to correspond exactly with my understanding of it', Descartes can be seen as making the conceivability entails possibility claim. The conceivability reference is not difficult to spot: if X 'corresponds exactly with my understanding of it', then I can conceive it. The possibility reference, on the other hand, it is more contentious. According to the interpretation I'm putting forward, X 'is capable of being created by God' equates to 'X is possible', at least, in Descartes' sense. Following Dugald Murdoch's (1993, 1999)[27] analysis of this issue, we can at least hold that Descartes has in mind a theistic sense of possibility. This in turn, allows him to claim (perhaps in a theistic way) the conceivability entails possibility principle as his first premise:

D1. It is conceivable that $\Phi \rightarrow$ it is possible that Φ

From D1, Descartes makes a further move by saying 'the fact that I can clearly and distinctly understand one thing apart from another is enough to make me certain that the two things are distinct, since they are capable of being separated'. From being 'capable of being separated'—the possibility of mind-body distinction—he is *certain* that they are distinct. Evidently, Descartes goes from possibility to actuality in a single inference:

[27] Murdoch, D. (1993). 'Exclusion and Abstraction in Descartes' Metaphysics' in *Philosophical Quarterly*, 44 (170), pp. 38-57.
Murdoch, D. (1999). 'The Cartesian Circle' in *Philosophical Review* 108 (2), pp. 221-244.
I am grateful to Dugald Murdoch and Robert Stoothoff for their bibliographical advice on this issue.

D2. It is possible that (mind ≠ body) → (mind ≠ body)

The '$\Diamond$' is dropped without an explanation. This is a tacit premise in his argument, and we can only speculate why Descartes thinks he can infer actuality from possibility. Perhaps Descartes was asserting $\Diamond R1{\neq}R2 \rightarrow R1{\neq}R2$, which is provable in S5. Hence he has some kind of S5 intuition, all those years ago.

After laying down the conditional premises, Descartes announces that mind-body distinction is conceivable by saying 'I have a clear and distinct idea of myself, in so far as I am simply a thinking, non-extended thing; and on the other hand I have a distinct idea of a body, in so far as this is simply an extended, non-thinking thing'. Hence, the third premise is:

D3. It is conceivable that (mind ≠ body)

From the three premises, he concludes 'it is certain that I am really distinct from my body, and can exist without it', and thereby establishes D4:

D4. Mind ≠ body

In a nutshell, the argument will be as follows:

Argument D

D1. It is conceivable that Φ → it is possible that Φ. Assumption

D2. It is possible that (mind ≠ body) → (mind ≠ body) Assumption

D3. It is conceivable that (mind ≠ body) Assumption

D4. Mind ≠ body 1, 2, 3, *modus ponens*

Formalised in this way, the argument is valid in the form of a conceivability argument. I will now present Kripke's argument and explain why it also has this logical form.

2. 1. 2. Argument K - the Kripkean way

Kripke devises a modal argument against identity theory.[28] The argument starts with an analysis of rigid and non-rigid designators. A rigid designator is defined as an expression that always refers to the same object in any possible world in which that object exists.

[28] Kripke , S. (1980). *Naming and* Necessity, pp. 146-55. A brief introduction of the same argument is offered in Kripke's (1977) 'Identity and Necessity' in S. Schwartz (ed.) *Naming, Necessity, and Natural Kinds*, pp. 66-101.

With this notion in hand, Kripke then proceeds to examine identity statements. His claim is that any identity holding between two terms that are rigid designators must be necessary identity.

In the mind-body problem case, if we have on one side of our identity statement an expression referring to a type of mental state rigidly, and on the other side, an expression referring rigidly to a type of brain state, C-fibre stimulations, the statement, ‘Pain = C-fibre stimulation’, would have to be necessarily true, if it were to be true at all. However, there is “a certain obvious element of contingency” of ‘Pain = C-fibre stimulation’ which cannot be explained away like that of ‘Heat = molecular motion’. For even if there is a strict correlation between pain and C-fibre stimulations, all the same, it is easy to imagine that a pain might exist without a C-fibre stimulation existing, and a C-fibre stimulation might exist without a corresponding pain. But, if that is so, then the identity statement is not necessarily true, and if it is not necessarily true, it cannot be true at all. Therefore, it is false. And what goes for the identification of pain with neurobiological events goes for any identification of conscious mental states with physical events. In a nutshell, ‘Pain = C-fibre stimulation’ is not necessary, and thus by *modus tollens*, it must be false.

A simplified formulation of the argument will look like this:

<u>Argument K (simplified)</u>

(A = “pain”, B = “C-fibre stimulation”)

1. A, B are rigid designators $\rightarrow ((A=B)\rightarrow\Box(A=B))$. Kripke’s assumption
2. A and B are rigid designators. Kripke’s assumption
3. $(A=B)\rightarrow\Box(A=B)$ 1, 2, *modus ponens*
4. $\Diamond(A\neq B)$ Kripke’s assumption
5. $A\neq B$ 3, 4, S5, *modus tollens*

The conclusion is drawn from two visible inferences and a hidden one. First, Kripke combines the rigidity of ‘pain’ and ‘C-fibre stimulation’ (i.e. premise 2) with his pre-established principle of necessity of identities between rigid designators, and arrives at sub-conclusion 3. Call this the essentialist inference. Second, the justification of premise 4 comes from Cartesian consideration, namely, that we seem to be able to conceive that the C-fibre stimulation existed without any pain, and conversely, that the pain existed without C-fibre stimulation.[29] Call this the Cartesian inference: $\Diamond_c(A\neq B)$. Introduce a new symbol - '$\Diamond_c$', which reads 'it is conceivable that …'. Thus, $\Diamond_c(A\neq B)$ reads 'it is conceivable that pain is not identical to C-fibre stimulation'. It is not hard to notice a gap needs to be filled between $\Diamond_{\mathrm{C}}$ $(A\neq B)$ and $\Diamond(A\neq B)$. What validates the transition from the former to the latter is a hidden, yet highly debated claim: that conceivability entails possibility (CEP for short). This transition is not given in the simplified argument, and now we can establish the proper formulation of Argument K:

Argument K

K1.	A, B are rigid designators → $((A=B)\rightarrow\Box(A=B))$	Assumption
K2.	A and B are rigid designators	Assumption
K3.	$(A=B)\rightarrow\Box(A=B)$	1, 2, *modus ponens*
K4.	$\Diamond_c(A\neq B)$	Assumption
K5.	$\Diamond_c\Phi \rightarrow \Diamond\Phi$	Assumption (CEP)
K6.	$\Diamond(A\neq B)$	4, 5, *modus ponens*
K7.	$\Diamond(A\neq B) \rightarrow \neg\Box(A=B)$	S5
K8.	$\neg\Box(A=B)$	5, 7, *modus ponens*
K9.	$A\neq B$	3, 8, *modus tollens*

The argument is very influential and has invited many discussions since it came out. To date, most criticisms of Argument K are directed against the essentialist inference – the inference from K1 and K2 to K3. Not long after Argument K came out, Feldman (1973, 1974)[30] proposed the claim that Kripke fails to show 'pain' is essentially pain. Such diagnosis has been followed up by the likes of Boyd (1980),[31] Mucciolo (1975),[32] Rocca (1993),[33] and less elaborately Lewis (1980a), where they also argued in various ways

[30] Feldman, F. (1973). 'Kripke's Argument against Materialism' in *Philosophical Studies*, Vol. 24, pp. 416-19.
Feldman, F. (1974). 'Kripke on the Identity Theory' in *Journal of Philosophy*, Vol. 71, pp. 665-76.

[31] Boyd, R. (1980). 'Materialism without Reductionism: What Physicalism Does Not Entail' in N. Block (ed.). *Readings in the Philosophy of Psychology*, Vol. 1, pp. 1-67.

[32] Mucciolo, L. (1975). 'On Kripke's Argument against the Identity Thesis' in *Philosophia*, Vol. 5, pp. 499-506.

[33] Rocca, M. D. (1993). 'Kripke's Essentialist Argument against the Identity Theory' in *Philosophical Studies*, Vol. 69 (1), pp. 101-12.

against the rigidity of 'pain', hence rejecting K2.[34] A slightly different route was pursued by Lycan (1974a, 1987),[35] McGinn (1977),[36] Sher (1977),[37] and less elaborately, Nagel (1986),[38] as these works focus more on the truth of K1. They deny the necessity of mental-physical identity statements on the grounds that some mental terms are fixed via contingently-associated description, hence are akin to terms like 'heat'. Furthermore,

In short, ever since the birth of Argument K, the opposition has had a predominant interest in the essentialist inference, namely, K1 and K2. It would be ungracious for me to remark critically on this common anti-Kripke chorus. To attack the essentialist inference is to tackle the core of Kripke's theory of reference, which usually initiates topics outside the mind-body discussions. I believe that this line of attack is fruitful, and more to the point, very ambitious – because the rationale behind the strategy is that the entire Kripkean system undermines his anti-physicalist argument. While retaining my admiration of this ambitious strategy, I do believe it is possible to leave Kripke's essentialism untouched and thereby uphold K1 and K2.

Byrne (2007)[39] and Papineau (2001, 2007)[40] have questioned whether Kripke's anti-physicalist argument really rests on a Cartesian premise. They argue that the standard interpretation, which attributes the justification of K6 to conceivability, is incorrect. This line of thought, if correct, would enable a possible refutation of the mirrored argument. As a result, I take this point to be a *de facto* defence of Kripke against my mirrored argument. I will return and access this defence in 2.3.1.

[34] This kind of criticism of K2 can be traced back to a debate in the philosophy of language, namely the debated of whether or not natural kind terms, such as mental state terms are rigid designators. The issue was extensively discussed between Kripke and Putnam, and Putnam's (1973) 'Meaning and Reference' remains to be one of the most comprehensive argumentation against the view that mental terms are rigid.

[35] Lycan, W. (1974a). 'Kripke and the Materialists' in *Journal of Philosophy*, Vol. 71, pp. 577-89. Lycan, W. (1987), *Consciousness*, pp. 15-8.

[36] McGinn, C. (1977). 'Anomalous Monism and Kripke's Cartesian Intuitions' in *Analysis*, Vol. 2, pp. 78-80.

[37] Sher, G. (1977). 'Kripke, Cartesian Intuitions, and Materialism' in *Canadian Journal of Philosophy*, Vol. 7 (2), pp. 227-38.

[38] Nagel, T. (1986). *The View from Nowhere,* pp. 47-8.

[39] Byrne, A. (2007). 'Possibility and Imagination' in *Philosophical Perspectives*, Vol. 21 (1), pp. 125–44.

[40] Papineau, D. (2001). *Thinking about Consciousness*, pp. 47-9.
Papineau, D. (2007). 'Kripke's Proof is Ad Hominem not Two-Dimensional' in *Philosophical Perspectives*, Vol. 21 (1), pp. 475–94.

This leaves us with K4, the Cartesian premise, and K5, the notion that conceivability entails possibility. The latter also seems to provoke constant attention.[41] One camp, best represented by Yablo (1993, 2000, 2002)[42], unveils a line of thought specifically designed to trouble all kinds of conceivability arguments against physicalism including Kripke's. Stephen Yablo distinguishes two notions of possibility: conceptual possibility, which corresponds to conceivability, and metaphysical possibility. An expression is about conceptual possibility if and only if it *could have expressed a true proposition*. An expression is about metaphysical possibility if and only if it *does express a proposition that could have been true*. It is, argues Yablo, a mistake to infer the latter from the former. This challenge has been greeted by replies from Chalmers (2002b)[43], who carefully categorizes conceivability into different kinds and asserts that only the right kind, namely, the 'ideal negative conceivability' is a useful guide to metaphysical possibility. According to Chalmers, an expression is negatively conceivable if it cannot be ruled out *a priori*; an expression is ideally conceivable if and only if it expresses a true proposition in a possible world w, *when w is considered to be actual.* So in regards to Argument K, Chalmers' view amounts to the following: $\Diamond_c(A\neq B)$ *can, but does not necessarily* entail $\Diamond(A\neq B)$. It follows that the inference from K4, K5 to K6 is unproblematic if and only if (a) the agent considers the possible world in which $A\neq B$ is true, to be the actual world, and (b) the agent evaluates the truth-value of $A\neq B$ on mere a priori grounds.[44]

I, myself, refrain from choosing a side to align with or making further comment on this discussion of conceivability and possibility, despite the popularity thereof. My reason for suspending doubt over Kripke's K5 is the same as that for K1 and premise K2. There is a rich list of valuable works devoted to this debate, but to bring the relevant subject—namely, the soundness of Argument K—to bear on some deeper metaphysical issues

[41] For general discussions on whether conceivability entails possibility, I rely on T. Gendler & J. Hawthorne, (2002). (eds.), *Conceivability and Possibility*.

[42] Yablo, S. (2002). 'Coulda, Woulda, Shoulda' in Gendler & Hawthorne (eds), *Conceivability and Possibility,* pp. 441-92.
Yablo, S. (2000). 'Textbook Kripkeanism & The Open Texture of Concepts' in *Pacific Quarterly* 81, pp. 98-122.
Yablo, S. (1993) 'Is Conceivability a Guide to Possibility?' in *Philosophy & Phenomenalogical Research* 53, pp.1-42.

[43] Chalmers, D. (2002b). 'Does Conceivability Entail Possibility?' in Gendler & Hawthorne (eds), *Conceivability and Possibility,* pp. 146-200.

[44] Details of Chalmers' view will be given in 3.1 where I will explain the two-dimensionalist component of his theory.

about the relation between conceivability and modality is, at best, an avoidable move. Therefore, what I am suggesting is an abandonment of usual strategies that focus on either the essentialist inference or the CEP inference. Instead, we shall shift our attention to the only remaining premise, K4. As I shall introduce soon, a trick can be done on this Cartesian premise, making Argument K an unpalatable one for Kripke himself.

2. 2. Mirroring Kripke's argument

The aim of this section is to consider an original way to criticise Kripke's modal argument against identity theory. I intend to show that Kripke's argument has significant pitfalls, not only because his account of rigid designation is prone to attack from philosophy of language's concern but also because his argument can be 'mirrored' and turned into one *in favour of* identity theory with his theory of reference unaltered. To be precise, unlike common criticisms, one need not quarrel about the falsity of Kripkean premises in order to 'mirror' the Kripkean conclusion (that pain is not identical to C-fibre simulation). This 'mirroring' approach, first mentioned in Bayne (1988)[45], is rarely discussed in contemporary literature. By revitalizing Bayne's argument I hope to accomplish two things: (1) to popularize this seemingly forgotten yet ingenious approach and transform it into a Kripkean conceivability argument against Kripke's own conclusion; (2) to generate a model for a new and original objection to Chalmers, which will be dealt with in the proceeding Chapter.

2. 2. 1. Bayne on Argument K

In his (1988), Steven Bayne puts forward a critique against Kripke's argument that does not appeal to the falsity of K1, K2 or K5.[46] Bayne starts by attributing Kripke's argument to Descartes, as reflected in his paper's title – 'Kripke's Cartesian Argument'. The Cartesian dualist view, as abstracted by Bayne, is built upon the following reasoning:

[45] Bayne, S. (1988). 'Kripke's Cartesian Argument' in *Philosophia,* Volume 18 (2-3), pp. 265-9.

[46] In fact, Bayne goes as far as claiming that other critics (those who deny K1, K2, or K5) "fall short in their appraisal" due to their "introduction of new theories or additional ontological commitments" (Bayne (1988), p. 265). However, he shows no arguments in support of this claim. Here, my stance differs from Bayne's, as I only argue that the critics *need not* debate about K1, K2, or K5. The mirroring approach, which I am going to present, is the neatest objection to Argument K, and it's among one of the *correct* ones.

Argument B

B1. The ideas of mind and body are different → $\Diamond_c$(mind exists in the absence of body).

B2. $\Diamond_c$ (mind exists in the absence of body) → mind and body are distinct substances.

B3. The ideas of mind and body are different.

B4. Therefore, mind and body are distinct substances.

B5. Mind and body are distinct substances → mind and body have distinct essences.

B6. Mind and body have distinct essences → $\Box$(mind $\neq$body).

B7. $\Box$(mind$\neq$ body).

Echoing my point given in 2.1.2, Bayne introduces the similarity between Argument B and Kripke's Argument K by citing their common enemy – mind-brain identity theory. According to Bayne, identity theorists' rejection of Argument B is based on a theory of reference, according to which mental and physical terms share the same reference even if they have very different and apparently unrelated 'meanings', or 'senses'. Due to this, the references are the sole measure of identifying what mind and body are.

> Modern materialists argue that mind and body are identical. They maintain that identifying the mental with the physical requires only an identity of what is referred to. Consider 'water=H_2O'. Just because the ideas of water and H_2O are different is no reason, it is argued, to deny that what the ideas refer to are identical.[47]

Whether or not the above-cited paragraph presents an accurate characterisation of the identity theory is a question that I will ignore here.[48] By spelling out the fact that the identity theory only considers the identity between referents of mental terms (such as 'pain') and physical terms (such as 'C-fibre stimulation'), Bayne wants to emphasize Kripke's opposing theory of reference and how that leads to Argument K. He proceeds to offer a brief explanation of Kripke's treatment of 'pain' and 'C-fibre stimulation' as rigid

[47] Bayne (1988). p. 265.

[48] Indeed, it is not. Now I must turn a critical eye towards Bayne. It is obvious that he thinks the referential theory of meaning is one of the main reasons (if not the only one) the identity theory was devised. However, not every identity theorist endorses the referential theory. People came up with this mental-physical identity answer to the mind-body problem for all sorts of reasons, and each has their own way of justifying their identity theory. Bayne's short summary of the identity theory is, therefore, over-simplified and incorrect. I suspect this is deliberately so, in order to bring out Kripke's own theory of reference and how that contributes to Argument K. But surely there is a finer way to introduce Kripke's argument and its similarity to the Cartesian view and dissimilarity to identity theory.

designators and why Kripke takes identity statements involving rigid designators to be necessary (i.e. Kripke's justification of K1 and K2), before calling attention to the Cartesian premise.

For Bayne, Kripke is similar to Descartes when he mentions that there is 'a certain element of contingency' to mental-physical identification. As Bayne puts, it is necessary for pain and C-fibre stimulation to be one and the same thing, if they are actually the same thing. But, it does not appear to be necessary. It appears that the terms 'pain' and C-fibre stimulation' might designate different things. Thus, the identity appears to be contingent. If the identity theorist's argument is to work, he must explain away the appearance of contingency. Moreover, in doing this the identity theorist must accommodate Kripke's point about how the pain concept secures reference to pain directly, rather than via an intermediate mode of presentation. Hence, the mode of presentation of the pain concept is pain itself. So the identity theorists cannot counter the Kripkean move by suggesting that the apparent contingency of the 'Pain = C-Fibre stimulation' statement can be explained away along the same line as the 'Water=H_2O' statement. Therefore, Bayne concludes that identity theory's biggest challenge is to address the apparent contingency of 'Pain = C-fibre stimulation', as clearly highlighted by Descartes and Kripke.

After stating Kripke's argument and noting the seemingly troublesome part, Bayne's next step is to turn it against itself. Firstly, he calls attention to K1-K3:

K1. A, B are rigid designators → $((A=B)\rightarrow\Box(A=B))$.

Bayne contends that given the definition of rigid designation, since A and B are rigid, then not only is the necessity of identity a correct notion, as exemplified in K3:

K3. $(A=B)\rightarrow\Box(A=B)$

But so is the necessity of *non*-identity, as reflected in the following MK3:

MK3. $(A\neq B)\rightarrow\Box(A\neq B))$

The trick is that we can infer a dual of K1 from logic alone without altering its truth.[49] Call this dual MK1:

[49] Note that MK1 is derivable from K1, but only in S5. This point leads to a possible refutation of Bayne's attack. It will be dealt with in detail in 2.3.1.

MK1. A, B are rigid designators $\rightarrow ((A \neq B) \rightarrow \Box(A \neq B))$.

MK1 and K1 are compatible with each other. Bayne's postulation MK1 is by no means an attack on Kripke's principle of necessity of identities. Instead, Bayne preserves that Kripkean notion. The same can be said for K3 and its counterpart MK3. The important message here is that Kripke must not argue against MK1 or MK3 should he want to preserve his K1 and K3.

Likewise, given Kripke's own acknowledgement of the apparent contingency of (A=B), he would, in all fairness, also accept that (A≠B) is apparently contingent. So we have another pair of premises:

K6. $\Diamond(A \neq B)$

and

MK6. $\Diamond(A = B)$

Recall the pre-established MK3, which says $(A \neq B) \rightarrow \Box(A \neq B)$. We now have a *modus tollens* with the conclusion that A=B. To spell things out, Kripke contends that if mental items and physical items are identical, then they must be necessarily identical. After all, A=B only appears to be necessary, because the terms might have been non-identical. On the other hand, Bayne shows the prospect that the mental and physical are necessarily diverse, if they are diverse. He then imitates Kripke's approach by asking what accounts for the apparent contingency of A≠B. Hence, Kripke must explain why $\Diamond(A=B)$ cannot be true. When confronted with this question, Kripke cannot say that we are merely conceiving of an epistemic counterpart of pain, for he has already argued that 'whatever feels like pain is pain'. Thus, he cannot say the possibility of A=B is illusory. If Kripke cannot explain away $\Diamond(A=B)$, then he has no grounds to establish $\Box(A \neq B)$, and consequently losing a key premise for his original *modus tollens*. More importantly, I will argue in 2.3.2 that should Kripke reject $\Diamond(A=B)$, via whatever means, he would in fact render Argument K redundant.

2. 2. 2. Argument MK

So much for Bayne's reply. I must clarify two points regarding Bayne's contribution. First, Bayne's final claim is that Kripke's modal argument leads to a paradox since the postulation of MK1 is consistent with, and derivable from K1.

> I conclude that Kripke's subtle argument against mind-body identity leads to paradox, and appears to be, therefore, without force.[50]

For this reason, I do not regard Bayne's conclusion as a resurrection of identity theory. Second, what Bayne has not done here, is to attribute this appearance of contingency to conceivability, at least not explicitly. However, I would argue that the postulation of B6 is inferred by the aforementioned Cartesian premise, with a subtle but fundamentally profound twist. To be precise, K4: the claim that $\Diamond_c(A \neq B)$ can be mirrored and turned into what I call MK4:

MK4. $\Diamond_c(A=B)$

MK4 is a counterpart of Kripke's Cartesian premise K4:

K4. $\Diamond_c(A \neq B)$

So far, we have two mirrored premises that Bayne devises in MK1 and MK2. Together, they derive MK3. Mirroring K4 gives us MK4. Let us also retain Kripke's K5, the CEP premise, and use it as MK5. From what we have gathered, a mirrored version of Argument K is yielded. Call it Argument MK:

Argument MK

(A = 'pain', B = 'C-fibre stimulation')

MK1. A, B are rigid designators → $((A \neq B) \rightarrow \Box(A \neq B))$. Assumption

MK2. A and B are rigid designators. Assumption

MK3. $(A \neq B) \rightarrow \Box(A \neq B)$ 1, 2, *modus ponens*

MK4. $\Diamond_c(A=B)$ Assumption

MK5. $\Diamond_c\Phi \rightarrow \Diamond\Phi$ Assumption (CEP)

MK6. $\Diamond(A=B)$ 4, 5, *modus ponens*

MK7. $\Diamond(A=B) \rightarrow \neg\Box(A \neq B)$ S5

[50] Bayne (1988). p. 268.

MK8. $\neg\Box(A\neq B)$ 6, 7, *modus ponens*

MK9. $\neg(A\neq B)$ 3, 8, *modus tollens*

The conclusion, MK9, is a negation of Kripke's conclusion K9. For this reason, Kripke needs to reject Argument MK. But what should he target? Argument K and Argument MK share the same form, so arguing for the invalidity of Argument MK is not a good option, for doing so would undermine the validity of his own Argument K. He must, therefore, hold the position that Argument K is sound and Argument MK is unsound, and thereby reject one or more premises of Argument MK. But which premise of Argument MK *can* Kripke reject? Let us go through each argument's premise one by one. The first premises in the two arguments are:

K1. A, B are rigid designators → $((A=B)\rightarrow\Box(A=B))$

MK1. A, B are rigid designators → $((A\neq B)\rightarrow\Box(A\neq B))$

As discussed above, the pair is comprised of logical duals, with K1 stating the necessity of identity between rigid designators and MK1 stating the necessity of non-identity between rigid designators[51]. Both are Kripkean notions, so it is impossible for Kripke to maintain one and reject the other. This is more obvious in the second premises. K2 and MK2 are identical:

K2. A, B are rigid designators.

MK2. A, B are rigid designators.

Since Kripke must accept MK1 and MK2, he must also accept MK3, for it is a *modus ponens* inference of the previous two premises, just as K3 is a *modus ponens* inference from K1 and K2.

The next line, namely K4 and MK4, is where the difference occurs:

K4. $\Diamond_c(A\neq B)$

MK4. $\Diamond_c(A=B)$

In the line after, we have an identical pair again in K5 and MK5:

K5. $\Diamond_c\Phi \rightarrow \Diamond\Phi$

MK5. $\Diamond_c\Phi \rightarrow \Diamond\Phi$

[51] An argument for MK1 is given on pg. 31.

In both arguments, every line after the fifth is a purely logical step. In short, among the four premises (1, 2, 4, and 5), 4 is the only plausible target for Kripke. In general, he could respond in two different ways: a) to reject MK4, and b) to reject the inference from MK4, MK5 to MK6, which amounts to saying that CEP works for A≠B and somehow fails for A=B. I will argue that neither is a viable option for Kripke. Before I explain my argument, I shall first discuss some of the concerns that have been voiced in the literature.

2. 3. Possible objections to Argument MK

2. 3. 1. Putative concerns

Although Argument MK is a seldom-discussed objection to Kripke, there have been fruitful considerations that might undermine the argument. Among these putative concerns, I take three of them to be valuable. But none of them, I contend, is effective. I am going to assess the trio in what I believe to be the ascending order of strength.

First, there is the question of whether Argument K really contains a Cartesian premise. More precisely, the critics allege that Kripke's argument should not be interpreted as a conceivability argument, and if so the inclusion of K4 needs to be avoided.[52] By the same token, not only is Argument K a wrong reconstruction of Kripke, as the critics claim, but Argument MK also fails to be a feasible mirrored version of Kripke for its inclusion of MK4, which states $\Diamond_c(A=B)$. It follows that to maintain the strength of the mirroring argument, I need to show evidence that Kripke's own argument has a premise that claims 'it is conceivable that pain is not C-fibre stimulations'.

Before we look at this evidence, one point needs to be recognised. That is, Kripke attempts to make it clear that his argument against identity theory is *not* a descendant of Cartesianism. In footnote 77 of *Naming and Necessity*, Kripke writes:

> Having expressed these doubts about the identity theory in the text, I should emphasize two things: … Second, rejection of the identity thesis does not imply acceptance of Cartesian dualism. In fact my view above that a person could not have come from a different sperm and egg from the ones from which he actually originated implicitly suggests a rejection of the Cartesian picture.[53]

[52] This criticism is offered in Byrne (2007) and Papineau (2001, 2008).

[53] Kripke (1980). p. 155. Author's italics.

Perhaps it is the kind of reluctance Kripke shows here that encourages the critics to argue for the removal of Cartesian references when reconstructing Kripke's argument. However, what Kripke tries to stay far away from is the Cartesian dualist theory as a whole. The presentation of the anti-identity theory argument, in his own words, echoes Descartes' crucial premise – namely, the claim that 'it is conceivable that mind is not the body'. Of course, in Kripke's argument that premise becomes 'it is conceivable that pain is not C-fibre stimulation'. The clearest textual evidence of Kripke making such a claim is in the following passage:

> To be in the same epistemic situation that would obtain if one had a pain *is* to have a pain. The apparent contingency of the connection between the mental state and the corresponding brain state thus cannot be explained by some sort of qualitative analogue as in the case of heat.[54]

The mention of 'epistemic situation' and 'apparent contingency' has been interpreted as Kripke's way of expressing '$\Diamond_c$'. For example, William Lycan reconstructs Kripke's key premise as follows:

> (*D*) If *a* and *b* are "distinguishable" in the sense that we seem to be able to imagine one existing apart from the other, then it is possible that $a \neq b$, *unless* (i) "someone could be, *qualitatively* speaking, in the same epistemic situation" *vis-à-vis a* and *b*, and still "in such a situation a *qualitatively* analogous statement could be false," or [let us add] (ii) there exists some third alternative explanation of the distinguishability of *a* and *b*.[55]

Likewise, in summarising the Kripkean argument, John Searle says:

> It does not seem right to say either that pains in general are necessarily brain states, or that my present pain is necessarily a brain state; because it seems easy to imagine that some sort of being could have brain states like these without having pains and pains like these without being in these sorts of brain states. It is even possible to conceive a situation in which I had this very pain without having this very brain state, and in which I had this very brain state without having a pain.[56]

It is hard to conjure up a notion other than conceivability that phrases like 'we seem to be able to imagine', 'it seems easy to imagine', and 'it is even possible to conceive a situation' are referring to. Moreover, even if Kripke really has in mind another notion that is different from conceivability, the mirroring approach can still go through. Suppose the

[54] Ibid. P. 152.

[55] Lycan (1987). p. 12. Author's italics and brackets.

[56] Searle, J. (1992). *The Rediscovery of Mind*, p. 39.

critics were right, and instead of $\Diamond_c$ let us use $\Diamond_x$ to notate whatever Kripke were to mean by 'epistemic situation' and 'apparent contingency'. The fourth premise in his argument would become:

K4*. $\Diamond_x(A \neq B)$

On the other side, the fourth line of the mirrored argument can be changed to:

MK4*. $\Diamond_x(A=B)$

The result remains the same as the original mirroring approach, unless Kripke and his defenders can prove that MK4* is false while maintaining the truth of K4*, and if they attempt to do so they would face the same difficulty as in denying MK4 while holding K4.

There is a subtle point about modal logic that needs to be addressed. Bayne was upfront about the fact that the postulation of MK1 could be vulnerable from a certain angle. As mentioned, MK1 is derivable from K1, but only in modal system S5. In response, Bayne explains that the modal system is of no importance as long as we stay in the Kripkean context.[57] For Kripke, the issue at hand is whether necessity of identity of pain and C-Fibre stimulation holds (i.e. whether K1 is true), and this is can be confirmed by a check of rigidity on the terms 'pain' and 'C-fibre stimulation'. Because to say that the a term is rigid, is by definition, just to say that the term always denotes the same thing in every possible world that it has a referent. Given the antecedent addressed in MK1 that A≠B, i.e. pain ≠ C-fiber stimulation, it follows trivially that 'pain' and 'C-fibre stimulation' denote different things in every possible world that they have a referent.

A *reductio* helps to elucidate this point. If we assume the falsity of MK1, then we would have a possible world in which (pain ≠ C-fibre stimulation) & (pain = C-fibre stimulation) is true despite the fact that 'pain' and 'C-fibre stimulation' are rigid terms. Therefore, the message is clear: when confronted with the challenge against the modal system S5, MK1 and K1 are in an equally bad situation. If this criticism knocks out MK1, it also knocks out Kripke's K1. Nevertheless, the S5 challenge does highlight a crucial point - that MK1 and K1 can escape such attack only because they are bound to a Kripkean antecedent. If Kripke's analysis of the necessity of identities between rigid designators collapsed, so would K1 and MK1. But this is not a problem for the mirroring

[57] Bayne (1988). p. 267.

strategy, since our initial motive is to construct an argument from Kripke's point of view, using his own theory of modality. After all, as one of the founders of modal semantics, Kripke will have no problem justifying S5.[58] What does seem alarming, is that if a conceivability argument of a similar kind is constructed outside the Kripkean context, and it happens to include a premise of the following form, then that argument would not be able to dodge the S5 challenge and must provide a direct justification for the application of a particular modal system, namely S5.[59]

Next, in one of the initial critical replies to Bayne, Alex Blum suggests that the dualists would never agree with our premise MK4—the claim that A=B is conceivable—because A=B is only apparently conceivable when in fact it is not, and more importantly, there is a way to explain why A=B appears to be conceivable.[60]

Blum continues by claiming that a physicalist must accept that the apparent contingency of pain=C-fibre stimulation can't be explained. In other words, the physicalists must agree with Kripke's K4, because the physicalists have no method to explain why A≠B appears to be conceivable when in fact it is not. Such talk of 'something appears to be conceivable when in fact it is not', and 'finding a way to explain away apparent contingency' leads us to the third objection to the mirroring of Kripke. Blum's brief idea is developed and enlarged in detail by Eddy Zemach in his (1994).[61] Zemach begins by revisiting Argument K and Bayne's mirrored version of it. I must stress that, Zemach's interpretation of Kripke and Bayne is slightly different to mine, as he does not explicitly attribute conceivability to either of the arguments. His criticism is centered around Bayne's postulation of $\Diamond(A=B)$, not my $\Diamond_c(A=B)$. Nonetheless, I think Zemach's objection, which considers the idea of epistemic counterparts of (A=B) and (A≠B), shall be better formulated with a subscript '$_c$'. Zemach points out that the vulnerable parts in both arguments are the fourth premises:

K4. $\Diamond_c(A\neq B)$

MK4. $\Diamond_c(A=B)$

[58] Copeland, B. J. (2002). 'The Genesis of Possible Worlds Semantics' in *Journal of Philosophical Logic*, Vol. 31, No. 2, pp.99-137. As pointed out by Copeland, Kripke has axiomatised S5 in his (1959) and proved completeness for S5 in his (1963a) and (1963b).

[59] I am grateful to Charles Pigden for pressing that ditching S5 is the solution to the problem.

[60] Blum, A. (1989). 'Bayne on Kripke' in *Philosophia,* Volume 19, Issue 4, pp. 455-6.

[61] Zemach, E. (1994). 'Identity and Epistemic Counterparts' in *Philosophia,* Volume 23, Issue 1-4, pp. 265-70.

Initially, Kripke targets the identity theory's claim that (A=B). By establishing K4, Kripke is suggesting two things: one, the identity claim appears to be contingent, and two, if the apparent contingency cannot be explained by the physicalists, then the claim that (A=B) is false. Kripke's whole argument is built around these two rationales. Zemach's objection re-interprets Kripke's rationale with an introduction of the idea of the *epistemic counterpart.* According to Zemach, the idea can be defined as follows

> Definition. X is an epistemic counterpart of Y iff X is phenomenally indistinguishable from Y.

Thus in Zemach's semantics, Kripke's line of thinking shows that the physicalists must find an epistemic counterpart to (A≠B), and if they fail to find such, Kripke wins. In other words, Zemach stresses the physicalists' task:

> Task P. Imagine X such that X is identical to pain and is phenomenally distinguishable from pain.

Our mirroring of Kripke, on the other hand, amounts to a quest for (A=B)'s epistemic counterpart. To spell it out, the establishment of MK4 suggests that Kripke, or the dualists who hold the same form of argument, must provide an explanation for the apparent contingency of their own premise - (A≠B). This can be achieved, says Zemach, by locating the epistemic counterpart to (A=B). If the dualists fail to do so, then the identity theorists win. The dualists' task is this:

> Task D. Imagine Y such that Y is not identical to pain and is phenomenally indistinguishable from pain.

Now, Zemach's objection to Bayne's can be simply characterised as follows: Kripke and his defenders have the upper hand because a) they can successfully finish Task D, and b) the defenders of the mirroring strategy cannot accomplish Task P.

Zemach's reason for b) is solely a Kripkean one: 'whatever feels like pain is pain'. Logically, the Kripkean notion also amounts to 'whatever is pain feels like pain'. Therefore, nobody is able to imagine pain as phenomenally distinct from pain. The backbone of this inability is that people epistemically grasp the concept of pain via feeling pain. So *prima facie*, Zemach is right about b). The mirroring advocates seem not to be able to find the epistemic counterpart of pain because an epistemic counterpart of pain is, and must be, pain.

The justification for a), however, is much more problematic. Can someone imagine Y such that Y is phenomenally indistinguishable from pain yet is not pain? Zemach suggests an affirmative answer and explains that for A and B to be phenomenally indistinguishable is just for them to have the same causal property – that is A and B have exactly the same causes and effects. For short, according to Zemach,

Definition. For concept A, concept B and causal property P, A is phenomenally indistinguishable from B if and only if $\neg(\exists P)(P(A) \& \neg P(B))$.

But this should not be confused with identity, because one cannot infer A=B from $\neg(\exists P)(P(A) \& \neg P(B))$ alone since

> One cannot observe that the things a and b are identical; all that can be established by observation is that they are empirically indistinguishable … identity to B is not a phenomenal feature of [A].[62]

Hence, $(A\neq B) \& \neg(\exists P)(P(A) \& \neg P(B))$ is a derivable well-formed formulae. Putting this into plain English we have the description of Task D: A is not identical to B but is conceived to be phenomenally indistinguishable from B. In a nutshell, by giving an analysis of phenomenal indiscernibility in terms of causal indiscernibility, Zemach presents a theoretic way of how the dualists can find the seemingly perfect epistemic counterpart of A=B.

So far so good. However, Zemach is wrong to argue from the accomplishment of Task D to the failure of physicalists (or the victory of Kripke). The pitfall of his objection is nothing to do with the abovementioned way of finding an epistemic counterpart of A=B. Nor is his analysis of phenomenal indiscernibility a flawed one. In fact, I am in full agreement with Zemach that phenomenal indiscernibility can be analysed in causal terms, and that identity shall not be confused with phenomenal indiscernibility. In short, I can grant that Zemach is right in saying that Task D can be completed and the dualists really can find a possible world in which A is phenomenally indistinguishable from B yet is non-identical to B. But, this inflicts little or no harm to the physicalists, unless 'epistemic counterpart of X' only means 'a thing that is phenomenally indistinguishable from X'.

[62] Zemach (1994). p. 266.

The trouble with Zemach's criticism, as I suggest, is due to his definition of 'epistemic counterpart', which can be put as follows:

> Definition. X is an epistemic counterpart of Y iff X is phenomenally indistinguishable from Y.

As explained above, according to this definition, $\neg(\exists P)(P(A) \& \neg P(B))$ would be a perfect epistemic counterpart for A=B. However, if I tried to imagine A=B, I would not have in mind a possible world in which $\neg(\exists P)(P(A) \& \neg P(B))$ is a true proposition, and I would not conceive of A and B being only phenomenally indistinguishable. Instead, I would have in mind a possible world in which A=B is a true proposition.[63] What follows is that in order to save Argument K, the dualists would have to do more than just conjure up the truthmaker for $\neg(\exists P)(P(A) \& \neg P(B))$.

Nonetheless, Zemach's objection to Bayne's critique of Kripke does bring out an important aspect of Argument K and its mirrored variant Argument MK. He presents the issue in a discussion about epistemic counterparts. One side can find the epistemic counterpart to its opponent's key claim, while the other side cannot. The backbone of the issue lies in the apparent incompatibility of K4 and MK4, in that it seems that one of the two premises, $\Diamond_c(A\neq B)$ or $\Diamond_c(A=B)$, must be false. The message is this: if we are allowed to apply a modal logic system that grants the joint truth of $\Diamond_c(A\neq B)$ and $\Diamond_c(A=B)$, then Kripke's argument would, again, need to address its reliance on S5.

2. 3. 2. Kripke's *quadlemma*

Having discussed the concerns voiced in existing literature, I now turn to analyse a final move Kripke and his defenders may attempt to make in order to circumvent the mirroring objection. By now it should be clear that MK4 is the only plausible target for Kripke, the other options being untenable for reasons already discussed. As mentioned, he could respond in two different ways: a) by questioning MK4, or b) by rejecting the inference from MK4 and MK5 to MK6, which amounts to say that while CEP works for A≠B it somehow fails for A=B. In order to accomplish a), he might take one of the two following positions:

[63] This refutation of Zemach was suggested to me by Douglas Campbell.

Position 1: Prove outright that MK4 is false, by providing a sound argument for its negation. This amounts to establishing the truth of the following proposition:

$$\neg\Diamond_c(A=B)$$

Position 2: Concede that both K4 and MK4 are true and show that our reasons for accepting MK4 is true are less secure than our reasons for accepting K4, so that K4 is a more reliable conceivability premise that MK4. This amounts to establishing the truth of the following:

$$\Diamond_c(A\neq B)\ \&\ \Diamond_c(A=B)\ \&\ (A\neq B \textit{ is more conceivable than } A=B)$$

On the other hand, should he opt for b), Kripke would take what I call *position 3*:

Position 3: Claim that the *modus ponens* on line 6 of Argument MK fails to work while the *modus ponens* on line 6 of Argument K works. In other words, CEP works for A≠B but fails to work for A=B, and it amounts to establishing the truth of the following:

$$(\Diamond_c(A\neq B) \rightarrow \Diamond(A\neq B))\ \&\ (\Diamond_c(A=B)\ \&\ \neg\Diamond(A=B))$$

I will now assess each of these three positions and explain why none of them is feasible.

First, *position* 1. Suppose Kripke can establish $\neg\Diamond_c(A=B)$, To do this he needs a justification, in the form of some sound argument for $\neg\Diamond_c(A=B)$. Call this Argument KZ. It will be possible to frame KZ in the following form where Ψ[64] can be substitute with whatever a good justification for $\neg\Diamond_c(A=B)$ might be:

Argument KZ

1. Ψ
2. $\Psi \rightarrow \neg\Diamond_c(A=B)$

----

3. $\neg\Diamond_c(A=B)$

What exactly instantiates Ψ is of little importance here. In all fairness, let us assume that there really is a reason for the inconceivability of A=B. *Prima facie*, once Kripke finds this reason, he wins. That is, if Argument KZ is a sound argument, MK4 is false and so the mirroring objection collapses. However by rebutting the mirroring objection in this

[64] In the current form of argumentation, Ψ is presented as an antecedent for $\neg\Diamond_c(A=B)$, hence the argument takes the form of *modus ponens.* This need not be the only form of the argument. An alternative construction might put Ψ as a consequent of $\Diamond_c(A=B)$ and contain a premise that rejects Ψ and derive the conclusion that $\neg\Diamond_c(A=B)$ via *modus tollens*. Nonetheless, the variety of the form of Argument KZ does not fix *position 1*'s problem.

way, Kripke would be creating a potential problem for his own Argument K. Recall Argument K's conclusion:

K9. $A \neq B$

We can add just one more premise to Argument KZ to derive Argument K's conclusion:

Argument KZ (extended)

1. Ψ
2. $\Psi \rightarrow \neg\Diamond_c(A=B)$

----

3. $\neg\Diamond_c(A=B)$
4. $\neg\Diamond_c\Phi \rightarrow \neg\Diamond\Phi$ Assumption (*IEI*)

----

5. $\neg\Diamond(A=B)$ 3, 4, *modus ponens*
6. $A \neq B$ 5, S5

The logic here isn't complicated as the reasoning is a simple and valid one: if it is inconceivable that A=B then it is impossible that A=B, because *whatever is inconceivable is impossible*. The issue at hand is around this newly added *inconceivability entails impossibility* principle (IEI for short). While CEP has always been a hotly debated topic, IEI seems much less contentious.[65] In proving CEP, one needs to first conjure up a conceivable scenario—that is, to eliminate the possibility that a state *entails a logical contradiction*—and then try to determine whether the *lack of logical contradiction* equates to possibility of that state. The fundamental difficulty of affirming CEP resides in the first step. The mental elimination process, in some cases, is enormously large. The apparent lack of logical contradiction in a state may be due to the incompleteness of the elimination process. Just because I can't think of a contradiction entailed by state X doesn't imply that there is no contradiction entailed by state X. Contra wise, affirming IEI is more like a proof by counter-example. One only needs to detect that a state entails a

[65] For someone who holds that conceivability is biconditional to possibility, both CEP and IEI will be resounding principles. I am grateful to Mark Steiner for suggesting to me that Hume can be interpreted as one philosopher who holds such view. In Book II, Section ii of *A Treatise of Human Nature*, Hume writes:

> *...nothing we imagine is absolutely impossible. We can form the idea of a golden mountain, and from thence conclude that such a mountain may actually exist. We can form no idea of a mountain without a valley, and therefore regard it as impossible.*

logical contradiction, then she can infer from the *occurrence of a single logical contradiction* to the impossibility of that state.[66]

However, our justification for IEI relies on a particular interpretation of conceivability, namely, *negative conceivability*, according to which, X is conceivable if X doesn't entail a logical contradiction. It must be noted here that this particular notion of conceivability isn't agreed by all. Chalmers argues that negative conceivability is not a reliable guide to possibility, and concludes on this basis that objections to conceivability arguments that use such a notion are flawed.[67] It will be unfair to neglect this possible way of rejecting the mirroring strategy. Discussion on the 'variety of conceivability' reply is given in Appendix.

It is for the aforementioned reason that $\neg\Diamond_c(A=B)$ entails $A\neq B$. Hence, Argument KZ yields the conclusion of Argument K. Thus, if Kripke is in a position to defend Argument K against Argument MK using Argument KZ, then this defence could itself be turned into a proof against identity theory, without using Argument K in the first place. Copeland, Campbell, and I call this logical phenomenon *self-sacrifice*. An Argument *U* is self-sacrificial iff refuting Argument *U*'s mirror *W* has the effect of establishing *U*'s conclusion.[68] In the present case, Kripke might construct an argument in the form of KZ to rebut our mirroring objection against his Argument K. If he can, he would break the mirror *and* establish the falsity of identity theory—the conclusion of Argument K—without so much as mentioning Argument K. Triumph! Not only would this defence rectify Argument K, it would also be another argument against identity theory. However, this is not good news for the Kripkean camp. Not yet, at least. The shortcoming of this kind of reply is subtle but undermining: the soundness of Argument K, hence the falsity of identity theory, relies on the soundness of Argument KZ – an argument that hasn't been produced yet. In all fairness, it is entirely possible for Kripke or anyone to conjure up the details of Argument KZ and thereby to successfully debunk the mirroring attack.

[66] For more detailed discussions of the role of IEI in Argument KZ, see 3.3.2.

[67] Chalmers, D. (2010), *The Character of Consciousness*, pp. 143-66.

[68] Copeland, B.J., Campbell, D.I. & Deng, Z. (2017), 'Mirroring, Zombies, and Non-Reductive Consciousness', *Digitalization for a Sustainable Society*: The 2017 Summit of the International Society for Information Studies, Chalmers University of technology, Gothenburg, Sweden (presented by Copeland).

But until Argument KZ has come to fruition, Argument K's soundness is *suspended*. Thus, by devising the mirroring attack, we have exposed a potential problem of Kripke's argument *and* we have kindly given away the recipe for fixing this problem. Whether or not Kripke can deliver the fix, the burden is on him.

Unlike the first option, *position 2* does not require denying the conceivability of A=B. Rather, the Kripkean defenders allege that both A=B and A≠B are conceivable but the latter is somehow more conceivable than the former. As noted, this is to claim the truth of a triadic conjunctive proposition that says:

$$\Diamond_c(A\neq B)\ \&\ \Diamond_c(A=B)\ \&\ (A\neq B \text{ is more conceivable than } A=B)$$

One *prima facie* concern arises immediately. Since A≠B and A=B are negations to each other, a conjunction involving the pair must be false, hence:

$$\neg((A\neq B)\ \&\ (A=B))$$

It follows that from *a priori* inspection alone we can conclude that (A≠B) & (A=B) is inconceivable:

$$\neg\Diamond_c((A\neq B)\ \&\ (A=B))$$

The absolute absurdity of the conceivability of something of the form 'P&¬P' should be recognised by all rational agents including friends of *position 2*. However, they need not worry about this. No doubt, in K4 and MK4, the contents of conceiving are two contradictory states of affairs. But this is not to say that one of them *entails detectable contradiction*. From *a priori* inspection, we cannot infer that neither A≠B nor A=B is conceivable from the fact that their conjunction is not conceivable. I contend, on behalf of proponents of *position 2*, that contrary to prima facie considerations, $\Diamond_c(A\neq B)\ \&\ \Diamond_c(A=B)$ is not an inconsistent pair of statements. The conceivability of A≠B and the conceivability of A=B can be jointly true, despite the obvious absurdity of the conceivability of (A≠B) & (A=B). In short, conceivability is not a *distributive property* – unlike the mathematical truth of a*(b+c) = a*b + a*c, the following is incorrect:

$$\Diamond_c((A\neq B)\ \&\ (A=B)) \leftrightarrow \Diamond_c(A\neq B)\ \&\ \Diamond_c(A=B)$$

It follows that the first two conjuncts in *position 2*'s claim need not have opposite truth-values and they *can* be jointly true. The drawback of *position 2*, instead, resides in the third conjunct:

A≠B is more conceivable than A=B

Here, the underpinning assumption is that something could be more conceivable another. I find this extremely puzzling, and I have two considerations that may cast doubt over its plausibility. The first point concerns the meaning of ‘more conceivable’. What exactly does it mean for X to be more conceivable than Y? According my interpretation of conceivability stated above, $\Diamond_c P$ obtains if there is no logical contradiction in P. Equally, $\Diamond_c Q$ obtains if there is no logical contradiction in Q. If so, proponents of *position 2* seem to be saying something like ‘it is less likely for P to entail a logical contradiction than Q’. This claim is outright false. Whether or not a proposition entails a contradiction is a yes or no matter, the talk of likelihood thereof is simply incomprehensible. Nonetheless, the talk of likelihood does make sense if the topic is about *our ability to detect logical contradiction in propositions*. Perhaps what friends of *position 2* have in mind is something like ‘it is less likely *for us to detect a logical contradiction in P than in Q*’. This claim is not outright false, but it is troublesome. If the justification for ‘A≠B is more conceivable than A=B’ rests on the stipulation that ‘it is less likely for someone to detect a logical contradiction in A=B than in A≠B’, then it paves the way for mirroring attack round two! That is, proponents of the mirroring strategy can say that ‘A=B is more conceivable than A≠B’ for ‘it is less likely for someone to detect a logical contradiction in A≠B than in A=B’. To reject this mirroring claim is therefore to assume a discrepancy between our epistemic accesses to identity and non-identity. To be more precise, foes of the mirroring strategy might allege that finding a contradiction in non-identity statements (e.g. A≠B) demands more work than finding one in identity statements (e.g. A=B). By the same token, they might suggest that proving an identity demands more work than proving a non-identity.[69] But this would be an independent metaphysical stipulation that requires justification of its own. The mirroring camp need not respond until it is proven. In short, the burden of proof is not on our side.

Alternatively, by saying ‘more conceivable’ the Kripkean defenders could be indicating some notion along the lines of ‘truer than’. In other words, they could suggest ‘the truth-value of $\Diamond_c A \neq B$ is greater than the truth-value of $\Diamond_c A = B$’. This leads to my second consideration. It is possible to read *position 2* in terms of fuzzy logic, and we can transform this conceivability contest between $\Diamond_c A \neq B$ and $\Diamond_c A = B$ into a truth-value

[69] I am grateful to John Bigelow and Cathy Legg for making this point.

contest. Using Lofti Zadeh’s framework[70], the following truth table illustrates such a view:

Table 2-a

$\Diamond_{c}A\neq B$	$\Diamond_{c}A=B$	‘$A\neq B$ is more conceivable than $A=B$’/’the truth-value of $\Diamond_{c}A\neq B$ is greater than the truth-value of $\Diamond_{c}A=B$’
1	0 ~ 1	1

One might propose a fuzzy logic defence of *position 2*, according to which the value of $\Diamond_{c}A=B$ is not 0 but between 0 and 1, while $\Diamond_{c}A\neq B$ takes the value of 1, thus $\Diamond_{c}A\neq B$ has a higher value than A=B. Then, since ‘A≠B is more conceivable than A=B’ simply means ‘the truth-value of $\Diamond_{c}A\neq B$ is greater than the truth-value of $\Diamond_{c}A=B$’ according to this defence, ‘A≠B is more conceivable than A=B’ will take the value of 1.

Unfortunately, this move is untenable. Recall the proposition that *position 2* aims to establish:

$$\Diamond_{c}(A\neq B)\ \&\ \Diamond_{c}(A=B)\ \&\ (A\neq B \text{ is more conceivable than } A=B)$$

Because the value of the second conjunct is between 0 and 1, it trumps the value of the whole conjunction to 0, as Table 2-b shows:

Table 2-b

$\Diamond_{c}A\neq B$	$\Diamond_{c}A=B$	$A\neq B$ is more conceivable than $A=B$’/’the truth-value of $\Diamond_{c}A\neq B$ is greater than the truth-value of $\Diamond_{c}A=B$’	$\Diamond_{c}(A\neq B)$ & $\Diamond_{c}(A=B)$ & ($A\neq B$ is more conceivable than $A=B$)
1	0 ~ 1	1	0

To sum up, I have given two reasons for the implausibility of *position 2*. Nevertheless, I concede that these two considerations of mine do not exhaust the scope of possible replies. It is entirely possible for the friends of *position 2* to conjure up new justifications.

[70] Zadeh, L. (1975). ‘Fuzzy Logic and Approximate Reasoning’ in *Synthesis*, Vol. 30 (3-4), pp. 407-28.

But until then, MK4 is not a guaranteed loser in this so-called conceivability contest against K4.

I now turn to examine *position 3*. Whereas the first two options target MK4, the third Kripkean reply to the mirroring argument is to circumvent this mirroring premise. Rather, it argues that the inference form MK4, MK5 to MK 6 is not a valid one and therefore Argument MK is invalid:

…

MK4. $\Diamond_c(A=B)$ Assumption

MK5. $\Diamond_c\Phi \rightarrow \Diamond\Phi$ Assumption (CEP)

MK6. $\Diamond(A=B)$ 4, 5, *modus ponens*

…

Meanwhile, this reply also has to maintain that the corresponding steps in Argument K is valid:

…

K4. $\Diamond_c(A\neq B)$ Assumption

K5. $\Diamond_c\Phi \rightarrow \Diamond\Phi$ Assumption (CEP)

K6. $\Diamond(A\neq B)$ 4, 5, *modus ponens*

…

It is visible that the derivations in both arguments use the same rule of logic - namely, *modus ponens*, and the same categorical premise - CEP. Furthermore, since *position 3* also wants to hold that the conceivability premises in both arguments are true, the position is not to claim the inconceivability of A=B; instead, it claims A=B is conceivable but not possible:

$$(\Diamond_c(A\neq B) \rightarrow \Diamond(A\neq B)) \ \&\ (\Diamond_c(A=B) \ \&\ \neg\Diamond(A=B))$$

This is the general strategy of *position 3*. It calls for a modification of CEP so that it works for A≠B but fails for A=B. In other words, it needs to claim that conceivability *does not always* entail possibility. It follows that Φ in $\Diamond_c\Phi \rightarrow \Diamond\Phi$ has a restricted membership. A≠B is one of the members, A=B is not.

Kripke can successfully deny Argument MK's validity only if it can offer a sound argument for this restricted-membership-account of CEP (RCEP for short). In adopting position 3, he must defend the following claim, where here *R* denotes a group of propositions excluded from CEP:

(RCEP). $(\Diamond_c\Phi \rightarrow \Diamond\Phi) \rightarrow \Phi \notin R$

A proposition is possible if it is conceivable, *only if* it doesn't belong to *R*.

Now, is RCEP a true principle? The short answer is 'yes'. RCEP specifies that there are counterexamples to CEP. This is true. One famous counterexample, according to Kripke, is the 'Water ≠ H_2O' case. It is conceivable that water is not H_2O, but as Kripke explains, this is merely an *a posteriori* impossibility for 'water' is a natural kind term that secures its reference by specifying an accidental property of its referent.[71]

Therefore, 'Water ≠ H_2O' is one example of a proposition belonging to *R*. More generally, any proposition, *p*, satisfying the following condition will be a member of *R*, and will therefore fall outside the scope of CEP:

Condition 0. *P is* a proposition that contains natural kind terms that secures reference by specifying an *accidental* property of its referent.

Now, with this in mind, let us ask whether A=B is a similar to 'Water = H_2O' in being a member of R. If the answer is 'yes' then it is excluded from CEP's scope, and so Kripke can safely grant that A=B is conceivable while still denying that it is possible. However, the answer according to Kripke himself is a resounding 'no'. Recall that one of his key premises is 'whatever feels like pain is pain':

> Pain, on the other hand, is not picked out by one of its accidental properties; rather it is picked out by the property of being in pain itself, by its immediate phenomenological quality. Thus, pain, unlike heat, is not only rigidly designated by 'pain' but the reference of the designator is determined by an essential property of the referent.[72]

Thus on Kripke's view A=B is a proposition that contains natural kind terms that secures reference by specifying not an accidental but *essential* property of its referent. For this reason, A=B doesn't satisfy Condition 0, and so if Condition 0 is the only restriction on CEP's scope, then A=B does not belong in R. Thus Kripke is confronted by a dilemma:

[71] Kripke (1980). pp. 128-142. According to Kripke, since the apparent contingency can be explained away, water might be qualitative identical to substance other than H_2O, say 'XYZ', that is not to say the qualitative identity is equivalent to identity. Thus, Water is necessarily H_2O if water is H_2O.

[72] Ibid. Pp. 152-3.

should he choose to reject Argument MK by claiming RCEP, his own premise that pain is picked out non-accidental property, would be undermined; should he retain his premise, then *position 3* fails to prevail.

Of course, it is logically possible to come up with another specification of *R* that A=B fits, since Condition 0 might not be the only restriction condition. *Position 3* might have a chance to succeed if Kripke can describe another restriction condition, that excludes A=B but not A≠B from CEP. This would be a major undertaking, and Kripke hasn't even begun to provide us with the details. The ball is in Kripke's court.

To sum up, three possible replies to Argument MK have been examined. *Position 1* is to reject MK4 with a sound argument, and that argument would itself be a sufficient proof against identity theory. Perhaps Kripke and his defenders can produce such argument, but until they have done so Argument K's soundness should at least be suspended and Argument MK stands.

Position 2 is to contend that A=B is less conceivable than A≠B. Anyone taking this position would need to explain what it means to say that one proposition is more, or less, conceivable than another. In so doing, they also face the challenge of justifying a metaphysical assumption that their position relies on.

Position 3 is to argue that A=B is conceivable but impossible. This requires restricting the CEP principle to exclude A=B. The problem with this approach is that Kripke has not himself described any way of describing the principle that does this, and were a defender of Kripke to come such new principle they would also need to show that it does not violate Kripke's crucial premise that 'whatever feels like pain is pain'.

In addition to the three option just discussed, there is a fourth option - *position 4*. What Kripke and his defenders might do is to accept that Argument K and Argument MK are both equally powerful. This is the option that I endorse. In Arguments K and MK, we have two conceivability arguments of the same form, and with premises that yield opposite conclusions. The pair of arguments therefore generates a paradox despite the individual coherency in each argument. Both arguments lose their force as a result, and they are mutually defeating. Together they show that *something has gone wrong*; that at least one of the arguments has a false premise but they don't tell us which premise is faulty. It must be stressed that this position—*Position 4* as we might call it—is *not* unpalatable for the mirroring camp. My position is not that Kripke's premises are false or that his conclusion is false. Rather, by putting forward this strategy, the aim is to expose a serious weakness in the form of the argument that Kripke uses, namely, the conceivability argument. Before I move on to give the same diagnosis to Chalmer's zombie argument, I shall conclude that self-sacrifice, metaphysical challenge, restricting CEP, and mutual defeat are the four horns that make up what I call Kripke's *quadlemma.*

Chapter 3. The ‘zombie’ argument against materialism

3. 0. Outline of Chapter 3

The previous Chapter examined Kripke’s modal argument against identity theory, namely Argument K. I offered an argument that Argument K is mirrorable. That is, we can retain all but one assumptions of Argument K and derive a conclusion directly opposite to it. I discussed that Kripke and his defenders can reject the mirroring attack only by repudiating the mirroring assumption, however, should they reply in this way they would be confronted by, as I put it, Kripke’s *quadlemma* – four untenable options. The purpose of the present Chapter is to apply the same strategy to criticise Chalmers’ ‘zombie’ argument.

There are three sections in this Chapter. 3.1 will summarise the Chalmers’ argument. 3.2 and 3.3 will recall the mirroring strategy and describe how it undermines the ‘zombie’ argument. Unlike what we have seen in the Kripkean case, to yield an opposite conclusion to Chalmers’ argument requires a slightly different technique. I will present three mirroring arguments to elaborate on this distinction. In 3.3, I will also explain the significance of the so-called ‘That’s all’-clause in Chalmers’ definition of materialism. I will show how it makes the third mirroring argument hard to refute.

3. 1. Chalmers’ argument

David Chalmers’ stance on the mind-body debate and problems concerning consciousness in general is originally expressed in his 1996 book *The Conscious Mind*.[73] At large, Chalmers holds that consciousness cannot be explained in physical terms, and materialism therefore fails to be a correct theory. Before I proceed any further, I must make one important clarification regarding the dialectics presented in my dissertation so far. There seems to be a mismatch between Chalmers’ target and the theory I’m trying to

[73] Chalmers, D. (1996), *The Conscious Mind*.

defend in this dissertation. The present work of mine is a defence of identity theory, as I claimed. My task is to rebut arguments against identity theory. On the other hand, Chalmers' *does not explicitly target* identity theory. Rather, he aims at the materialist theory. By revealing what Chalmers takes materialism to be, I hope the apparent mismatch can be explained away.

Materialism, in general, amounts to an ontological thesis about all things in the world. It holds that all things are ultimately physical or made up of physical parts. As a result, according to the materialists' picture, the basic constituents of reality are fixed by our best scientific theory, and given that our best theory deals in basic physical elements, materialism holds that at root, the universe is physical. Resembling Braddon-Mitchell and Jackson's approach,[74] Chalmers defines materialism in terms of possible worlds:[75]

> Definition. A minimal physical duplicate of the world is a duplicate simpliciter.

Phenomenal facts are part of the universe. Thus, a materialist theory of mind holds the following:

> Definition. A minimal physical duplicate of the world is a psychological duplicate.

Thus, for any world, if all physical facts obtain and nothing else obtains in that world, phenomenal facts must obtain in that world. We can formalise this as C1 below, where P denotes all physical facts about the universe, T denotes a 'nothing else, that's all clause', and Q denotes phenomenal facts. The T-clause is a very important element in Chalmers' argument that requires special attention. In 3.2.2, I will explain how the involvement of this T-clause allows Chalmers to formalise materialism and allows me to resurrect the mirroring attack against Chalmers. For now, Chalmers' first premise can be put as follows:

C1. Materialism $\leftrightarrow \Box(PT \rightarrow Q)$

In this formalisation, the central tenet of materialism is a necessary conditional statement. On the other hand, mind-brain identity theory states that mental states are brain states,

[74] Braddon-Mitchell & Jackson (2007), pp. 28-30. The definition evolves from Jackson's (1994).

[75] Chalmers' exact formalisation is 'materialism is true if for any logically possible world *W* that is physically indiscernible from our world, all the positive facts true of our world are true of *W*' (1996, p. 42). As confirmed by Chalmers, this matches Braddon-Mitchell's and Jackson's definition stated above.

hence, $M=B$. In other words, the central tenet of identity theory is an identity statement. What makes Chalmers' rejection of materialism—a necessary conditional statement—relevant to my rectification of the identity statement? The answer is simple in terms of logic. Rejecting materialism means negating C1, which can be represented by C6:

C6. $\neg\Box(PT \rightarrow Q)$

Adding a S5 assumption in the form of C5:

C5. $\Diamond(PT \,\&\, \neg Q) \rightarrow \neg\Box(PT \rightarrow Q)$

C6 can be entailed by C4:

C4. $\Diamond(PT \,\&\, \neg Q)$

C4 says it is possible for all physical facts obtain and nothing else obtains, *and no phenomenal facts obtains*. Since facts about brain states are physical facts, and facts about mental states are phenomenal facts, the possibility of C4 being true thus allows the possibility of $B\&\neg M$, which contradicts $M=B$. In short, rejecting materialism suffices to reject identity theory. On the flip side, by attacking Chalmers' argument against materialism, I am effectively defending identity theory.

In clarifying the relevance of the present Chapter to the overall aim of my dissertation, I have already shown four crucial lines of Chalmers' argument in C1, C4, C5 and C6. Together, they seem to generate the denial of materialism:

C1. Materialism $\leftrightarrow \Box(PT \rightarrow Q)$ Assumption

…

C4. $\Diamond(PT \,\&\, \neg Q)$?

C5. $\Diamond(PT \,\&\, \neg Q) \rightarrow \neg\Box(PT \rightarrow Q)$ S5

C6. $\neg\Box(PT \rightarrow Q)$ 4, 5, *modus ponens*

C7. Materialism is incorrect. 1, 6, *modus tollens*

However, as they currently stand, they do not form a valid argument. The precise reason is the lack of justification for C4. C4 cannot be simply assumed, for that would render the reasoning circular: since materialism is defined in terms of $\Box(PT\rightarrow Q)$, one mustn't assume its negation, namely, $\Diamond(PT\&\neg Q)$, to yield the denial of materialism (barring the possible denial of S5). Consequently, in order to form a valid argument against materialism, there must be premises in between C1 and C4, and these premises will need to entail C4. In what follows, I will gradually introduce and explain these currently missing lines in Chalmers' argument.

Chalmers' whole argument starts with a description of the notion of *supervenience*. Supervenience, generally speaking, is a relation between two sets of facts. If 'one set of facts fully determine another set of facts',[76] then the latter set is said to be supervening on the former set. In elaboration, Chalmers claims that supervenience is better formalised in terms of properties: *B-properties*, the high-level properties, supervene on *A-properties*, the physical properties, if 'no two possible situations are identical with respect to their A-properties while differing in their B-properties'.[77] He then introduces the idea of *logical supervenience*, which he defines in the following way:

> B-properties supervene *logically* on A-properties if no two *logically possible* situations are identical with respect to their A-properties but distinct with respect to their B-properties.[78]

The crucial addition here is 'logically possible', which Chalmers describes as below:

> One can think of it loosely as possibility in the broadest sense, corresponding roughly to conceivability, quite unconstrained by the laws of our world. It is useful to think of a logically possible world as a world that it would have been in God's power (hypothetically!) to create if he had so chosen. God could not have created a world with male vixens, but he could have created a world with flying telephones. In determining whether it is logically possible that some statement is true, the constraints are largely conceptual. The notion of a male vixen is contradictory, so a male vixen is logically impossible; the notion of a flying telephone is conceptually coherent, if a little out of the ordinary, so a flying telephone is logically possible.[79]

The mention of conceptual constraints being the only constraints and the example of male vixens clearly suggest that Chalmers' notion of logical possibility is no different to the conventional understanding of it. Logical supervenience is therefore a modal variant of supervenience, and we can state the notion by adding '$\neg\Diamond$' in front of the definition of supervenience, which in turn yields:

$$\Box(\text{A-properties} \rightarrow \text{B-properties})$$

[76] Chalmers (1996), p. 32.

[77] Ibid. P. 33.

[78] Ibid. P. 35. Author's italics. Chalmers also distinguishes between *local* and logical supervenience. Only the latter notion is relevant to the zombie argument.

[79] Ibid. P. 35.

This is one definition of supervenience. Moreover, since A-properties, as Chalmers announces, are 'the fundamental properties that are invoked by a completed theory of physics' [80] we can substitute 'A-properties' in the above definition with the aforementioned 'P', which denotes all physical facts:

$$\Box(P \rightarrow \text{B-properties})$$

In other words, given all the physical facts, B-properties must occur, in all possible worlds.

With the help of the notion of logical supervenience, materialism can be defined in terms of logical supervenience between physical facts and phenomenal fact. Up to this point, 'B-properties' is understood as a generic term. Swapping it with the aforementioned 'Q', we get a definition of materialism:

$$\Box(P \rightarrow Q)$$

However, there is one small problem with this definition. Negative existential facts such as 'There are no unicorns' do not logically supervene on physical facts:

$$\neg\Box(P \rightarrow \text{there are no unicorns})$$

But the fact that there are no unicorns does not post threat to the truth of materialism. They are, in fact, compatible. To tweak this problem coming from negative existentials, Chalmers adds that there needs to be a 'That's all' clause on P, namely, the aforementioned 'T'.[81] This completes Chalmers' definition of materialism,[82] which is C1 as aforementioned:

C1. Materialism $\leftrightarrow \Box(PT \rightarrow Q)$

As I explained, denying materialism is thus denying $\Box(PT \rightarrow Q)$, which in turn amounts to establishing $\Diamond(PT \,\&\, \neg Q)$ (i.e. C4).What might be an antecedent to C4? The statement asserts that *it is possible that PT & ¬Q*, or in other words, *it is possible that for any world, if all physical facts obtain and nothing else obtains in that world, phenomenal facts must obtain in that world.* Putting aside the ◇-operator for a moment, let us first look at what, according to Chalmers, is an example of PT&¬Q:

[80] Chalmers (1996), p. 33.

[81] Ibid. p. 41.

[82] Kim (2011, pp. 8-10) also uses this logical supervenience-definition. He calls it *strong supervenience* instead.

> So let us consider my zombie twin. This creature is molecule for molecule identical to me, and identical in all the low-level properties postulated by a completed physics, but he lacks conscious experience entirely.[83]

Zombie, thus, is a creature that is physically identical to us that nonetheless lack phenomenal consciousness. Chalmers' Zombie is indeed one way to describe $PT\&\neg Q$. However, the mere description of one thing does not entail the possibility of it. To fill this inference gap, Chalmers argues:

> I confess that the logical possibility of zombies seems equally obvious to me. A zombie is just something physically identical to me, but which has no conscious experience—all is dark inside. While this is probably empirically impossible, it certainly seems that a coherent situation is described; I can discern no contradiction in the description. In some ways an assertion of this logical possibility comes down to a brute intuition, but no more so than with the unicycle. Almost everybody, it seems to me, is capable of conceiving of this possibility. Some may be led to deny the possibility in order to make some theory come out right, but the justification of such theories should ride on the question of possibility, rather than the other way around.[84]

Three messages are lucidly conveyed in this passage. First, Chalmers concedes that zombie is 'probably empirically impossible'. In my view, this is Chalmers' way of saying that zombies do not exist in the actual world, and subsequently the truth-value of $PT\&\neg Q$ is false. Second, the truth of $PT\&\neg Q$ does not matter to the argument against materialism, so long as zombies are logically possible, hence, $\Diamond(PT\&\neg Q)$ is true. Third, zombies are logically possible because what seems consistent and coherent is logically possible and nothing in the zombie case seems contradictory or incoherent.

Thus, these words of Chalmers' readily transform into the missing steps in his arguments. They are C2 and C3:

<u>Argument C</u>

(P = all microphysical facts about the universe, T = 'That's all', Q = qualia)

C1. Materialism $\leftrightarrow \Box(PT \rightarrow Q)$ Assumption

C2. $\Diamond_c(PT \,\&\, \neg Q)$ Assumption

C3. $\Diamond_c\Phi \rightarrow \Diamond\Phi$ Assumption (CEP)

[83] Chalmers (1996), p. 94.

[84] Ibid. P. 96.

C4. $\Diamond(PT \mathbin{\&} \neg Q)$ 2, 3, *modus ponens*

C5. $\Diamond(PT \mathbin{\&} \neg Q) \rightarrow \neg \Box(PT \rightarrow Q)$ S5

C6. $\neg \Box(PT \rightarrow Q)$ 4, 5, *modus ponens*

C7. Materialism is incorrect. 1, 6, *modus tollens*

With all premises in place, we have a complete reconstruction of Chalmers' argument against materialism in Argument C. It requires little effort to recognise that Argument C and Argument K have the same form. To be precise, they are both conceivability arguments with a conceivability premise and a conditional premise in CEP.

3. 2. Mirroring Chalmers – round one

Criticisms of the above argument have predominantly targeted the argument's premises. In my opinion, one need not argue for the falsity of the premises to show that the argument is toothless. Instead, we can attack its form. As established, Argument C is a version of conceivability argument with a conceivability premise (C2) and a CEP premise (C3). If we can retain all premises except the conceivability premise, and yield a conclusion opposite to Chalmers' conclusion, then something is wrong. If we can also show that conceivability premises in both arguments do not contradict each other, then the flaw lies in the form of both arguments. This is the mirroring objection to Chalmers. In what follows, I will present two ways of replacing C2 in Argument C to deduce the conclusion that materialism is correct.

3. 2. 1. Argument RC

To begin with, recall Chalmers' first premise:

C1. Materialism $\leftrightarrow \Box(PT \rightarrow Q)$

C1 is a definition of materialism using the notion of logical supervenience. If this definition is good for Chalmers, it's good for the materialists. So the first premise of the mirrored argument, RC1, is identical to C1:

RC1. Materialism $\leftrightarrow \Box(PT \rightarrow Q)$

The same applies to C3 and its mirrored counterpart RC3:

RC3. $\Diamond_c\Phi \rightarrow \Diamond\Phi$

Hence, both arguments share CEP as their crucial conditional premise. The question of whether conceivability entails possibility has invited most discussions on the 'zombie' argument. One of the upshots of the mirroring attack on conceivability arguments, in my view, is the convenience that we don't need to enter this muddy battleground about the truth of CEP. In saying that, including CEP in the mirrored argument does not suggest that I endorse this principle. Rather, the strategy is that Chalmers must not reject RC3 should he wants to preserve his C3.

The only change to be made is on C2:

C2. $\Diamond_c(PT \ \& \ \neg Q)$

To mirror it, we might repeat the technique we use to mirror Kripke, that is, to assert the conceivability of the negation of PT & ¬Q. Thus, we might try RC2:

RC2. $\Diamond_c(PT \rightarrow Q)$

Note that RC2 and C2 do not contradict each other, despite the fact that the contents of conceivability expressed in these two premises contradict each other. This is because conceivability is not a distributive property. From the absolute falsity of $\Diamond_c(S \ \& \ \neg \ S)$, one cannot infer $\Diamond_c S$ and $\Diamond_c \neg S$ are mutually exclusive.[85]

We now have the first three lines of the first mirrored argument:

RC1. Materialism $\leftrightarrow \Box(PT \rightarrow Q)$

RC2. $\Diamond_c(PT \rightarrow Q)$

RC3. $\Diamond_c \Phi \rightarrow \Diamond \Phi$

Next, from *modus ponens*, we can derive $\Diamond(PT \rightarrow Q)$ from RC2 and RC3. Thus, RC4:

RC4. $\Diamond(PT \rightarrow Q)$

With these four lines in hand, we seem to have compiled a mirrored argument against Argument C, namely, Argument RC:

Argument RC

(P = all microphysical facts about the universe, T = 'That's all', Q = qualia)

RC1. Materialism $\leftrightarrow \Box(PT \rightarrow Q)$ Assumption

[85] I have explained this point in detail in 2.3.2 when I discussed why MK4 is not mutually exclusive to Kripke's K4.

RC2. $\Diamond_c(PT \rightarrow Q)$ Assumption

RC3. $\Diamond_c\Phi \rightarrow \Diamond\Phi$ Assumption (CEP)

RC4. $\Diamond(PT \rightarrow Q)$ 2, 3, *modus ponens*

Argument RC, in this presentation, ends with $\Diamond(PT \rightarrow Q)$ as its conclusion. However, this is too weak to be a necessary and sufficient condition for materialism. Also, it is compatible with Chalmers sub-conclusion C6:

C6. $\neg\Box(PT \rightarrow Q)$

which yields his conclusion C7 from C1 via *modus tollens*:

C7. Materialism is incorrect.

Therefore, the mirrored argument cannot end at RC4. In order to establish a conclusion that negates Chalmers' conclusion—materialism is incorrect—Argument RC needs to claim more. It needs a conclusion that states $\Box(PT \rightarrow Q)$:

Argument RC

(P = all microphysical facts about the universe, T = 'That's all', Q = qualia)

RC1. Materialism $\leftrightarrow \Box(PT \rightarrow Q)$ Assumption

RC2. $\Diamond_c(PT \rightarrow Q)$ Assumption

RC3. $\Diamond_c\Phi \rightarrow \Diamond\Phi$ Assumption (CEP)

RC4. $\Diamond(PT \rightarrow Q)$ 2, 3, *modus ponens*

RC5. $\Box(PT \rightarrow Q)$?

This way, Argument RC seems to satisfy the purpose of mirroring Chalmers. It has the same form, retains all premises except RC2, and RC2 is not a negation of C2. More importantly, it has a conclusion that negates Chalmers' conclusion. Triumph! It seems. Unfortunately, a fatal problem occurs at the step from RC4 to RC5. For the inference to go through, an obscure modal logic rule is required. To be precise, it relies on the plausibility of $\Diamond S \rightarrow \Box S$, which is too bizarre to be accepted as an inference rule just for the purpose of mirroring Chalmers. In short, unless this extraordinary modal claim can be justified, Argument RC collapses.

This failure shows that to derive $\Box(PT \rightarrow Q)$, the mirroring camp has to find another antecedent to replace $\Diamond(PT \rightarrow Q)$. Moreover, this problem is rooted in RC2, the postulation of the conceivability of $\Diamond(PT \rightarrow Q)$, because it yields $\Diamond(PT \rightarrow Q)$ via CEP.

Thus, the message is clear: what needs to go is RC2. The old mirroring technique that works against Kripke doesn't work against Chalmers since we can't simply establish the truth of materialism by conceiving the negation of zombies. The mirroring camp need to go back to the drawing board and think of a proposition *S* such that the conceivability of *S* can entail $\Box(PT \rightarrow Q)$ via CEP.

3. 2. 2. Argument RC* - double modality

How about the conceivability of $\Box(PT \rightarrow Q)$? Perhaps the conceivability of materialism can be the mirroring conceivability premise. Hence, we might be able to mirror the zombie argument with a simple five-lines *modus ponens*:

Argument RC*[86]

(P = all microphysical facts about the universe, T = 'That's all', Q = qualia)

RC1*. Materialism $\leftrightarrow \Box(PT \rightarrow Q)$ Assumption

RC2*. $\Diamond_c \Box(PT \rightarrow Q)$ Assumption

RC3*. $\Diamond_c \Phi \rightarrow \Diamond \Phi$ Assumption (CEP)

RC4*. $\Diamond \Box(PT \rightarrow Q)$ 2, 3, *modus ponens*

RC5*. Materialism is correct. 1, 4, *modus ponens*

Noticeably, this mirrored argument has a distinctive feature, namely, there is double modality in it. Specifically, RC4* is in the form of $\Diamond\Box S$. This, from modal logic's point of view, does not post any threat to the validity of Argument RC*. In S5, the '$\Box$'-operator trumps the '$\Diamond$'-operator. From $\Diamond\Box(PT \rightarrow Q)$, it follows that $\Box(PT \rightarrow Q)$. However, this way of mirroring seems to be giving away the symmetrical relation between Argument C and its mirrored argument, because there is no double modality in Argument C. Nevertheless, if we can reformulate Argument C so that it contains double modality, then the symmetry can be restored. The following suggests a doubly-modal variant of Argument C:

[86] This version of the mirrored argument resembles what Chalmers calls *the conceivability of materialism* reply to the zombie argument (2010, p. 180). However, Argument presented here differs from the mirrored arguments proposed by Marton (1998), Frankish (2007), and Brown (2010), on the basis that Argument RC is doubly modal and in turn it commits to a doubly-modal interpretation of Argument C. Among those cited by Chalmers, Sturgeon (2000) is the only one that endorses the double-modal interpretation.

Argument C*

(P = all microphysical facts about the universe, T = 'That's all', Q = qualia)

C1*. Materialism $\leftrightarrow \Box(PT \rightarrow Q)$ Assumption

C2*. $\Diamond_c\Diamond(PT \,\&\neg Q)$ Assumption

C3*. $\Diamond_c\Phi \rightarrow \Diamond\Phi$ Assumption (CEP)

C4*. $\Diamond\Diamond(PT \,\&\neg Q)$ 2, 3, *modus ponens*

C5*. $\Diamond\Diamond(PT \,\&\neg Q) \rightarrow \neg\Box(PT \rightarrow Q)$ S5

C6*. $\neg\Box(PT \rightarrow Q)$ 4, 5, *modus ponens*

C7*. Materialism is incorrect. 1, 6, *modus tollens*

Just as in Argument RC*, the double modality post no threat to Argument C*'s validity. Formalised in this way, the pair of arguments reenact the situation we have seen in Argument K and MK.

First of all, just as Argument MK does not wish to show the truth of identity theory, Argument RC* is not an attempt to prove materialism. Instead, the blueprint here is what I call a 'partner's in crime' strategy. It is to show that we have no better grounds for judging Argument RC* to be unsound than we have for judging Argument C* to be unsound. The reasoning proceeds as follows:

Chalmers: The conceivability of $\Diamond(PT\&\neg Q)$ entails the possibility of $\Diamond(PT\&\neg Q)$. Materialism is incompatible with $(PT \,\&\neg Q)$, therefore, materialism is false.

Mirroring reply: The conceivability of $\Box(PT\rightarrow Q)$ leads to the possibility of $\Box(PT\rightarrow Q)$ which in turn by S5 leads to $\Box(PT\rightarrow Q)$, therefore materialism is true.

The two sides share the same form of argumentation. To be precise, the falsity of a theory *T* is derived from the incompatibility between *T* and a conceivable state of affair.

1. If *S* is possible, then *T* is false.
2. It is conceivable that *S*.
3. The conceivability of *S* entails the possibility of *S*.
4. Therefore, *T* is false.

In Argument C*, $S = \Diamond(PT\&\neg Q)$, and in Argument RC*, $S = \Box(PT\rightarrow Q)$. The contents of conceivability are two contradictory states of affairs, because $\Diamond(PT\&\neg Q)$ is the negation of $\Box(PT\rightarrow Q)$. But as I explained, this is not to say that one of them is necessarily false. From *a priori* inspection, we can not infer that neither $\Diamond(PT\&\neg Q)$ nor $\Box(PT\rightarrow Q)$ is conceivable from the fact that $\Diamond(PT\&\neg Q)$ & $\Box(PT\rightarrow Q)$ is not conceivable.

For this reason, the second premises in both arguments *can* be jointly true. This point is reinforced by Scott Sturgeon's words:

> It's coherent to suppose zombies are genuinely possible. It's coherent to suppose zombies are not genuinely possible. A full grasp of [$\Diamond(PT\&\neg Q)$] reveals nothing to preclude [its] truth. A full grasp of its negation reveals nothing to preclude truth.[87]

This creates a serious problem for both arguments. The form of these arguments, namely, the argument from conceivability, seems to work well for rival theories. each argument only differs from the other in regards to the second premises – one stresses that we can conceive of situation *S*, the other asserts that we can conceive of situation $\neg S$. Since, contrary to prima facie considerations, this is not an inconsistent pair of statements, it follows that the second premises in both arguments need not have opposite truth-values. However, this consequence is alarming, since the two arguments aim to establish the truth of two mutually contradictory propositions. Sturgeon observes this point and concludes that conceivability arguments yield what he calls *symmetric defeat.*[88] We have *a priori* reason, namely, our ability to conceive of $\Diamond(PT\&\neg Q)$, for accepting $\Diamond(PT\&\neg Q)$. At the same time we also have *a priori* reason, namely, our ability to conceive of $\Box(PT\rightarrow Q)$, for accepting $\Box(PT\rightarrow Q)$. The two conceivability premises therefore generate a paradox, despite their individual coherency. Both arguments for and against physicalism, therefore, lose their force.

In general, how can a pair of valid arguments with contradictory conclusions be plausible at the same time? We have reason to accept the truth of each premise. As things stand now, our capacities to get epistemically engaged with both propositions are equally strong, or equally weak. In other words, we have equal weights on each side of this conceivability scale. The good thing is that in the current situation Argument C* and Argument RC* are equally plausible. The bad news is, however, neither of them has adequate force, because the opposite side has just the same level of credibility. They

[87] Sturgeon (2000), p. 115.

[88] Ibid. P. 116.

cancel out. In short, if the equilibrium of epistemic possibilities of $\Diamond(PT\&\neg Q)$ and $\Box(PT\rightarrow Q)$ obtains, then both arguments cease to attain any kinds of strength, and are therefore, toothless. For this reason, the fourth horn of Kripke's *quadlemma*, namely, mutual defeat, haunts Chalmers too.

Alternatively, should Chalmers try to show that one of the two arguments attain enough force to demonstrate the truth of its conclusion? In other words, he might respond that Argument C* is a compelling sound argument. But that would leave an even more unpleasant outcome for him and his defenders. Suppose Argument C* is not without force, hence, materialism really is incorrect, then Argument RC*, by virtue of definition, is unsound, because no sound argument has false conclusion.[89] However, the failure of Argument RC* does not really need to bother us here. Rather, we should consider Argument C*'s status – that of being a sound and compelling argument. What does this imply? By the definition of soundness, it implies that C2*'s truth value must be confirmed. Our reason for accepting C2*'s truth is, still, our capacity to get epistemically engaged with $\Diamond(PT\&\neg Q)$. But now we cannot say the same for the other side, specifically, we cannot hold that $\Box(PT\rightarrow Q)$. At this point, the conceivability scale is tipped.

The trouble for Chalmers starts here. To say that the conceivability scale is tilted and RC2* is false, is to say that $\Box(P\rightarrow Q)$ is inconceivable, which is equivalent to the following:

$$\neg\Diamond_c\Box(PT\rightarrow Q)$$

From IEI (inconceivability entails impossibility), the logical impossibility of $\neg\Diamond_c\Box(PT\rightarrow Q)$ follows:

$$\neg\Diamond\Box(PT\rightarrow Q)$$

which by S5, entails:

$$\Diamond(PT\&\neg Q)$$

$\Diamond(PT\&\neg Q)$ is Argument C*'s conclusion. Thus, if Argument C is sound, the presence of C2* means the denial of its mirrored counterpart RC2*. But this just amounts to another way of expressing Argument C*'s conclusion. In other words, if one wants to show

[89] This example goes both ways, you can suppose the opposite scenario where C* is unsound and R* is sound and powerful.

Argument C* is sound, she must deny that the second premises in both arguments yield mutual defeat. What in turn constitutes the problem of mutual defeat is the equilibrium of epistemic possibilities of $\Diamond(PT\&\neg Q)$ and $\Box(PT\rightarrow Q)$. This means that she needs to reject $\Diamond_c\Diamond(PT\&\neg Q)$ & $\Diamond_c\Box(PT\rightarrow Q)$. She must claim that the pair is an exclusive disjunction, and yields single defeat as a result. Therefore, she needs to claim the following proposition:

$$(\Diamond_c\Diamond(PT\&\neg Q) \vee \Diamond_c\Box(PT\rightarrow Q)) \& \neg(\Diamond_c\Diamond(PT\&\neg Q) \& \Diamond_c\Box(PT\rightarrow Q))$$

If so, then the mere postulation of the left-hand side of the disjunction, $\Diamond_c\Diamond(PT\&\neg Q)$, is trivially a denial of the right-hand side disjunct, $\Diamond_c\Box(PT\rightarrow Q)$. This amounts to what can be called CZ:

CZ3. $\Diamond_c\Diamond(PT\&\neg Q) \rightarrow \neg\Diamond_c\Box(PT\rightarrow Q)$

Chalmers and his defenders can dodge the problem of mutual defeat by making the case for CZ3. But this way of avoiding the 'partner's in crime' situation comes at a hefty price. As shown, they need to claim 'the conceivability of $\Diamond(PT\&\neg Q)$ entails the inconceivability of $\Box(PT\rightarrow Q)$', and thus has to assert 'it is not the case that we can conceive of $\Diamond(PT\&\neg Q)$ and we can conceive of $\Box(PT\rightarrow Q)$, albeit no obvious contradiction in $\Diamond_c\Diamond(PT\&\neg Q) \& \neg\Diamond_c\Box(PT\rightarrow Q)$'.

Furthermore, claiming CZ3 brings out the first horn of Kripke's *quadlemma*, namely, the problem of self-sacrifice. Suppose CZ3 is true, proponents of Argument C* can use it as a premise to construct another argument against materialism:

Argument CZ

(P = all microphysical facts about the universe, T = 'That's all', Q = qualia)

CZ1. Materialism $\leftrightarrow \Box(PT\rightarrow Q)$	Assumption
CZ2. $\Diamond_c\Diamond(PT\&\neg Q)$	Assumption
CZ3. $\Diamond_c\Diamond(PT\&\neg Q) \rightarrow \neg\Diamond_c\Box(PT\rightarrow Q)$	Assumption
CZ4. $\neg\Diamond_c\Phi \rightarrow \neg\Diamond\Phi$	Assumption (IEI)
CZ5. $\neg\Diamond\Box(PT\rightarrow Q)$	2, 3, 4, *modus ponens*
CZ6. $\neg\Box(PT\rightarrow Q)$	5, S5
CZ7. ¬Materialism	1, 6, *modus tollens*

From Chalmers' standpoint, Argument CZ will have two advantages over Argument C*. First, by having CZ3 as a premise it will avoid the 'partner's in crime' problem. Argument CZ is explicit in attempting to make the apparent conceivability of RC2* fade away, whereas Argument C* does not make this move. Second, as I suggested in 2.3.2, IEI is a far less controversial claim than CEP. For this reason, proponents of Argument CZ are on a safer ground than that of Argument C*.

Can we ring the victory bell for Argument CZ, based on these two upshots? If they can show that CZ3 is true, then the answer is yes. In fact, any argument that has the following logical form will be a better and simpler argument against materialism than the original zombie argument:

Argument CZG

1. Ψ
2. $\Psi \rightarrow \neg\Diamond_c\Box(PT \rightarrow Q)$

----

1. $\neg\Diamond_c\Box(PT \rightarrow Q)$
2. IEI

----

3. $\neg\Box(PT \rightarrow Q)$

Here, Ψ stands for whatever a good justification for $\neg\Diamond_c\Box(PT \rightarrow Q)$ might be. If foes of materialism find that there really is a good justification, then they can claim that materialism is false by Argument CZG. Therefore rejecting RC2* in the mirror argument renders Argument C* self-sacrificial. More importantly, the ball is put to Chalmers' court, and it is up to him to show us the detail of Argument CZG. Before he has done so, the soundness of Argument C* is suspended.

On the other hand, if they want to save Argument C* from rsuspension, then the epistemic equilibrium between C2* and RC2* is upheld, and the two arguments are mutual defeating. In short, the mirror argument creates a dilemma for Chalmers.

However, this mirroring objection depends largely on the assumption that Argument C* is a correct interpretation of Chalmers. One could say that the original zombie argument does not contain double modality, and it follows that the Argument RC* is mirroring a straw man argument. Moreover, since it has been conceded that Argument

RC fails to mirror Argument C—the standard interpretation of the zombie argument, the mirroring strategy fails to work against Chalmers.

3. 3. Mirroring Chalmers – round two

In reply to the above, whether or not the doubly-modal interpretation is correct is an issue that I do not need to pursue here, so long as I can think of alternative way to mirror the standard interpretation. This brings us to round two of mirroring Argument C.

3. 3. 1. Argument MC

A lesson learnt from round one is that the mirroring camp need to think of a proposition *S* such that the conceivability of *S* can entail $\Box(PT\rightarrow Q)$ via CEP. This will be a two-step inference. First, from the conceivability of *S* and CEP, the possibility of *S* is entailed. Second, the argument need to show that the possibility of *S* entails $\Box(PT\rightarrow Q)$. This prospective mirroring argument will have the following form:

1. Materialism $\leftrightarrow \Box(PT \rightarrow Q)$
2. $\Diamond_c S$
3. $\Diamond_c \Phi \rightarrow \Diamond \Phi$
4. $\Diamond S$

----

5. $\Box(PT \rightarrow Q)$

In searching for the suitable candidate for *S*, three conditions need to be met:

Condition 1. *S* must not contain logical contradiction, for that would prevent the conceivability of *S*.

Condition 2. *S* needs to be mirroring the idea of zombies, namely, $PT\&\neg Q$.

Condition 3. $\Diamond S$ needs to be a sufficient condition for $\Box(PT\rightarrow Q)$.

The first condition is relatively easy to be complied with. The real hurdle is the other two conditions. Starting with Condition 2, what exactly does it mean to 'mirror' the idea of zombies? To find out, let us recall what we have mirrored so far. In the Kripkean case, the original Kripkean conceivability premise and its mirroring counterpart are K4 and MK4:

K4. $\Diamond_c(A=B)$

MK4. $\Diamond_c(A \neq B)$

In mirroring Chalmers, we have postulated RC2 to mirror C2:

C2. $\Diamond_c(PT \; \& \; \neg Q)$

RC2. $\Diamond_c(PT \rightarrow Q)$

And also RC2* to mirror C2*:

C2*. $\Diamond_c\Box(PT \; \& \; \neg Q)$

RC2*. $\Diamond_c\Box(PT \rightarrow Q)$

From these three mirrored-pairs, a pattern is visible. The content of the mirroring conceivability is the negation of the content of the original conceivability. If we maintain this approach when we mirror Argument C, we need *S* to be PT→Q, since it negates PT&¬Q. But this has been proven wrong. *S* cannot be PT→Q because it fails Condition 3. To be precise, $\Diamond(PT \rightarrow Q)$ does not entail $\Box(PT \rightarrow Q)$, for already established reason. Therefore, we shall perhaps change the mirroring pattern. Maybe we don't need to postulate the negation of P&¬Q, in order to mirror it.

Returning to the big picture will give us a clearer idea. Why does Chalmers postulate PT&¬Q in Argument C? The answer is that Argument C tries to show that a minimal duplicate of our world *need not* be a phenomenal duplicate of it by asking us to conceive of a minimal physical duplicate of our world *that is not* a phenomenal duplicate of it. It follows that to mirror Argument C, the strategy should be to conceive of a minimal physical duplicate of our world *that is* a phenomenal duplicate of it. What would such a world be like? It would be a world in which PT and Q are both true, i.e. PT&Q is true.

Thus, PT&Q meets Condition 2. It takes little effort to recognise that it also meets Condition 1. The remaining question is whether it meets Condition 3. Let us use it to construct the mirror argument and investigate from there:

Argument MC

(P = all microphysical facts about the universe, T = 'That's all', Q = qualia)

MC1. Materialism $\leftrightarrow \Box(PT \rightarrow Q)$ Assumption

MC2. $\Diamond_c(PT \; \& \; Q)$ Assumption

MC3. $\Diamond_c\Phi \rightarrow \Diamond\Phi$ Assumption (CEP)

MC4. $\Diamond(PT \; \& \; Q)$ 2, 3, *modus ponens*

MC5. $\Box(PT \rightarrow Q)$?

MC6. Materialism. 1, 5, *modus ponens*

The argument retains Argument C's logical form. From the conceivability of PT&Q, the possibility of PT&Q is entailed via CEP. If we can justify that $\Diamond(PT \,\&\, Q)$ entails $\Box(PT \rightarrow Q)$, then Chalmers' conclusion will be mirrored. So is $\Diamond(PT\&Q) \rightarrow \Box(PT\rightarrow Q)$ a true proposition? The answer is yes, and the reason resides in the occurrence of 'T' – the 'That's all'-clause. To elucidate this point, we need a proper understanding of 'T".

3. 3. 2. 'That's all!'

As I briefly mentioned in 3.1.1, the inclusion of the 'That's all'-clause (T for short, henceforward) in Argument C and its mirrored version Argument MC arises from Chalmers' definition of materialism. To define materialism, one might suggest the following definition:

(DP) Any world that is a physical duplicate of our world is a psychological duplicate of our world.

The trouble is that DP does not exclude possible worlds that are physically identical to ours but contain extra non-physical phenomenology. In those possible worlds, the mental does not supervene on the physical, because physical phenomenology is not present. Thus DP fails to establish logical supervenience as Chalmers wishes. Another way to highlight DP's inadequacy of defining materialism is to treat the phenomenal facts as negative facts. As mentioned, this is noted by Chalmers. In short, the trouble with DP is that it misrepresents the materialists' claim as more wide-ranging than it in fact is. What we need is something that limits itself to worlds more nearly like ours. Hence, DT:

(DT) Any world that is a *minimal* physical duplicate of our world is a psychological duplicate of our world.

In notation, DP asserts:

$$\Box(PT \rightarrow Q)$$

Here we conjoin all the microphysical facts in our world, P, with T. This addition eliminates worlds that are physically identical to ours but contains extra non-physical things as well.

In his 2012 book *Constructing the World*, Chalmers offers his insights on T.[90] The idea of T originated in Russell. To be precise, Russell (1985)[91] discusses the question of whether or not there must be general facts in order for general truths to be true. Or, are general truths made true by general facts? Russell's answer, to my understanding, can be outlined as follows:

1. There are true general sentences such as:

   (A). All men are mortal.

2. There are particular facts corresponding to the truth of (A) such as:

   (F1). *a* is a man that is mortal.

   (F2). *b* is a man that is mortal.

   ……

   (Fn). *n* is a man that is mortal.

3. If general truth (A) is made true by particular facts (F1) … (Fn), then you have to know that (F1) … (Fn) are *all and the only* particular facts corresponding to the truth of (A).
4. If you know that (F1) … (Fn) are *all and the only* particular facts corresponding to the truth of (A), then you have to know (B):

   (B). *a, b … n* are all the men.

Therefore,

5. General truths are not made true by particular facts alone.
6. (B) is a general fact.

Therefore,

7. General truths are made true by general facts.

We can reveal the essence of T by analysing this Russellian argument.

Russell's argument starts by mentioning general truths. General truths are true general propositions. They are universally quantified propositions that say things like all things are such and such. (A) is an example of general truth. One way to confirm (A)'s truth is by looking at some features about the world. Intuitively, this verification process can be done by induction, namely, by simple enumeration. That is, if we count every individual man in the world, we would discover every one of them is mortal. Each and every one of

[90] Chalmers, D. (2012). *Constructing the World*, pp. 151-6.

[91] Russell, B. (1985). *The Philosophy of Logical Atomism*, pp. 101-8.

these individual mortalities is a particular fact about the world. Hence, (F1) … (Fn). Notice here that the number of these particular facts is finite, so it is possible for someone to complete the enumeration*;* and let us assume there really are some people or devices good enough to complete this simple, but tremendous enumeration. It seems that we can infer (A) from (F1) … (Fn).

Not quite, according to Russell. It is not the credibility of the result we got, in fact, (M1) … (MN) could be all true, and yet '*you cannot ever arrive at a general fact by inference from particular facts however numerous...*'[92]

Russell's reason is presented as the third and fourth premises of the above argument. In my view, Russell insists an epistemic point here: to infer (A) from (F1) … (Fn), one must have already known (B), for otherwise she would not know she had completed the enumeration. Thus, according to Russell, '*you will always have to have at least one general proposition in your premises*'.[93]

Moreover, from the point of view of logic, (A) cannot be the consequent of (F1) … (Fn). As Soames explains, it is right to universally instantiate (F1) … (Fn) from (A), while it is wrong to universally generalise (A) from (F1) … (Fn). This is because (A) has the logical form of $\forall x(Fx \rightarrow Gx)$, and each of (F1) … (Fn) has the form of *a is an F that is G*. It is logically possible for things to possess *F-ness* yet lacks *G-ness*. Thus, it is logically possible that all particular facts *a is an F that is G* are true, while $\forall x(Fx \rightarrow Gx)$ is false.[94]

Particular facts, therefore, cannot entail general truths. So Russell's sub-conclusion 5 is true. But up to this line, he merely shows that 'no general truths from particular facts'. How does he reach conclusion 7, namely, 'general truths from general facts'? This is where (B) kicks in, as my version of the Russellian argument suggests. (B) is a example of T-clause, The important question is: in order to infer (A), do we need a statement to assert the fact that 'all we have enumerated is all there are and nothing else are'? Isn't this statement redundant?

[92] Russell (1985), p. 101.

[93] Ibid.

[94] Soames, S. (2003). *Philosophical Analysis in the Twentieth Century*, Vol. 1, pp. 188-190.

As explained by Colin Cheyne and Chrles Pigden (2006)[95], it is not redundant. T has the role of providing a boundary or a limit to the universal discourse. It indicates that *a, b ... n* are the *totality* of individuals who are man and mortal, and consequently shows (M1) … (MN) are the *totality* of facts concerning *a, b ... n's* individual mortalities.[96]

Now this will rule out the occurrence of the scenario Soames proposed. Hence, $\forall x(Fx \rightarrow Gx)$ would not be false while all particular statements of the form *a is an F that is G* are all true. A T-clause imposes a limit on the number of individual constants that are eligible to be assigned to the variable *x.* Since all we enumerated is all there are and nothing else is**,** and all we enumerated are having both *F-nes*s and *G-ness,* it follows that there is no valuation that can simultaneously satisfy *Fx* and fail to satisfy *Gx.* Thus $\forall x(Fx \rightarrow Gx)$ is true by adding a T-clause to the particular facts we got. In the case of mortal men, it can be said that without a T-clause such as (B), (F1) … (Fn) do not *necessitate* (A)'s truth, whereas adding (B) to the inference process makes (A) a true conclusion necessarily. Thus, conclusion 7 of the Russellian argument can be derived.

Now we can go back to the mirroring case. In both Argument C and Argument MC, the scope of possible worlds that P obtains is limited by the addition of T. We might say that a P world is a world in which P is true. Then a PT world is a *minimal* world in which P is true. Thus, a PT world is, by definition, a *minimal* P world. With this in mind, we can provide a justification for $\Diamond(PT\&Q) \rightarrow \Box(PT\rightarrow Q)$.

Firstly, let us translate $\Diamond(PT\&Q)$ and $\Box(PT\rightarrow Q)$ as follows:

$\Diamond(PT \& Q)$ = There is at least one minimal world W_m in which P is true, and Q is also true in $W_{m.}$

$\Box(PT \rightarrow Q)$ = Every minimal world in which P is true has Q true in it.

Given the translations, we can prove the inference from the former to the latter:

Proof. $\Diamond(PT\&Q) \rightarrow \Box(PT\rightarrow Q)$

1. Suppose P is deductively closed. Since a minimal P world is a world in which P is true and nothing but P, Ω is true:

   Ω. Q is true in a minimal world = Q is logically entailed by P.

[95] Cheyne, C. & Pigden, C. (2006). 'Negative Truths From Positive Facts' in *Australasian Journal of Philosophy*, Vol. 84, No. 2, pp. 249-65.

[96] Ibid. P. 254.

2. Suppose Q is true in minimal P world W_m, then by Ω, Q is logically entailed by P.
3. Let W_v be a minimal P world such that $W_v \neq W_{m.}$ Then Q is true in minimal P world $W_{v.}$
4. Since W_v is arbitrarily chosen, if PT→Q is true in W_v then PT→Q is true in all worlds. Therefore, □(PT→Q).
5. From 1~4, ◇(PT&Q) → □(PT→Q)

In addition, if we translate PT as P *and its entailments* exhaust all the fact in the world, then we can have a second proof that looks as follows:[97]

Proof. ◇(PT&Q) → □(PT→Q)

1. Let W_1 be a world in which PT is true and Q is true. Since P and its entailments exhaust all the facts in a world, □(P→Q).
2. Let W_2 be a different world in which PT is true. Since □(PT→P), PT→P is true in W_2. Since □(P→Q), P→Q is true in W_2.
3. PT→P, P→Q ⊢ PT→Q. Thus, PT→Q is true in W_2.
4. Since W_2 is arbitrarily chosen, if PT→Q is true in *W*1 then PT→Q is true in all worlds. Therefore, □(PT→Q).

To sum up, the justification for ◇(PT&Q) → □(PT→Q) is the presence of T. Any world in which PT is true is a world where P exhausts all the fundamental facts. So any two possible worlds in which PT is true must be qualitatively identical. Therefore, if Q is true in one PT world, it is true in every PT world. In other words, if PT&Q is true in some possible worlds, then PT→Q is true in all possible worlds.

For this reason, the inference from MC4 to MC 5 can be explained, and Argument MC is thereby valid:

Argument MC
(P = all microphysical facts about the universe, T = ‘That’s all’, Q = qualia)

MC1. Materialism ↔ □(PT → Q) Assumption

MC2. $◇_c$(PT & Q) Assumption

MC3. $◇_c$Φ → ◇Φ Assumption (CEP)

[97] Thanks to Jack Copeland who is the mastermind behind this proof.

MC4. $\Diamond(PT \,\&\, Q)$ 2, 3, *modus ponens*

MC5. $\Box(PT \rightarrow Q)$ *Proof.* $\Diamond(PT\&Q) \rightarrow \Box(PT\rightarrow Q)$

MC6. Materialism. 1, 5, *modus ponens*

Lastly, having PT&Q as the content of conceivability in the mirroring argument has one more upshot. Since every other premise is identical to Chalmers' own premises, should he try to avoid mutual defeat, the only option is to reject MC2. Specifically, he needs to prove the inconceivability of PT&Q. Since there are no modal operators in front of PT&Q, this task is incredibly difficult. It is equivalent to claiming that the actual world is either physical or phenomenal but not both, which is outright wrong! Chalmers is too good a philosopher to have really been guilty of making such an absurd claim. In reply to this mirroring argument, he would claim that PT&Q is conceivable but not possible, which amounts to taking *position 3* in Kripke's *quadlemma*. But that too, leads to an unpalatable result, as explained in the previous Chapter.

Appendix. The conceivability fallacy

Co-authored with Douglas Campbell and B. Jack Copeland, and published as 'The Inconceivable Popularity of Conceivability Arguments' (2017), in *Philosophical Quarterly*, 67 (267), 223--240.

1. Introduction

Consider the following familiar situation. Someone alleges that a certain philosophically significant proposition, ϕ, is true. One would dearly like to refute this claim. Moreover, perhaps by way of rubbing salt into the wound, one would like to show, not just that ϕ isn't *actually* true, but that it *can't possibly be true*. That is, one's aim is to prove $\neg\Diamond\phi$. How to proceed?

The obvious method is as follows:

> *Reductio method:* First, prove ϕ is contradictory (i.e., that for some p, $\phi\rightarrow(p\wedge\neg p)$). Second, apply modal logic's Necessitation Rule (which lets $\neg\Diamond\phi$ be derived from $\phi\rightarrow(p\wedge\neg p)$).

Voila! Out pops $\neg\Diamond\phi$. Mission accomplished.

But the *reductio* method has a hitch. Proving that ϕ is contradictory can be, well … difficult! Sometimes, rack one's brains though one will, no contradiction springs to mind. Perhaps no contradiction is there to be found in ϕ in the first place. How to proceed in such a case? How to proceed, that is, when one badly wants to prove $\neg\Diamond\phi$ but ϕ seems to be *conceivable*? (By 'conceivable' we simply mean 'consistent' or 'does not entail any contradictions'. This species of conceivability corresponds to what Chalmers calls *ideal negative conceivability*.[98])

Here is an oft-tried approach:

First, identify two other propositions, ψ and μ.

[98] Chalmers (2010), pp. 143-8.

Second, show that ψ is (probably) conceivable, by *trying but failing* to detect a contradiction in it.[99]

Third, use the conceivability of ψ to infer $\Diamond\psi$, by applying the general principle that *conceivability entails possibility* (CEP).

Fourth, show that $\Diamond\psi\rightarrow\mu$, and infer μ by *modus ponens*.

Finally, show that $\Diamond\phi\rightarrow\neg\mu$, and infer $\neg\Diamond\phi$ by *modus tollens*.

The form of argument—henceforth the 'conceivability argument' (CA)—is as follows. ($\Diamond_c\psi$ represents the claim that ψ is conceivable.)

C1. $\Diamond_c\psi$

C2. $\Diamond_c\psi\rightarrow\Diamond\psi$ CEP

C3. $\Diamond\psi\rightarrow\mu$

C4. $\Diamond\phi\rightarrow\neg\mu$

C5. $\neg\Diamond\phi$

In what follows we begin by reviewing several important arguments of this form. We then show that there is something profoundly rotten in their logic.

2. Examples of CA

If the values of ϕ, ψ, and μ are set as follows,

ϕ: Pain=C-fibre stimulation

ψ: Pain≠C-fibre stimulation

μ: $\Box$(Pain≠C-fibre stimulation),

then CA becomes the following version of Kripke's 'modal argument' against psychophysical identity theory:

K1. $\Diamond_c$(Pain≠C-fibre stimulation)

K2. $\Diamond_c$(Pain≠C-fibre stimulation)$\rightarrow\Diamond$(Pain≠C-fibre stimulation)

[99] Since (for Gödelean reasons, among others) a contradiction might still be lurking somewhere among ψ's implications even if our best efforts to find it have so far been unsuccessful, a demonstration that ψ is conceivable will generally be *defeasible*.

K3. $\Diamond(\text{Pain}\neq\text{C-fibre stimulation})\rightarrow\Box(\text{Pain}\neq\text{C-fibre stimulation})$

K4. $\Diamond(\text{Pain}=\text{C-fibre stimulation})\rightarrow\neg\Box(\text{Pain}\neq\text{C-fibre stimulation})$

K5. $\neg\Diamond(\text{Pain}=\text{C-fibre stimulation})$

Here K1 is justified by the apparent absence of contradictions in the idea of pain being non-identical to C-fibre stimulation.[100] K2 is an instance of CEP. K3 is justified by the fact that both 'pain' and 'C-fibre stimulation' are rigid designators, and by Kripke's principle that all identities and non-identities between rigid designators are necessary. K4 is trivial. K5 is bad news for anyone wishing to identify mental types with physical or functional types, for the argument readily generalizes.

Kripke presents Argument K as a modern take on Descartes' argument for mind-body dualism.[101] It is therefore no surprise that Descartes' argument can itself be shoehorned into CA's form. Setting ϕ, ψ, and μ as follows:

ϕ: Mind=Body

ψ: Mind≠Body

μ: $\Box(\text{Mind}\neq\text{Body})$,

we get:

D1. $\Diamond_c(\text{Mind}\neq\text{Body})$

D2. $\Diamond_c(\text{Mind}\neq\text{Body})\rightarrow\Diamond(\text{Mind}\neq\text{Body})$

D3. $\Diamond(\text{Mind}\neq\text{Body})\rightarrow\Box(\text{Mind}\neq\text{Body})$

D4. $\Diamond(\text{Mind}=\text{Body})\rightarrow\neg\Box(\text{Mind}\neq\text{Body})$

D5. $\neg\Diamond(\text{Mind}=\text{Body})$

Argument D appears a plausible rational reconstruction of Descartes' argument in *Meditation VI*:

> I know that everything which I clearly and distinctly understand is capable of being created by God so as to correspond exactly with my understanding of it. Hence the fact that I can clearly and distinctly understand one thing apart from another is enough to make me certain that the two things are distinct,

[100] Kripke frames his argument in terms of 'Pain=C-fibre stimulation' being *a posteriori*, rather than in terms of its denial, 'Pain≠C-fibre stimulation', being conceivable. But the former implies the latter, since *p* can be *a posteriori* only if ¬*p* is contradiction-free, and thus only if ¬*p* is conceivable.

[101] Kripke (1980) pp. 144-50.

> since they are capable of being separated, at least by God... [O]n the one hand I have a clear and distinct idea of myself, in so far as I am simply a thinking, non-extended thing; and on the other hand I have a distinct idea of a body, in so far as this is simply an extended, non-thinking thing. And accordingly, it is certain that I am really distinct from my body, and can exist without it.[102]

Here Descartes uses a (theistic) version of CEP to infer that it is *possible* for mind and body to be distinct. From this he infers they are *actually* distinct. Why does he think he can make this move from mere possibility to actuality? Presumably because he is assuming D3, or something like it, as a tacit premise.[103]

Chalmers' 'zombie argument' against materialism can also be cast as a version of CA. Following Chalmers, let *P* represent a conjunction of all the microphysical facts. Let *T* be a 'totality operator' (or 'that's all' clause), which, when tacked onto the end of *P*, yields a proposition that says *P* provides a *complete* description of all the non-supervenient facts that obtain in the world. Let *Q* represent a conjunction of all the phenomenal facts. Thus '$\Box(PT\rightarrow Q)$' represents the materialist thesis that the phenomenal facts supervene metaphysically on the bare microphysical facts (i.e., that any 'PT-world' must also be a 'Q-world').

Plugging the following values for ϕ, ψ, and μ into CA,

ϕ: $PT\wedge \mathrm{Q}$

ψ: $PT\wedge\neg Q$

μ: $\neg\Box(PT\rightarrow Q)$,

we get this version of the zombie argument:

Z1. $\Diamond_c(PT\wedge\neg Q)$

Z2. $\Diamond_c(PT\wedge\neg Q)\rightarrow\Diamond(PT\wedge\neg Q)$

Z3. $\Diamond(PT\wedge\neg Q)\rightarrow\neg\Box(PT\rightarrow Q)$

Z4. $\Diamond(PT\wedge \mathrm{Q})\rightarrow\Box(PT\rightarrow Q)$

Z5. $\neg\Diamond(PT\wedge \mathrm{Q})$.

Here Z1 claims that a PT-world could conceivably fail to be a Q-world. It is justified by the apparent absence of contradictions in the idea of a PT-world being, say, a *zombie*

[102] Cottingham, Stoothoff, and Murdoch (1984), p. 54.

[103] Robinson, H. (2012) 'Dualism', in Zalta E.N. (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2012.). URL http://plato.stanford.edu/archives/win2012/entries/dualism/

world (a world wherein some or all human beings lack phenomenal consciousness). Z2 is an instance of CEP. Z3 is trivial.

Z1, Z2 and Z3 suffice by themselves to give Chalmers the result he is after, namely $\neg\Box(PT\rightarrow Q)$ (the denial of materialism). Chalmers doesn't need Z4, and so Z4 doesn't feature in the zombie argument as Chalmers himself presents it. Hence Chalmers' own version of the zombie argument doesn't quite fit the form of CA.

However, Z4 is a harmless addition to Chalmers' zombie argument because it is *analytic*. Why so? Well, recall that a *PT*-world is a logically possible world where *P* offers a *complete* description of the supervenience base. This being so, any two *PT*-worlds must be alike in all respects. No fact obtaining in one could fail to obtain in the other, every such fact being entailed by *PT*. So if *some PT*-world is a *Q*-world (i.e., a world that is a phenomenological duplicate of the actual world), then (by the meaning of the *T* operator) *every PT*-world must be a *Q*-world. In short, if $PT\wedge Q$ is true at some possible world, then $PT\rightarrow Q$ will be true at all possible worlds. This is what Z4 says.

Since Z4 is analytic Chalmers can't object if we add Z4 to his zombie argument as an extra premise, to produce Argument Z, which *does* fit CA's form. If Chalmers' version of the zombie argument is sound, then Argument Z is sound too. Contrariwise, if Argument Z is problematic, as we will show below, then so too is Chalmers' zombie argument.

The above examples all come from the philosophy of mind, but CA also crops up elsewhere. A case in point is the modal ontological argument[104] of which one formulation, obtained by setting the values of ϕ, ψ and μ as follows,

ϕ: God doesn't exist
ψ: God exists
μ: $\Box$(God exists),

is this:

O1. $\Diamond_c$(God exists)
O2. $\Diamond_c$(God exists)$\rightarrow\Diamond$(God exists)

[104] Hartshorne, C. (1965) *Anselm's Discovery: A Re-Examination of the Ontological Proof for God's Existence*. La Salle, IL: Open Court.
Malcolm, N. (1960) 'Anselm's Ontological Arguments', *Philosophical Review* 69: 41–62.
Plantinga, A. (1974) *The Nature of Necessity*.

O3. $\Diamond$(God exists)→$\Box$(God exists)

O4. $\Diamond$(God doesn't exist)→¬$\Box$(God exists)

O5. ¬$\Diamond$(God doesn't exist).

Here the justification for O3 turns on the idea that the concept of God is (in part) the concept of a necessarily existent being. The rest of the argument is self-explanatory.

3. Why CA is problematic

Let a 'conceivabilist' be a proponent of some version of CA. That is, she is someone who, for certain values of ϕ, ψ, and μ, defends the claim that $\neg\Diamond\phi$ is true by arguing that C1, C2, C3 and C4 are true. For example, proponents of Argument K, D, Z and O are conceivabilists.

We now show that the conceivabilist's position is untenable, and that by relying on CA to argue for $\neg\Diamond\phi$ she reasons fallaciously. To see the problem, consider the following 'mirror argument':

M1. $\Diamond_c\phi$

M2. $\Diamond_c\phi\rightarrow\Diamond\phi$ CEP

$\Diamond\phi$

For reasons to be explained in a moment, the arguments the conceivabilist uses to justify two of her own premises, C1 and C2, also justify M1 and M2. But M1 and M2 jointly entail $\Diamond\phi$, a conclusion that flatly contradicts $\neg\Diamond\phi$, the conclusion she is herself arguing for. And so she is caught in the jaws of an inconsistency.

Why must the conceivabilist accept M1? M1 is modeled on C1: where C1 says *ψ is conceivable*, M1 says *ϕ is conceivable*. The conceivabilist's reasons for accepting C1, will, if they are any good, consist of the fact that ψ *appears* to be conceivable, in the sense that ψ's logical implications appear to be contradiction-free. Now, suppose ϕ *also appeared to be conceivable*—i.e., that its implications also appeared to be contradiction-free. In this case the conceivabilist's reasons for accepting C1 would be matched by equally good reasons for accepting M1, and so C1 and M1 would stand or fall together:

the conceivabilist could reject M1 only on pain of admitting that C1 might just as easily have been rejected, instead.

This being so, a conceivabilist who opts to reject M1 must first of all break the symmetry between M1 and C1 by showing that whereas ψ appears to be conceivable, ϕ does not. In order to do this she needs to have some argument—call it *T*—at her disposal, which justifies her in doubting that ϕ is conceivable. *T* might consist of an outright demonstration of a contradiction in ϕ's implications, in which case it will provide *absolute certainty* that ϕ is inconceivable. But *T* needn't be quite so conclusive as this. For instance, it might merely consist of various forceful intuitions to the effect that ϕ is contradictory. If these intuitions have not yet been borne out by the actual detection of the putative contradiction in question, then they won't justify the conceivabilist in being 100% certain that ϕ is contradictory. But they might still justify her in being relatively confident that ϕ is contradictory, and thus relatively confident that ϕ is inconceivable.

Now if the conceivabilist possesses some such *T* then she can, perhaps, reject M1 without undermining C1 in the process. But here's the rub. In using *T* to reject M1 the conceivabilist will be protecting CA's premises from being refuted by M1 and M2, but only at the expense of exposing these selfsame premises as being *logically redundant*. To see this, notice that if ϕ entails a contradiction (i.e., if ϕ is inconceivable), then $\neg\Diamond\phi$ can be proved by the *reductio* method (described in §1), instead of by using CA. Hence, just to the degree that *T* provides the conceivabilist with reason to think that ϕ is contradictory, it also provides her with a direct, *reductio*-method-based proof of $\neg\Diamond\phi$, a proof that is logically independent of CA itself. In other words, just in so far as *T* justifies the conceivabilist in rejecting M1, it also justifies her in thinking that CA is otiose and dispensable. What *T* gives the conceivabilist with one hand, by enabling her to defend CA from M1 and M2, it takes back with the other, by rendering CA superfluous.

Compare. A burglar attempts to enter a house through a small window. Finding the window securely latched, he breaks down a door, unlatches the window from the inside, exits through the door, then successfully squeezes back in through the window. Triumph! Of course having broken down the door he no longer needed to bother with entering through the window. Similarly, a conceivabilist who can defend CA's premises from the mirror argument by using *T* to argue against M1 doesn't need to bother anymore with using CA to prove $\neg\Diamond\phi$. *T* does the job by itself. CA is surplus to her requirements.

From what has just been said it follows that if CA *is not* logically redundant—i.e., if the conceivabilist is genuinely reliant on CA to prove $\neg\Diamond\phi$—then she can't have any such argument as *T* at her disposal. She therefore won't be in a position to defend CA's premises from the mirror argument by rejecting M1. Needless to say, if Descartes, Kripke, Chalmers or proponents of the modal ontological argument were able to prove their respective conclusions using the *reductio* method, without relying on CA, then they would be first to recognize and loudly trumpet this fact. They resort to using CA only because they have no such argument as *T* up their sleeves. Hence they are not in a position to repudiate the mirror argument by rejecting M1.

Before we move on, three brief clarifications are in order. The first concerns logical redundancy. Redundancy in one's arguments can, of course, be a useful and desirable thing. If one has two arguments, *G* and *H*, that share the same conclusion, *p*, then *G* can be used as backup in cases where *H* fails to convince, and *vice versa*. Notice, however, that such redundancy is useful only if *G* and *H* are *logically independent* of each other, in the sense that there are good reasons for accepting *G*'s premises that don't presuppose the truth of *H*'s premises, and *vice versa*. If *G*'s premises were vulnerable to some counterargument, and if it were necessary to rely on *H*'s premises in order to defend *G*'s premises from this counterargument, then *G* wouldn't provide any genuine support for *p* over and above the support already provided by *H*. *H* would be doing all the real logical work, and G would be otiose.

Unfortunately for the conceivabilist she would, in using *T* to defend CA's premises from the mirror argument, be making CA logically dependent on *T*. The resulting logical redundancy is therefore of the useless variety, not the useful variety. By using *T* to defend CA's premises and then using CA's premises to argue for $\neg\Diamond\phi$ she would be relying on a complete set of premises comprised of all of *T*'s premises and all of CA's premises. But since *T* shows ϕ is contradictory, *T*'s premises suffice by themselves to provide a *reductio*-method-based proof of $\neg\Diamond\phi$. This makes the other premises she is invoking—namely, those that belong to CA but not to *T*—extraneous where the goal of proving $\neg\Diamond\phi$ is concerned. They add nothing but pointless complexity to the overall case for thinking $\neg\Diamond\phi$ is true. They are like cogs in a clockwork that can be removed without disturbing the clockwork's function.

The second clarification concerns the conceivabilist's goal. It is of course *part* of the conceivabilist's goal to prove $\neg\Diamond\phi$. She would succeed in attaining at least this part of her goal were she to use *T* and the *reductio* method to prove $\neg\Diamond\phi$. This is not in dispute. We have no objection to the idea that $\neg\Diamond\phi$ might be proved using the *reductio* method. But it is the conceivabilist's goal, not to prove $\neg\Diamond\phi$ using *some method or other*, but to prove it *using CA*. It is distinctive of the conceivabilist that she thinks CA is capable of providing us with good reason to accept $\neg\Diamond\phi$. If a conceivabilist were to construct an argument, *T*, with which to attack M1, and furthermore acknowledge that *T* renders CA pointless and superfluous, then we would no longer be in any disagreement with her: but she would have renounced her position and be a 'conceivabilist' no more.

The third clarification concerns *degrees of confidence*. Clearly if *T* provides *absolute certainty* that ϕ is contradictory, then it also enables $\neg\Diamond\phi$ to be proved outright by the *reductio* method, so rendering CA *entirely* redundant. But what if *T* only justifies the conceivabilist in being x% confident that ϕ is contradictory, where $0<x<100$? Might CA have a useful logical role to play in this case? No. To see why not let's distinguish two epistemic possibilities. Possibility 1 is that ϕ is contradictory and M1 is false. Possibility 2 is that ϕ is non-contradictory and M1 is true. *T* enables the conceivabilist to assign a credence of x% to Possibility 1. This leaves a credence of (100-x)% to be assigned to Possibility 2 (Possibility 2 being simply the logical complement of Possibility 1). In being x% confident that Possibility 1 obtains, the conceivabilist can also be x% confident that CA is logically redundant (because if Possibility 1 obtains then the *reductio* method proves $\neg\Diamond\phi$). In being (100-x)% confident that Possibility 2 obtains, she can also be (100-x)% confident that CA's premises cannot be successfully defended from the mirror argument by attacking M1 (because if Possibility 2 obtains, then M1 is true). Putting these two results together, we obtain the conclusion that she can be $x\%+(100-x)\%=100\%$ confident *that either CA is logically redundant or CA's premises cannot be successfully defended from the mirror argument by attacking M1*. In other words, if there is epistemic uncertainty as to which of these two possibilities obtains then there will be corresponding uncertainty as to which of the two horns of a dilemma the conceivabilist will be impaled by, but this should be of cold comfort to the conceivabilist *because it is still 100% certain that she will be impaled by one horn or the other*.

So much for M1. Next, why can't the conceivabilist reject M2? M2 is modeled on C2, for where C2 says *that if ψ is conceivable then ψ is possible*, M1 says instead *that if ϕ is conceivable then ϕ is possible*. The conceivabilist justifies C2 by invoking CEP, the general principle that if a proposition is conceivable then it is possible. But unfortunately for the conceivabilist this principle justifies M2 every bit as much as it justifies C2. Were the conceivabilist to reject M2 then she would, in effect, be holding that ϕ is conceivable but impossible. If this were right, then ϕ would be a counterexample to CEP, which would raise the possibility of ψ being another such counterexample. The conceivabilist's own CEP-based argument for C2 would thereby be severely undermined: for if the conceivabilist herself grants that CEP fails where ϕ and M2 are concerned, then why should CEP be trusted where ψ and C2 are concerned?

Summary. On the one hand the conceivabilist uses C1, C2, C3 and C4 to argue for $\neg\Diamond\phi$. On the other hand, her reasons for accepting C1 and C2 also support M1 and M2, which together entail $\Diamond\phi$. Specifically, just as CEP supports C2, so too it supports M2. And just as an apparent absence of contradictions in ψ suggests that C1 is true, so too an apparent absence of contradictions in ϕ would suggest that M1 is true. The conceivabilist is in no position to deny that ϕ appears to be contradiction-free, since in denying this she would be setting up a *reductio*-method-based proof of $\neg\Diamond\phi$, which would render CA logically redundant. The conceivabilist's premises and the principles she uses to justify these premises therefore 'prove too much'. They generate a contradiction, in the form of $\neg\Diamond\phi \wedge \Diamond\phi$.

Other critics of the various different versions of CA have noticed that they are vulnerable to being 'mirrored' along the above lines (although the term 'mirroring' is ours). For example, Bayne (1988) argues that Kripke's argument can be mirrored to yield a conceivability argument *for* (rather than *against*) psychophysical identity theory. Frankish (2007) constructs an 'anti-zombie argument', which amounts to a mirrored version of Chalmers' zombie argument. Marton (1998), Yablo (1999), Sturgeon (2000), and Brown (2010) present similar criticisms of the zombie argument. Where the modal ontological argument is concerned, various authors have observed that it seems possible to run it backwards, starting from the conceivability of a (necessarily-existent) God *not* existing and then inferring that it is *not possible* for there to be such a God.[105]

However connections are seldom drawn between these disparate literatures. It appears to have gone unrecognized that the logical issues being encountered in each of the cases are, at root, the same, and that the problem is a general one that afflicts all versions of CA identically. Moreover, at least to our minds, none of these authors have exposed the true depth of CA's logical bankruptcy. Their critiques of the various versions of CA suggest the presence of loopholes though which a proponent of CA might escape. (See our discussion of (Zemach 1994) and (Chalmers 2010), below.) We don't think these loopholes are real. By way of showing this we now return to Arguments K, Z, D and O, and examine how the general mirroring objection we have just outlined plays out in each case.

4. Against Argument K

As explained in §2, Argument K is a version of CA wherein ϕ's value is 'Pain=C-fibre stimulation'. Plugging this value for ϕ into the mirror argument produces the following mirrored version of Argument K:

K′1. $\Diamond_c$(Pain=C-fibre stimulation)

K′2. $\Diamond_c$(Pain=C-fibre stimulation)$\rightarrow\Diamond$(Pain=C-fibre stimulation)

$\Diamond$(Pain=C-fibre stimulation)

K′1 and K′2 conjointly entail $\Diamond$ (Pain=C-fibre stimulation), thereby flatly contradicting K1—K4, which conjointly entail $\neg\Diamond$(Pain=C-fibre stimulation). Hence to save his premises Kripke must reject K′1 or K′2.

Can Kripke reject K′1? Suppose he knew of good reasons for thinking that the *Pain=C-fibre stimulation* hypothesis is contradictory. Then he could immediately refute psychophysical identity theory using the *reductio* method, and so Argument K would be redundant. But of course he doesn't use the *reductio* method to prove this result; he relies on Argument K instead. Great philosopher that he is, he wouldn't keep a simple,

[105] McGarth, P.J. (1990) 'The Refutation of the Ontological Argument', *Philosophical Quarterly* 40/159: 195–212.

knockdown, *reductio*-based refutation of psychophysical identity theory secret if he had one up his sleeve. We may therefore conclude that he can't tender good reasons for thinking that the *Pain=C-fibre stimulation* hypothesis is contradictory. This being so, the *Pain=C-fibre stimulation* hypothesis and the *Pain≠C-fibre stimulation* hypothesis are, for Kripke, on a par, in the respect that to the best of his knowledge *neither one* of them entails a contradiction and *both* appear to be conceivable. Kripke uses the apparent conceivability of the *Pain≠C-fibre stimulation* hypothesis to justify K1, which says that *Pain≠C-fibre stimulation* is in fact conceivable.[106] By parity of reasoning, the apparent conceivability of the *Pain=C-fibre stimulation* hypothesis likewise justifies K′1, which says that *Pain=C-fibre stimulation* is conceivable. If Kripke is warranted in concluding that the one hypothesis is conceivable based on its appearing to be contradiction-free, then he is warranted in concluding that the other hypothesis is conceivable on the same grounds. And so Kripke is obliged to accept K′1.

His only other option is to reject K′2. Kripke (famously) recognizes that there are certain exceptions to CEP, in the form of propositions, like, say, *Water≠H_2O*, which are conceivable but impossible (or, in his terminology, *a posteriori* but necessarily false). However he notes that such exceptions to CEP involve rigid designators (like the natural-kind term, 'water') that secure reference via accidental properties of their referents. He points out that neither 'Pain' nor 'C-fibre stimulation' is such a rigid designator, since both these terms pick out their referents via *essential* properties—the way pain *feels* in the one case, and the essential scientific nature of C-fibre stimulation in the other. He therefore concludes, albeit somewhat tentatively,[107] that the *Pain≠C-fibre stimulation* hypothesis is *not* an exception to CEP. This gives him his premise K2 (which says that if *Pain≠C-fibre stimulation* is conceivable then it is possible). But if the *Pain≠C-fibre stimulation* hypothesis if free of rigid designators that refer via accidental properties of their referents, then so too is the *Pain=C-fibre stimulation* hypothesis: for the two hypotheses differ only in that one, being the denial of the other, includes an additional negation concept. Hence the restricted version of CEP used by Kripke to justify acceptance of K2 also justifies acceptance of K′2.

[106] Kripke (1980) speaks, variously, of *Pain≠C-fibre stimulation* being 'epistemically possible', or of it being '*a posteriori*', or of *Pain=C-fibre stimulation* 'appearing contingent'. He does not, as we do, speak of *Pain≠C-fibre stimulation* being *conceivable*. But this is a mere difference of terminology.

[107] Kripke (1980), p. 148, 150.

In short, although Kripke must reject K′1 or K′2, he can reject them only on pain of admitting that the reasoning he uses to justify K1 and/or K2 cannot be trusted.

This mirroring objection to Kripke's modal argument is partly anticipated by Bayne (1988), who, like us, points out that Kripke's argument is susceptible to being turned on its head.[108] Bayne's argument has received scant attention, but is critiqued by Zemach (1994), who defends Kripke. Since Zemach's argument against Bayne might be adapted to make trouble for us, it will be instructive to examine it.

Zemach argues, in effect, that there is an asymmetry between Kripke's premise, K2, and the corresponding premise of the mirror argument, K′2. Specifically, he contends that the latter is vulnerable to a mode of attack against which the former is invulnerable.

Let's start with the vulnerability of K′2. To attack K′2, it would be necessary to show that *pain=C-fibre stimulation* might be conceivable even if it were in fact impossible, which is to say, even if pain and C-fibre stimulation were non-identical. Zemach points out that even if pain and C-fibre stimulation were non-identical, we could still imagine them always co-occurring *as if they were identical*. He holds that we would thereby, in effect, *be imagining them being identical*. This opens the door to rejecting K′2. In arguing along these lines, Zemach is tacitly relying on the following principle:

> P: If A and B are two non-identical states, then in order to conceive of A=B being true, it suffices for one to conceive of A occurring whenever B occurs, and *vice versa*.

Next, why according to Zemach is K2 invulnerable to the same style of attack? Well, to attack K2 it would be necessary to show that *pain≠C-fibre stimulation* might be conceivable even if it were in fact impossible, which is to say, even if pain and C-fibre stimulation were in fact identical. But if pain and C-fibre stimulation were identical, then how could one coherently imagine them being non-identical? Zemach notes that this cannot be done by simply imagining that pain and C-fibre stimulation sometimes *fail* to

[108] There are both terminological and substantive differences between our argument and Bayne's. On the terminological front, Bayne frames his discussion in terms of the 'apparent contingency' of the *Pain≠C-fibre stimulation* hypothesis, not, as we do, in terms of the conceivability of the *Pain=C-fibre stimulation* hypothesis. On the substantive front, Bayne doesn't point out, as we do, that Kripke *must* concede that *Pain=C-fibre stimulation* appears conceivable, on pain of rendering his modal argument against psychophysical identity theory redundant. And unlike us Bayne doesn't present his objection to Kripke as being a mere instance of a much more general objection against *all* conceivability arguments.

co-occur *as if they were non-identical*, because on the operative assumption that pain and C-fibre stimulation are in fact identical, this is incoherent. (If A and B are identical states, then to imagine A occurring *just is* to imagine B occurring, and *vice versa*.)

That's Zemach's argument. What's wrong with it? Our answer is that P is not remotely credible. Two states can reliably co-occur without being identical, as when they are non-identical states that co-occur by chance alone, or as when they share a common cause. Thus it is simply not the case that conceiving of A and B reliably co-occurring suffices for conceiving of A=B being true.

Zemach might respond by conceding the point, but then switching targets from K′2 to K′1. He might claim that when *we think* we are conceiving of *pain=C-fibre stimulation* we are *really* just conceiving of pain and C-fibre stimulation co-occurring as if they were identical (a kind of mistake that has no analogue where *pain≠C-fibre stimulation* is concerned). He might deny on this basis that *pain=C-fibre stimulation* is genuinely conceivable. In reply we note that: (i) at best Zemach would thereby have explained why K′1 might appear to be true even if it were in fact false. He would not have demonstrated that *pain=C-fibre stimulation* entails a contradiction, and so he would not have shown that K′1 is in fact false. And (ii) even if this argument did show that K′1 is false (which it doesn't) then it would thereby save Argument K from being mirrored only at the expense of exposing Argument K as being logically redundant: for, as we have seen, if K′1 is false then Argument K's conclusion can be proved by the *reductio*-method, without using Argument K's premises at all.

5. Against Argument Z

Substituting $PT \wedge Q$ for ϕ within the mirror argument yields the following mirrored version of Argument Z:

Z′1. $\Diamond_c(PT \wedge Q)$

Z′2. $\Diamond_c(PT \wedge Q) \rightarrow \Diamond(PT \wedge Q)$

$\Diamond(PT \wedge Q)$

Z′1 and Z′2 together entail $\Diamond(\mathrm{PT}\wedge\mathrm{Q})$, which contradicts Argument Z's conclusion, $\neg\Diamond(\mathrm{PT}\wedge\mathrm{Q})$, and therefore also contradicts its premises, Z1—Z4.[109] To save his premises Chalmers must reject Z′1 or Z′2.

Can Chalmers reject Z′1? Suppose he could show that Z′1 is false by showing $\mathrm{PT}\wedge\mathrm{Q}$ is contradictory. Having done this he could then use the *reductio* method (of §1) to prove $\neg\Diamond(\mathrm{PT}\wedge\mathrm{Q})$, from which the denial of materialism can then be derived *via* the following conditional:

COND: $\neg\Diamond(PT\wedge Q)\rightarrow\neg\Box(PT\rightarrow Q)$

To see why COND must be accepted, consider a PT-world, *w* (a possible world that is a minimal physical duplicate of our actual world). (Notice that such a *w* certainly exists. If materialism is true then *w* will be identical to the actual world. If materialism is false then *w* will be a bare physical duplicate of the actual world from which non-physical things have been subtracted.) COND's antecedent says, in effect, that no PT-world is a Q-world. Assume this is true. Then it follows that *w* is not a Q-world. Thus there is at least one world—namely, *w*—that is a PT-world but *not* a Q-world. Thus it is not the case that *every* PT-world *is* a Q-world. This is what COND's consequent says. And so, assuming COND's antecedent is true, its consequent is true too. Thus, COND itself is true. Q.E.D.[110]

In short, if Chalmers were able to refute Z′1 by showing that $PT\wedge Q$ is contradictory, he could then go on to provide an immediate, slam-dunk, *reductio*- (and COND-) based refutation of materialism. Since Chalmers relies on the zombie argument to refute materialism instead of providing any such *reductio*-based refutation, we can reasonably assume that he is unable to show that $PT\wedge Q$ is contradictory. But this means that for Chalmers $PT\wedge Q$ and $PT\wedge\neg Q$ are on a par, in the respect that to the best of his knowledge neither entails a contradiction, so that they both appear conceivable. His argument for Z1 rests on the apparent conceivability of $PT\wedge\neg Q$. By parity of reasoning,

[109] □$(PT\wedge Q)$ also entails □$(PT\rightarrow Q)$ (i.e., the truth of materialism) via Z4, which was shown to be analytic in §2, above.

[110] COND and Z4 together entail the bi-conditional, □$(PT\wedge \mathrm{Q}) \leftrightarrow$ □$(PT\rightarrow Q)$ (i.e. □$(PT\wedge \mathrm{Q}) \leftrightarrow$ materialism).

the apparent conceivability of $PT \wedge Q$ supports an argument for Z′1. Hence Chalmers is in no position to reject Z′1.

This means he must instead reject Z′2. Chalmers shares Kripke's reservations about CEP. Like Kripke, he acknowledges that conceivability is not a reliable guide to possibility when rigid designators that secure reference via accidental properties of their referents (e.g., names, natural kind terms and indexicals) are in play.[111] However he maintains that $PT \wedge \neg Q$ is free of such rigid designators, and thus that conceivability entails possibility at least where $PT \wedge \neg Q$ is concerned. This gives him his premise, Z2. But if $PT \wedge \neg Q$ is free of such rigid designators, then so too is $PT \wedge Q$, since the latter proposition differs from the former only in respect of containing one less negation concept. Hence by Chalmers' own reasoning we can conclude that if $PT \wedge Q$ is conceivable then it is possible. This gives us Z′2.

And so Chalmers is caught in the same trap as Kripke. He must reject Z′1 or Z′2, but can do this only by admitting that there is something wrong in his own arguments for Z1 and Z2.

Several authors—e.g., Marton (1998), Yablo (1999), Sturgeon (2000), Frankish (2007) and Brown (2010)—have argued, similarly to us, that the zombie argument's logic can be hijacked to produce a conclusion inconsistent with its own conclusion. Chalmers has two main counter-arguments. First he notes that whereas his own Argument Z trades on the conceivability of a non-modal claim—namely, $PT \wedge \neg Q$—Marton, Yablo and Sturgeon instead rely on the conceivability of a modal claim—namely, $\Box(PT \rightarrow Q)$ (materialism). He then points out that: (i) there is room for him to deny that CEP applies to such modal claims; and (ii) such modal claims concern the constitution of the entire space of possible worlds, not just the constitution of a single possible world, which makes their conceivability difficult to evaluate.[112] This counter-argument is of no help to Chalmers where Argument Z′ is concerned, since Argument Z′ is just like Argument Z in that it trades on the conceivability of a non-modal claim—namely, $PT \wedge Q$.

Chalmers' second counter-argument is as follows:

[111] In Chalmers' (1996, 2010) terminology, conceivability is an unreliable guide to possibility when the primary and secondary intensions of a proposition diverge.

[112] Chalmers (2010), p. 179-80.

> It may be prima facie negatively conceivable that materialism is true about consciousness, but it is not obviously conceivable in any stronger sense. Many people have noted that it is very hard to imagine that consciousness is a physical process. I do not think that this unimaginability is so obvious that it should be used as a *premise* in an argument against materialism, but likewise the imaginability claim cannot be used as a premise either.[113]

Chalmers is here attempting to negotiate a safe path between two horns of a dilemma. On the one hand he doesn't want to grant that materialism is manifestly *inconceivable* (i.e., that it clearly entails a contradiction) because then materialism could be refuted by a simple *reductio* argument and so his zombie argument against materialism would be superfluous. Call this *the threat of redundancy*. On the other hand, he mustn't grant that materialism is *conceivable* either, because then CEP could be used to infer that materialism is *possible*, and from this it would follow (via the fact that materialism is itself a modal claim) that materialism is *true*. Call this *the threat of absurdity*.

Chalmers' way out of the dilemma, as intimated in the above passage, is to maintain that it is *epistemically uncertain* whether materialism is conceivable or inconceivable. If Chalmers is right then: (i) uncertainty about whether materialism is inconceivable prevents materialism being proved false with a *reductio* argument and so saves the zombie argument from the threat of redundancy; and (ii) uncertainty about whether materialism is conceivable prevents the zombie argument's logic from being hijacked to prove materialism true, thereby saving it from the threat of absurdity.

That's Chalmers' general strategy. He could bring it to bear against our mirror argument by maintaining that it is epistemically uncertain whether $PT \wedge Q$ is conceivable or inconceivable, and thus epistemically uncertain whether $Z'1$ is true or false. But this approach fails for reasons previewed in §3, above. Viz., uncertainty about whether $Z'1$ is true or false merely translates into uncertainty about *which* horn of the dilemma Chalmers will be skewered by, not into uncertainty as to *whether* he will be skewered by one horn or the other. To see this, suppose Chalmers has reasons for being *x*% confident that $PT \wedge Q$ is inconceivable, and thus for being (100-*x*)% confident that $PT \wedge Q$ is conceivable. Suppose, furthermore, that these reasons leave him epistemically uncertain whether $PT \wedge Q$ is conceivable or inconceivable. Thus $0 \ll x \ll 1$. Given Chalmers is *x*% confident that $PT \wedge Q$ is inconceivable, he should also be *x*% confident that $\neg\Diamond(PT \wedge Q)$

[113] Chalmers (2010), p. 180.

can be proved by the *reductio* method, instead of by using CA, and so he should be *x*% confident that Argument Z succumbs to the threat of redundancy. Furthermore, given he is (100-*x*)% confident that $PT \wedge Q$ is conceivable, he should be (100-*x*)% confident that Argument Z succumbs to the threat of absurdity; for if $PT \wedge Q$ is conceivable then Z′1 is true, and if Z′1 is true then Z′1 and Z′2 together entail $\Diamond(PT \wedge Q)$, which contradicts Argument's Z's conclusion. Putting these two results together, the credence he assigns to the proposition that Argument Z succumbs *either* to the threat of absurdity *or* to the threat of redundancy should be *x*%+(100-*x*)%=100%.

6. Against Arguments D and O

We will keep our comments on Arguments D and O brief.

Substituting *Mind=Body* for ϕ in the mirror argument yields this mirrored version of Argument D:

D′1. $\Diamond_c$(Mind=Body)

D′2. $\Diamond_c$(Mind=Body)→$\Diamond$(Mind=Body)

$\Diamond$(Mind=Body)

Similarly, substituting *God doesn't exist* for ϕ in the mirror argument yields the following mirrored form of Argument O:

O′1. $\Diamond_c$(God doesn't exist)

O′2. $\Diamond_c$(God doesn't exist)→$\Diamond$(God doesn't exist)

$\Diamond$(God doesn't exist)

D′1 and D′2 conjointly entail $\Diamond$(Mind=Body), thereby contradicting Argument D's conclusion. Hence a Cartesian proponent of Argument D must repudiate D′1 or D′2. By the same token a theist proponent of Argument O must repudiate O′1 or O′2.

Might the Cartesian reject D′1, or might the theist reject O′1? If the Cartesian could refute D′1 by demonstrating the existence of a contradiction in the *Mind=Body* hypothesis then she could go on to offer a simple, knockdown proof of ¬$\Diamond$*(Mind=Body)* using the *reductio* method, and so Argument D would be logically redundant. Since she relies on Argument D to prove $\neg\Diamond$*(Mind=Body)* instead of using a *reductio* argument, she presumably knows of no such contradiction. This being so, the *Mind≠Body* and *Mind=Body* hypotheses will be on a par for the Cartesian, in the respect that both appear to be contradiction-free. Her grounds for accepting the conceivability of the latter hypothesis are therefore just as good as her grounds for accepting the conceivability of the former. This means she is obliged to accept D′1 for the same reasons she accepts D1. Were she to fail to accept D′1 she would be tacitly acknowledging a weakness in her argument for D1.

What has just been said about the Cartesian holds equally for the theist. If the theist could show that *God doesn't exist* entails a contradiction, then she could prove God's existence by *reductio*, and so she wouldn't need Argument O in the first place. Given that she instead relies on Argument O, we can infer that she can't show that *God doesn't exist* entails a contradiction. But this being so, *God exists* and *God doesn't exist* are for the conceivabilist on a par: they both appear conceivable. The same considerations that drive her to accept O1 should also drive her to accept O′1. She could baulk at accepting O′1 only at the cost of acknowledging a weakness in her argument for O1.

If the Cartesian doesn't reject D′1, she must instead reject D′2. Likewise, if the theist does reject O′1, she must instead reject O′2. But both D′2 and O′2 are instances of CEP. If D′2 is false, it is a counterexample to CEP; and likewise for O′2. Neither the Cartesian nor the theist can afford to concede that CEP has such counterexamples, because the Cartesian relies on CEP to justify D2, and the theist relies on it to justify O2.

Thus the Cartesian and the theist are caught in the same trap as Kripke and Chalmers: they must reject one of the mirror argument's premises in order to avoid an outright contradiction, but in rejecting either one of the mirror argument's premises they would be conceding that their own premises are inadequately supported.

7. Conclusion

We began with the question as to how one might prove $\neg\Diamond\phi$ if, due to the fact that ϕ appears to be contradiction-free, one is unable to prove it by the *reductio* method. The conceivabilist thinks CA provides an answer. We hope to have persuaded the reader that it does not. The problem is really perfectly simple. If ϕ appears to be contradiction-free, then ϕ appears to be conceivable, and if ϕ appears to be conceivable then the conceivabilist's own logic can be easily adapted to prove $\Diamond\phi$, a conclusion that exactly contradicts the conclusion the conceivabilist is aiming for! This adapted version of CA, which proves $\Diamond\phi$, is our 'mirror argument'.

We have shown that four famous arguments—namely, Kripke's modal argument, the Cartesian argument for dualism, Chalmers' zombie argument and the modal ontological argument—can each be cast as versions of CA, and that each are susceptible to being mirrored. We conclude that all four of these arguments are logically bankrupt.

Part II:

Indeterminacy and the multiple realisability argument

Chapter 4. The multiple realisability argument against identity theory

4. 0. Outline of Part II

Part I examined conceivability arguments against identity theory and materialism proposed by Kripke, Chalmers, respectively. I argued that all these arguments are 'mirror-able' and concluded on this basis that mind-brain identity cannot be refuted by conceivability arguments. A general diagnosis of the formal fallacy committed within conceivability arguments was also given. What's left for the foes of identity theory? Is there any extant criticism against mind-brain theory that is not fallacious and thereby more compelling? The aim of Part II is to consider one of the earliest objections to identity theory, which many agree remains the biggest threat to it. This is the so-called 'argument from multiple realisability'. The idea that mental states can be multiply realised has been widely discussed by philosophers since its popularisation by Putnam. Today the common reply to this counter-argument is to take the functionalist approach – the dominant mind-brain theory nowadays. This requires rejecting at least the identity theory.

Part II is divided into four Chapters. In Chapter 4, an exposition of the notion of multiple realisability will be provided. It summarises both Putnam's argument and existing objections against it. Chapter 5 revisits Putnam's argument, focusing on a vital rule of inference that Putnam relies on and highlighting why the rule is problematic. In doing so, it draws on the idea that phenomenal terms are *vague* and the idea that identity claims containing vague terms are sometimes *indeterminate*. Chapter 6 addresses some foreseeable objections to this approach of mine. Finally, Chapter 7 attempts to formalise, via proper proofs, the logical relation between vagueness and indeterminacy – the two key elements in my argument against the multiple realisability objection, and in so doing makes my indeterminacy reply to Putnam's multiple realisability argument fully robust.

4. 1. Multiple realisability and functionalism

There is a can of baked beans sitting to the left of my laptop as I am typing this. It will soon be my lunch today, and I will need a can opener to cut off the tin lid. I will use the one sitting in my kitchen drawer, which is a rotating wheel opener, commonly seen in every household. But, I could instead choose another tool that would deliver the same result. For example, my little Swiss multi-tool has a key-sized blade that can be folded out to pierce the can lid. Or, I could use a kitchen knife, a pair of scissors, or even my teeth to poke a number of little holes on the lid. In fact, there are numerous tools other than the rotating wheel opener that can *realise* the *role* of a can opener. It does not matter what they are made out of as long as they do the job. In short, can openers are 'multiply realised'; it is possible for can openers to be instantiated in many different ways.

One way to find out whether mental states are brain states is to ask whether it is possible for mental states to be instantiated by things other than brain states. If the answer is 'yes', then just as can openers need not have a rotating wheel design, so too mental states need not be brain states. If a theory has as its central tenet the claim that can openers necessarily are of rotating wheel design, then it is a false theory. By the same token, if mental states can be realised by things other than brain states, the idea that mental states are identical to brain states is false. Since the beginning of the 1960s a number of philosophers have attacked identity theory on this basis. They have argued, via various argumentations, for the possibility of mental states being instantiated in things other than brain states. Putnam's 'Minds and Machines'[114] is widely recognised as having pioneered this objection to identity theory.

4. 1. 1. Turing machines and minds

Putnam's strategy, as seen in this paper, is to investigate a question that he claims is 'logically analogous'[115] to whether it is acceptable to identify minds with brains—namely, the question of whether or not Turing-machine states are multiply realised. Putnam argues 'yes', and concludes, by analogy, that mental states are multiply realised too. Consequently, there are two steps in Putnam's strategy. The first step is to explain

[114] Originally published in 1960, reprinted as Putnam, H. (1975b), 'Minds and Machines' in *Mind, Language and Reality*, Vol. 2, pp. 362-85.

[115] Ibid. P. 362.

why and how any given Turing machine can be physically instantiated in multiple different ways. The second step is to provide a framework under which mental states can be conceptualised as Turing machine states.

To understand Putnam's argument, it is first necessary to understand what a Turing machine is. Alan Turing, in his famous and hugely influential 1936 paper 'On Computable Numbers', proposes a formalism that aims to provide a solution to Hilbert's decidability problem.[116] He introduced the name of '*computing machines*' – now known as *Turing machines*.[117] A Turing machine is an abstract device. It can be formally defined in several equivalent ways. The following definition follows Ned Block's description.[118] A Turing machine computes function from inputs to outputs. The key components of the machine are a *head*, and a *tape* of infinite length that is divided into an infinite number of cells. These cells on the tape contain symbols which are usually numerals like '0' and '1'. The symbols inscribed in the cells of the tape when the machine first begins operating together comprise the machine inputs. Figure 4-a is an illustration of a Turing-machine tape:

Figure 4-a

…	0	1	1	1	0	0	1	0	…

The machine operates through a sequence of individual steps. At each step the head is situated on one cell of the tape, which we may call the current cell. The head *reads* the inputs and proceeds in accordance with the *Turing machine table*. The machine table is a finite list of rules that govern the action of the machine. Each instruction is in the form of a conditional statement as follows: if the machine is in state S_1 and receives input I_1, then emits output O_1 and goes to S_2 (or stays in S_1).[119] For example, some instructions that the table tells the head to perform are:

[116] Turing, A. (1936), 'On Computable Numbers, with an Application to The Entscheidungsproblem' in *Proceedings of the London Mathematical Society*, 42, Vol. 1, pp. 230-65.

[117] Block, N. (1980a), 'Functionalism' in N. Block (ed.) *Readings in Philosophy of Psychology*, Vol. 1, pp. 173-5.

[118] Block, N. (1980c), 'Troubles with Functionalism' in N. Block (ed.) *Readings in Philosophy of Psychology*, Vol. 1, pp. 268-305.

[119] Ibid. Pp. 231-9.

1. If the machine is in state S_1 and receives input 0, then write 1 on the cell of the tape and move head left by one cell along the tape.
2. If the machine is in state S_2 and receives input 1, then write 1 on the cell of the tape and stay.

In cases where the machine table does not have any rule that is applicable to the machine state (e.g. if there are only two numerals 0 and 1 on the tape, and the table contains a rule that says ‘ receive B then write 1’), the Turing machine *halts*.

In general, a Turing machine has a head that travels back and forth along the infinitely long tape to read, write, and perhaps halt. Given this behaviour pattern, what exactly does it mean to say a Turing machine is like a can opener, in respect of being multiply realisable? I think the answer to this question is two-fold.[120] First, it is the multiple realisability of the tape. It is widely accepted that the tape on a Turing machine can be made of any material, for example an infinitely long roll of toilet paper with symbols written on it, or a gigantic spaghetti noodle with the same set up. Likewise, the head of the machine can be in any form, such as a dial or a slider. Of course, these would have to be idealised realisations, given the fact that Turing-machine tape is infinite and thereby idealised. Putnam expresses his endorsement of this view in his 1967 paper ‘The Mental Life of Some Machines’:[121]

> We can still draw no inference whatsoever to the physical-chemical composition of T_1, for the reason that the *same* Turing Machine (from the standpoint of the machine table) may be physically realized in a potential

[120] Turing, to whom history often fails to give enough credit, smartly confirms what could be seen as the third meaning of the ‘multiple realisability’ of the Turing machine. In the following passage (Turing 1936, p. 252), he notes that the mental state of a *computer* (which at that time denoted a person who computes) when computing can be multiply realised by the computing machine state’s:

> We may now construct a machine to do the work of this computer. To each state of mind of the computer corresponds an “*m*-configuration” of the machine. The machine scans *B* squares corresponding to the *B* squares observed by the computer. In any move the machine can change a symbol on a scanned square or can change any one of the scanned squares to another square distant not more than *L* squares from one of the other scanned squares. The move which is done, and the succeeding configuration, are determined by the scanned symbol and the *m*-configuration. The machines just described in § 2, and corresponding to any machine of this type a computing machine can be constructed to compute the same sequence, that is to say the sequence computed by the computer.

[121] Putnam, H. (1967a), ‘The Mental Life of Some Machines’ in H. Castaneda (ed.) *Intentionality, Minds, and Perception*, pp. 177-200.

> infinity of ways. Even if in fact a machine belonging to our community prefers A to B when and only when flip-flop 57 is on, this is a purely contingent fact. Our machine might have been exactly the same in all "psychological" respects without consisting of flip-flops at all.[122]

Apart from this common belief, following Lawrence Shapiro's line of thought I think the multiple realisability of the Turing machine has a second meaning.[123] Utilising the concept of Turing-machine tables, we can theoretically construct the following abstract machine table to unpack Turing's idea. Recall that all Turing-machine tables can be represented in the form of conditional statements:

> If the machine is in state S_1 and receives input ❦, then it emits output ❡ and stays in S_1.

The meaning of ❦ and ❡ are entirely irrelevant to how the machine performs. In other words, Turing machines are purely syntax-dependent and semantic-independent. Any machines that share table A (or any other machine tables) will behave the same way, or as Turing puts, 'compute the same sequence'[124].

I have just explained why, on Putnam's account, Turing machines can be multiply realised. Let us now turn to the second and final step of Putnam's argument. Namely, the step where he carries the lesson over from Turing machines to the human mind or brain. Having in mind Turing's groundbreaking brainchild, namely, the multiply realised Turing machine, Putnam wants to extend the same feature to minds by contending that mental states and Turing-machine states are categorically alike. Putnam presents his reasoning in the following passage:

> To obtain such an analogue, let us identify a scientific theory with a 'partially-interpreted calculus' in the sense of Carnap. Then we can perfectly well imagine a Turing machine which generates theories, tests them (assuming that it is possible to 'mechanize' inductive logic to some degree), and 'accepts' theories which satisfy certain criteria (e.g. predictive success). In particular, if the machine has electronic 'sense organs' which enable it to 'scan' itself while it is in operation, it may formulate theories concerning its own structure and subject them to test. Suppose the machine is in a given state (say, 'state A') when, and only when, flip-flop 36 is on. Then this statement: ' I am in state A when, and only when, flip-flop 36 is on', may be one of the theoretical principles concerning its own structure accepted by the machine … Now all of the usual considerations for and against mind-body identification can be

[122] Putnam (1967a), p. 187. Author's italics.
[123] Shapiro, L. (2004), *The Mind Incarnate*, pp. 14-5.
[124] Turing (1936), p. 14.

> paralleled by considerations for and against that state A is in fact *identical* with flip-flop 36 being on.[125]

Citing Rudolf Carnap, Putnam firstly tries to offer a foundation upon which two different scientific theories can be regarded as analogues of one another. This Carnapian foundation that Putnam has in mind can be found in Carnap's 1953 article 'The Interpretation of Physics'[126] in which he explains that mathematical geometry and physics are kindred theories:

> By the interpretation, the theorems of the calculus of mechanics become physical laws, i.e., universal statements describing certain features of events; they constitute physical mechanics as a theory with factual content which can be tested by observations. The relation of this theory to the calculus of mechanics is entirely analogous to the relation of physical to mathematical geometry.[127]

For Putnam, this Carnapian argument readily generalises and is thereby easily applied to Turing-machine theory and the theory of mind and body. To exemplify the analogous nature of the two theories, Putnam notes that both theories are prone to Chisholm's problem of criterion.[128] Roderick Chisholm, who famously remarks upon the definability of mental states, draws attention to the fact that a certain type of mental state, say 'state X', cannot be defined without mentioning other mental states say 'state Y'.[129] Turing machine theory exhibits the same shortcoming: a certain Turing-machine state, cannot be defined without the mention of other Turing-machine states. In short, Chisholm's problem manifests in both theories as they both involve inter-defined states. To remedy this, Putnam establishes that the following PB and PT are both statements asserting *theoretical identification*:

(PB) Mental state X = brain state Y

(PT) Turing-machine state A = flip-flop 36

For theoretical identifications, inter-definitions can still exhibit uniqueness.[130] The precise recipe is the use of Ramsey sentences – a technique that Lewis also employed in his

[125] Putnam (1975b), p. 363.

[126] Carnap, R., (1953) 'The Interpretations of Physics' in H. Feigl & M. Brodbeck (eds.), *Readings in the Philosophy of Science*, pp. 309-18.

[127] Ibid. P. 309.

[128] Rey (1997), p. 172.

[129] Chisholm, R. (1957), *Perceiving: a Philosophical Study*, pp. 43 -66.

[130] Putnam (1975b), pp. 379-382.

defence of the identity theory. Thus, Chisholm's problem does not post a threat to Turing machine theory and mind-body theory.

Of course, this does not exhaust the scope of the link between the two theories. Putnam explains in detail why mental states can be conceptualised as Turing-machine states by saying:

> It is interesting to note that just as there are two possible descriptions of the behavior of a Turing machine – the engineer's structural blue print and the logician's 'machine table' – so there are two possible descriptions of human psychology. The 'behavioristic' approach … This corresponds to the engineer's or physicist's description of a physically realized Turing machine. But it would also be possible to seek a more abstract description of human mental processes, in terms of 'mental states' ... This description, which would be the analogue of a 'machine table', it was in fact the program of classical psychology to provide!
> ……
>
> The analogy which has been presented between logical states of a Turing machine and mental states of a human being, on the one hand, and structural states of a Turing machine and physical states of a human being, on the other, is one that I find suggestive.[131]

The proposal that Putnam puts forward here is to define mental states by a 'mental machine table' that governs the operation of human psychology. Just as symbols on the tape of a Turing machine serve as inputs and outputs, mental states have their corresponding inputs and outputs in sensory stimulations and behaviour, respectively. Consequently, just as Turing-machine tables contain instructions that specify the relations between inputs and outputs, mental-state tables specify how to behave given a certain stimulation. What Putnam proposes is thus 'a machine table for a human', as summed up by Block.[132]

By providing reasons for the analogous nature of Turing-machine theory and mind-body theory, the rest of Putnam's argument follows uncomplicatedly. Just as 'flip-flop 36' cannot be identical to Turing machine's 'state A' due to the multiple realisability of Turing-machine states, mental state x cannot be identical to brain state y, for mental states are also multiply realised. Therefore, mind-brain identity theory is false. In a nutshell, Putnam's argument against identity theory as presented in 'Minds and Machines' can be formalised as follows:

[131] Putnam (1975b), pp. 372-383.

[132] Block (1980a), p. 178.

Argument T

1. Mental states are analogous to Turing-machine states.
2. Turing-machine states are multiply realised.
3. Therefore, mental states are multiply realised.
4. Therefore, mind-brain identity theory is an incorrect theory.

Argument T has been widely discussed and criticised by many, including Putnam himself who went on to reclaim his own functional theory of mind. In his 'Philosophy and Our Mental Life', Putnam rejects premise 1 on the basis that Turing-machine states are not lucid models of mental states.[133] The objection was later strengthened in his 1988 book *Representation and Reality*, in which he elaborates on the mismatch between the two kinds of states. Specifically, he argues that in specifying mental states, one mustn't define them purely in terms of functional descriptions, because social and environmental aspects must also be taken into account, whereas in specifying machine states, isolating the social and environmental consideration is acceptable. In short, at least on the view of the later Putnam, Turing-machine states are functional states, but mental states are not exclusively so. This mismatch suffices to falsify premise 1 and as Putnam himself claims, his earlier argument as a whole.[134]

In presenting a refined version of the multiple realisability argument, Block and Fodor (1972)[135] also attempt to break Putnam's purported link between Turing machines and minds. Block and Fodor's rejection of premise 1 requires distinguishing the difference between what they call 'machine table states' and 'computational states' of an automaton. According to them, the former is a sub-set of the latter. The 'computational states' of an automaton can refer to any states that are defined in terms of inputs and outputs, and 'machine table states' are just one kind of these states. Making this distinction allows us to discover another mismatch in Putnam's analogy: that if mental states are to be conceived as machine states at all, they seem to be more analogous to 'computational states' than Turing-machine table states.[136]

[133] Originally presented in 1973, and later published as Putnam, H. (1975c), 'Philosophy and Our Mental Life' in *Mind, Language and Reality*, Vol. 2, pp. 291-303.

[134] Putnam, H. (1988), *Representation and Reality*, pp. 73-89.

[135] Block, N. & Fodor, J. (1972), 'What Psychological States Are Not' in *The Philosophical Review*, Vol. 81, No. 2, pp. 159-81.

[136] Ibid. Pp. 178-9.

Another challenge to premise 1 can be found in Kim (1992, 2011).[137] Kim notes a potentially undermining implication of the Turing machines—minds analogy. If mental states are to be entirely conceptualised as Turing machines, then two subjects with identical mental states are deemed to be realisations of the same Turing machine. If two subjects, minds or Turing machines, are instantiations of the same Turing machine, then they must share all their Turing-machine states. It follows that if two subjects share *any* mental states (Turing-machine states), then they must share *all* mental states (Turing-machine states). For example, if Jack Copeland and I have *one* belief in common, say that 'Donald Trump is the current president-elect of the U.S.', Jack and I must also share *every* other belief. As Kim remarks, to endorse Argument T's premise 1 will be to endorse this absurd consequence.[138]

As shown, discussions of Argument T have been well covered by existing literature. I, however, want to explore another aspect of the argument. It is not difficult to realise that Argument T is not deductively valid since it has the form of an argument from analogy. Nevertheless, this shall not be fatally worrying for Putnam as long as Argument T is convincing. Unfortunately it fails to be a convincing argument unless the transition from sub-conclusion 3 to conclusion 4 is explained. In other words, the mere multiple realisability of mental states does not analytically falsify mind-brain identity statements. What Putnam needs is thus an extra premise that claims 'if mental states are multiply realised then it is incorrect to identify mental states with brain states', and thereby revamping the argument as follows:

Argument T*

1. Mental states are analogous to Turing-machine states.
2. Turing-machine states are multiply realised.
3. Therefore, mental states are multiply realised.
4. If mental states are multiply realised, then it is incorrect to identify mental states with brain states.
5. Therefore, mind-brain identity theory is an incorrect theory.

[137] Kim, J. (1992), 'Multiple Realization and the Metaphysics of Reduction' in *Philosophy and Phenomenological Research*, Vol. 52, No. 1, pp. 1-26.
Kim, J. (2011), *Philosophy of Mind*.

[138] Kim (2011), p. 152.

This is, I contend, Putnam's overall argument against identity theory. Putnam's multiple realisability objection must have this form should it be considered a pervasive one. In so doing, Putnam also has to justify this newly added premise 4 – which he did not do in 'Minds and Machines'. Seven years after 'Minds and Machines' was originally published, 'Psychological Predicates'[139] was out and Putnam presents in this later paper his full argument for why he thinks premise 4 is true. I will call that argument the 'octopus argument'. From 4.2 onwards, my thesis will turn to analyse the 'octopus' argument and introduce my original reply to it. My reply, if successful, would render premise 4 false and thereby post a threat to Putnam's overall argument. Before we look at that, let me say a few words about functionalism – the seemingly inevitable entailment of the multiple realisability of minds.

4. 1. 2. Multiply realised minds and functional minds

Suppose Putnam's sub-conclusion 3 (of Argument T*) is right, that is, multiple realisability of minds is assumed, a question quickly emerges: in virtue of what is it true that two agents, say a human and an octopus, can have the same mental state, say pain? Let me start by rephrasing the question with the can opener analogy: what do the rotating wheel openers and my Swiss tool have in common that make them both physical realisations of can openers? Surely the answer cannot be the rotating wheel design, because the Swiss tool does not have a rotating wheel. By the same token, what makes the human and the octopus both in pain surely cannot be C-fibre firing, as purported by identity theorists, because octopuses do not have C-fibres. In a nutshell, given the multiple realisability of mental states, what is the nature of mental states such as pain? Assuming that by establishing the multiple realisability of mental states one is able to falsify identity theorists' answer to this question[140], one also has to provide one's own answer to it. Putnam's answer to this question *was*[141] what has been coined *functionalism*[142], as he asserts here:

[139] Putnam, H. (1967b), 'Psychological Predicates' in W. H. Capitan & D. D. Merrill (eds.) *Art, Mind, and Religion*, pp. 37-54.

[140] I will argue against this assumption from 4.2 onwards.

[141] As mentioned, Putnam later revised his view in his (1988). See the block on the next page.

[142] Similar views have been offered by many others, notably in Fodor (1968, 1974), Lewis (1972, 1980b), and Shoemaker (1975, 1981). However, these alternative versions of functionalism are not motivated by the 'octopus argument' and are thereby outside my target zone.

> Many years ago, I published a series of papers in which I proposed a model of the mind which became widely known under the name “functionalism.” According to this model, psychological states (“believing that *p*,” “desiring that *p*,” “considering that *p*,” etc.) are simply “computational states’ of the brain. The proper way to think of the brain is as a digital computer. Our psychology is to be described as the software of this computer—its “functional organization.”[143]

A good way to understand this functional account of mind is to look at the functional account of can openers so described. What the rotating wheel opener and the Swiss tool have in common is that they both can accomplish the same job, namely to take the lid off a tin can. Both tools can fulfil the role of opening cans. It is in virtue of this role that they are can openers. Factors such as their design, the materials they are made out of, are entirely irrelevant to whether or not they are can openers. The only factor that matters is whether or not they can fulfil the role. Likewise, all that matters for what counts as pain is whether or not there is an internal state that occupies the role of pain. Physical features of this internal state, such as what this realiser is made out of (i.e. C-fibre or not), are irrelevant factors. Also, it must be highlighted here that this functional account applies not only to sensations like pain and hunger, but also to other mental states including propositional attitudes, as Putnam clarifies in the above quotation.

I do not wish to reiterate this nowadays-dominant approach to the mind-brain problem, because the core of my argument, which I will soon reveal, does not target functionalism directly. Rather, my target is the foundation upon which functionalism is built – namely, Putnam’s multiple realisability argument. Nevertheless, it is worthwhile to acknowledge here that the functionalist approach has been subjected to some very famous criticism since its first inception, such as the ‘China brain’ argument and the ‘Blockhead’ argument, both due to Block.[144]

4. 2. Octopuses and their pain

Let us now examine Putnam’s argument for premise 4 of Argument T*. First, it is worthwhile to revisit Putnam’s own words. To date, the clearest presentation of this ‘octopus argument’ remains that given by Putnam in his 1967 paper ‘Psychophysical

[143] Putnam (1988), p. 74.
[144] Block (1980c), pp. 275-80.

Predicates', where he criticises the 'theoretical identification of mental states with physical states':

> Consider what the brain-state theorist has to do to make good his claims. He has to specify a physical-chemical state such that *any* organism (not just mammal) is in pain if and only if (a) it possesses a brain of a suitable physical-chemical structure; and (b) its brain is in that physical-chemical state. This means that the physical-chemical state in question must be a possible state of a mammalian brain, a reptilian brain, a mollusc's brain (octopuses are mollusca, and certainly feel pain), etc. …… Finally, the [brain-state] hypothesis becomes still more ambitious when we realize that the brain-state theorist is not just saying that *pain* is a brain state; he is, of course, concerned to maintain that *every* psychological state is a brain state. Thus if we can find even one psychological predicate which can clearly be applied to both a mammal and an octopus (say 'hungry'), but whose physical-chemical 'correlate' is different in the two cases, the brain-state theory has collapsed. It seems to me overwhelmingly probable that we can do this.[145]

4. 2. 1. The argument proper

What is Putnam claiming here? First, by saying:

> Thus if we can find even one psychological predicate which can clearly be applied to both a mammal and an octopus (say "hungry")

Putnam notes that the following identity claim appears true, where P_h denotes human pain and P_o denotes octopus pain:[146]

(I) $$P_h = P_o$$

Putnam asks us to imagine two creatures, a human ($_h$) and an octopus ($_o$), who share the same mental state or psychological predicate. Hence, $P_h = P_o$ represents the claim that the mental states of the human being denoted by P_h are of the same type as the mental states of the octopus denoted by $P_{o.}$ The identity in question is type identity, not token identity, as explained in Chapter 1. Furthermore, (I) appears true to Putnam because, as he remarks, 'octopuses certainly feel pain', and because human pain and octopus pain, both being pain, will have the same essential experiential feel, and thus be of one and the same phenomenal state type.

[145] Putnam (1967b), pp. 44-5.

[146] For convenience, I choose pain instead of Putnam's own example of hunger. I hope it is obvious that the swap is of no significant philosophical importance.

Secondly, Putnam notes that, very likely, the physical-chemical correlates of octopus pain will not match the physical-chemical correlates of human pain. Clearly, he is proposing a non-identity claim between the two physical-chemical 'correlates'. To elucidate Putnam's point, let us firstly list the physical-chemical 'correlates' of P_h and $P_{o.}$ According to identity theory, the mind is identical to the brain, and mental states, say pain, are nothing but a specific type of brain state, namely, C-fibre firing. That is, the following identity holds, where C denotes C-fibre firing (or whatever physical-chemical correlate human pain is purportedly identical to[147]):

(II) $P_h = C$

Here, (II) means every given token of P_h is a token of C.

On the other hand, since octopuses do not have C-fibres at all, we cannot call their pain correlate C-fibre firing. Instead, let us call them 'jelly firing' in recognition of the jelly-ish texture of the octopus. Let J be the physical-chemical correlate of octopus pain; thus, according to identity theory, we have

(III) $P_o = J$

Here, again, the identity holds between every token of the type P_o and every token of the type J.

Having established what each of the physical-chemical correlates is, we can state the non-identity that Putnam attributes to them. C and J have nothing relevant in common. Even the hard-core identity theorist might have agreed that J- and C-fibre firing are very different states, since octopuses and human beings have radically dissimilar brain. (Octopuses do not have C-fibres at all.) So:

(IV) $C \neq J$

Here, (IV) represents the claim that no token of C-fibre firings is identical to (i.e. the same thing as) any token of a jelly-firing.

Under standard logic, (I), (II), (III), and (IV) form an inconsistent tetrad. It therefore seems that at least one of them must be rejected. The difficulty is which. Putnam rejects (II) and (III) and concludes that '… the brain-state theory has collapsed.'

[147] Traditionally, identity theorists have suggested that this brain state that is identical to pain is a state of 'C-fibre firing', although nothing of philosophical importance depends on this being true. See 6.2 for my discussion on this.

Conclusion. $P_h \neq C$

The above is a general outline of Putnam's argument, but the exact form of the argument can, however, be understood in a number of different ways. That is to say, there are many ways to formalise the argument. In the next Chapter, I will introduce a novel attack against the argument based on a specific formalisation. For now, let us consider the following formalisation, which does not involve all of (I)-(IV) in its premises:[148]

1. If identity theory is correct, then $(P_h = P_o) \rightarrow (C = J)$.
2. $P_h = P_o$
3. $C \neq J$

----

4. The identity theory is incorrect.

This reading of Putnam's argument targets an implication of identity theory – *physical distinctness entails phenomenal distinctness.* Two agents can differ in phenomenal attributes only if they differ in physical attributes. In our example, a human and an octopus would have different phenomenal states, i.e., one in pain and the other not, only when they have different physical states, hence, one have C-fibre firings and the other doesn't. By contraposition, if the pair shares the same states, for instance both are in pain, then their corresponding physical states must be the same. In other words, *phenomenal sameness entails physical indiscernibility* –

$$(P_h = P_o) \rightarrow (C = J)$$

This is premise 1, the conditional premise. Following Putnam's rationale, the following premises proceed to assert (I), the claim that mental states such as pain are multiply realisable across species:

$$P_h = P_o$$

and (IV), that it is obvious that C-fibre firings are not identical to jelly firings:

$$C \neq J$$

Combining premise 2 and 3 we will have:

$$(P_h = P_o) \& (C \neq J)$$

[148] Specifically, this formalisation omits (III) $P_o = J$.

which is the negation of $(P_h = P_o) \rightarrow (C = J)$**,** the consequent of the conditional premise. Then by *modus tollens*, the antecedent in the conditional premise is falsified. Thus, identity theory is incorrect.

4. 2. 2. Alleged objections to the 'octopus' argument

Since its inception, this 'octopus' argument has been subject to critical replies. By and large, there have been two types of replies, stemming from two different sorts of considerations. The first type *does not* criticise Putnam's argument *per se*, and instead questions its force against identity theory. This reply alleges that Putnam's argument is attacking a straw man. The second type questions the truth of the premises, with a predominant focus on Putnam's first claim – (I). To begin with, Plantinga notably provides the first published diagnosis of the straw-man fallacy to Putnam's argument.[149] Plantinga acknowledges that the theory which Putnam purportedly attacks is indeed a false one, but denies that it is identity theory:

> I doubt, however, that the identity theorist would wish to dispute Putnam's conclusion; for I am inclined to think that when he says *being in pain is really being in a certain neurological state S′*, the identity theorist does not mean to assert the identity of any universals at all. What he means to assert is that every instance of the universal *being in pain* is contingently identical with some instance of the universal *possessing neurological state S′*; and that every instance of the universal *preferring A to B* is contingently identical with some instance of the universal *possessing S* … This theory is not open to any of the objections Putnam deploys against modern materialism. I may be wrong in supposing that this is the theory the modern materialists mean to put forth; nonetheless, Putnam should have considered it, for the words in which he himself states the identity theory are plainly open to this interpretation.[150]

An identity theorist need not argue with Putnam, says Plantinga, because there is a possibility that her theory is *not* a type-type identity theory. This solution offered by Plantinga depends to a great extent on a fundamental stipulation, one that Plantinga himself concedes could be wrong: that the theory in question bears an uncanny resemblance to a token-token theory. As Plantinga sums up above, according to the token-token mind-brain identity theory, to say that my pain is identical with C-fibre firing is to say that my pain at a certain time is identical with my C-fibre firing at that certain

[149] Plantinga, A. (1967), 'Comments' in H. Castaneda (ed.) *Intentionality, Minds, and Perception*, pp. 201-5.
[150] Ibid. Pp. 203-4. Author's italics.

time. It does not make generalisation such as 'pains are C-fibre excitations'. Although this has been regarded by many as a weak version of identity theory, it is indeed immune to Putnam's objection.

But is Plantinga right about this stipulation of identity theory? There are two considerations that point toward the negative answer. First, Putnam refutes the idea that there can be a genuine contingent identity of *instances* of mental states and brain states. He argues that Plantinga's assertion that '*being in pain* is contingently identical with some instance of the universal *possessing neurological state S*', could be formalised as the following general statement:

(AT) For every agent *A*, temporal instant *t*, *A* is in a mental state at *t* iff *A* is in a brain state at *t.*

AT, as Putnam emphasises, amounts to a mind-brain correlation account of minds, which is an even weaker notion than the token-token theory, and most importantly it is far from what identity theorists would agree with.[151] This leads to the second consideration as to why Plantinga is wrong. Even if the correlation account is out of the picture, a genuine token-token theory is still some distance away from the identity theory that Putnam's argument supposed to reject. By 'brain-state theorist', Putnam clearly refers to the likes of Smart and Lewis who undoubtedly does not embrace the token-token interpretation. So, Plantinga's criticism that Putnam is attacking a straw man is in fact a straw-man argument itself.

Coming from a different angle, Lewis offers another criticism that aims to render the 'octopus argument' a straw man reasoning.[152] In 'Review of Putnam',[153] Lewis argues that what identity theorists advocate must be species-bound identity should their theory be considered a plausible one:

> A reasonable brain-state theorist would anticipate that pain might well be one brain state in the case of men, and some other brain (or nonbrain) state in the case of mollusks.[154]

[151] Putnam, H. (1967c), 'Rejoinder' in H. Castaneda (ed.) *Intentionality, Minds, and Perception*, pp. 206-13.

[152] Lycan (1974) holds that Lewis' and Plantinga's objections to Putnam are the same. I disagree on the basis that Lewis' appeals to species-bound identity and Plantinga's doesn't.

[153] Lewis, D. (1980b), 'Review of Putnam' in N. Block (ed.) *Readings in Philosophy of Psychology*, Vol. 1, pp. 232-3.

[154] Ibid. P. 233.

Will this move save identity theory from Putnam's argument? *Prima facie*, the answer is yes. By removing the tenet that humans and octopuses have the exact same kind of pain, a refined (restricted) identity theory thereby makes no commitment to Putnam's (I) – the proposition that fuels the 'octopus argument'. On closer inspection, however, Lewis' point starts to show weaknesses. Elaborating on the species-bound theory, Lewis writes:

> No mystery: that is just like saying that the winning number is 17 in the case of this week's lottery, 137 in the case of last week's. The seeming contradiction (one thing identical to two things) vanishes once we notice the tacit relativity to context in one term of the identities. Of course no one says that the *concept* of pain is different in the case of different organisms (or that the *concept* of the winning number is different in the case of different lotteries). It is the *fixed* concept expressed by 'pain' that determines how the denotation of 'pain' varies with the nature of the organism in question. Moral: the brain-state theorist cannot afford the old prejudice that a name of a necessary being (such as a state) must name it necessarily and independently of context.[155]

Here Lewis undoubtedly treats mental state terms as anything but rigid terms. The lottery analogy even seems to suggest that he takes mind-brain identity to be statements involving definite descriptions. In short, the referent of 'pain' varies from species to species, according to this Lewisian view of identity theory. Interpreted this way, identity theory is prone to scrutiny from anyone who doesn't adhere to the theory of the meaning of natural kind terms that Lewis hints at. Nevertheless, language matters aside, the species-bound account just doesn't seem to possess adequate explanatory power in addressing an important question in philosophy of mind. As I expressed earlier, one of the key enquiries that a credible theory of mind should answer is 'Given a type of mental state M, what is in common to all Ms in virtue of which they are Ms?' Echoing Lycan's comment, I hold that the species-bound theory is untenable because it offers little insight regarding the aforementioned question.[156] 'What's common to all pains?' The answer from the species-bound identity theorist would be 'There is nothing in common!' At this point, the cost of evading Putnam's objection is simply too great.

Despite possessing what I believe to be obvious shortcomings, this restricted rendition of identity theory has been taken up by many as the preferred option to sidestep the

[155] Lewis (1980b), p. 233.

[156] Lycan, W. (1974b), 'Mental States and Putnam's Functionalist Hypothesis' in Australasian Journal of Philosophy, 52:1, pp. 48-62. Lycan's view is stronger than mine since he thinks that the species-bound theory in effect doesn't address the nature of mental states at all, and for this reason he thinks the theory is equivalent to a version of eliminative materialism.

‘octopus’ argument.[157] More recently, another approach has been proposed to question Putnam’s argument. In his (2004), Shapiro explains that the actual realisations of a mental state are often quite unpredictable, and confirming the sameness of psychological states is problematic, given the different realisations thereof. In other words, it is possible for differently realised mental states to have nothing in common on the psychological level.[158] Sharing Shapiro’s position are Bechtel & Mundale (1999)[159] and Couch (2004)[160] who claim that the multiple realisability of mental states occurs at a level of abstraction, but whether or not sameness obtains at the psychological level is entirely subject to empirical research. In a nutshell, this type of criticism challenges the truth of (I) head-on by demanding empirical evidence of the multiple realisability of mental states.

This concludes my outline of putative objections to Putnam’s ‘octopus’ argument. Before I present my own reply, I must highlight here that *all* of the objections I have so far discussed take for granted that the argument is valid. These putative criticisms *do not* investigate the inference steps Putnam’s argument commits to. For this reason, my criticism of Putnam is radically different to theirs.

[157] Notably by Braddon-Mittchell & Jackson (2007), Kim (1972, 2011) and Polger (2002).

[158] Shapiro (2004), pp. 35-66.

[159] Bechtel, W. & Mundale, J. (1999), ‘Multiple Realizability Revisited: Linking Cognitive and Neural States’ in *Philosophy of Science*, 66, pp. 175-207.

[160] Couch, M. (2004), ‘Discussion: A Defense of Bechtel and Mundale’ in *Philosophy of Science*, 71, pp. 198-204.

Chapter 5. Indeterminate identity and Putnam's octopus

5. 0. Outline of Chapter 5

The previous Chapter looked at how Putnam argues from the multiple realisability of the mental to the falsity of identity theory:

Argument T*

1. Mental states are analogous to Turing-machine states.
2. Turing-machine states are multiply realised.
3. Therefore, mental states are multiply realised.
4. If mental states are multiply realised, then it is incorrect to identify mental states with brain states.
5. Therefore, mind-brain identity theory is an incorrect theory.

I emphasised that premise 4 requires an independent argument, which I dubbed the 'octopus' argument. The last Chapter also examined some putative objections to the 'octopus argument' that solely target the soundness (assuming its validity). The present Chapter develops a new reply to the 'octopus' argument that targets the argument's validity. 5.1 will propose and explain a particular way of interpreting Putnam's argument in order to expose one of its vital inference rules. My refutation of Putnam focuses on the inapplicability of this inference rule. There are three premises in my argument, which will be discussed in 5.2, 5.3, and 5.4, respectively.

5. 1. Putnam's argument reconstructed

As sketched in 4.2.1, the conventional way of interpreting Putnam's argument amounts to the following, which does not involve all of (I)-(IV) in its premises:

1. If identity theory is correct, then ($P_h = P_o$) → ($C = J$).
2. $P_h = P_o$
3. $C \neq J$

---

4. The identity theory is incorrect.

5. 1. 1. The seemingly inconsistent tetrad

For my purpose, there is another formalisation that better serves to express the weakness in the argument. I contend that the better way to highlight the tension between identity theory and multiple realisability is to reframe the argument so that it includes all of (I)-(IV):

(I)	$P_h = P_o$
(II)	$P_h = C$
(III)	$P_o = J$
(IV)	$C \neq J$

It needs little explanation that this group of four statements cannot be jointly true, meaning that if three of them are true the remaining one must be false. For example, accepting the truth of (I), (II), and (III) will imply rejecting (IV). Likewise, accepting any three will give you the negation of the remaining one. Consequently, to avoid the inconsistency you need to reject at least one statement of the tetrad (rejecting more than one will, of course, do the job just fine). It follows that there are four possible ways to retain consistency, these being to each of the four statements. Let us recap what each component of the tetrad amounts to and have a quick overview of our options:

(I)	$P_h = P_o$	Multiple realisability
(II)	$P_h = C$	Identity theory
(III)	$P_o = J$	Identity theory
(IV)	$C \neq J$	Plain truth

Option A: As explained above, both (I) and (IV) are part of Putnam's initial setup as he argues for multiple realisation and he also notes that human pain's and octopus pain's ‘physical-chemical “correlate” is different in the two cases’. So Putnam’s way out is to get rid of (II) and (III), the statements of identity theory.

Option B**:** Alternatively, one can retain (II), (III), and (IV), and subsequently reject (I). This solution sees the idea of multiple realisability as the weakest link in the tetrad. As I explained in 4.2.2, Kim, Lewis, and Plantinga take this position; Braddon-Mitchell and Jackson arguably take it too.

Option C**:** The third option is to keep (I), (II), and (III), and reject (IV) (i.e. reject the seemingly plain truth of $C \neq J$.

No doubt, these three ways out (if they are correct) can lead us out of the inconsistency. However, none of them can reconcile identity theory and multiple realisability. They either endorse one idea and reject the other (option A and B), or appeal to a weaker notion of identity that hints at backdoor functionalism (option C). In fact, rejecting the components of the tetrad will get us nowhere near accepting both identity theory and multiple realisability. To crystalise this point, let us spell out what exactly we are trying to resolve here. The goal of reconciling identity theory and multiple realisability means fulfilling two conditions:

Condition (A): To find a way to assign truth to (I), (II), and (III).

Condition (B): The truth-values assigned to the proponents of the tetrad still need to be consistent.

Satisfying (A) gives us two possible combinations of truth-values:

Combination one

(I)	$P_h = P_o$	TRUE
(II)	$P_h = C$	TRUE
(III)	$P_o = J$	TRUE
(IV)	$C \neq J$	TRUE

Combination two

(I)	$P_h = P_o$	TRUE
(II)	$P_h = C$	TRUE
(III)	$P_o = J$	TRUE
(IV)	$C \neq J$	FALSE

To satisfy (B), however, we will have to cut Combination One, for it is the one that yields inconsistency. This leaves us with Combination Two, which resembles the view advocated by taking option C. As sketched above, option C amounts to backdoor functionalism, which does reconcile multiple realisability and the identity theory, but unfortunately only at the price of weakening the identity theory. For this reason, Combination Two is ruled out as well.

The message is clear: to fulfill both (A) and (B), specifically, to attain consistency of the tetrad and claim joint truth of (I), (II), and (III), one shall not go the usual route of rejecting the remaining statement. But wait a minute! Isn't the only way out of an inconsistent tetrad to reject one or more components? The situation at hand is really puzzling:

Step one: we start by having four statements, (I), (II), (III), and (IV).

Step two: when grouped together these statements appear to form an inconsistent tetrad.

Step three: we want to find a way accepting (I), (II), and (III).

Step four: we learn from standard logic that we can do this by rejecting (IV).

Step five: However, (IV) is not rejectable because a) it seems to be a plain truth, and b) in rejecting it we would fall prey to established objections and are derailing from the common philosophical notion of 'identity'.

Step six: we are stuck!

5. 1. 2. Transitivity and Putnam's argument

Is there a new solution to this? I will show you there is! So far we have been building our investigation on the assumption that the second step is undeniable – that the tetrad really is inconsistent. If this assumption is wrong, then the whole problem is solved - we would not have to argue for the truth or falsity of each components of the tetrad just to restore the consistency to the tetrad, because the tetrad would be regarded as consistent in the first place! This likely solution requires an insight into what makes the tetrad appear inconsistent, which in turn calls attention to the logical form of these four statements – identity statements.[161]

Statements (I), (II), and (III) share the form of *x* = *y* – a form of statement that can be read in various ways in English. Simply put, it says *x equals* y, or *x is identical to y*. More precisely, it aims to tell *something called x is one and the same thing as something called y*, which can also be understood as saying *there is one thing which can be known as 'x' in such and such circumstances and as 'y' in such and such circumstances.* Alternatively, we can say *x bears a relation to y that x really is just y.* Under this reading, *x* = *y* stresses a relation of *being the same as* between *x* and *y.* Let us conjure up an example using this reading:

> 'Hesperus = Phosphorus' reads 'Hesperus bears the relation to Phosphorus of being the same thing as Phosphorus'.

- Symmetry

Now consider swapping the terms flanking the identity symbol.

> 'Phosphorus = Hesperus' reads 'Phosphorus bears the relation of being the same as Hesperus'.

Notice that the swapping does not change the truth-value of the statement. In other words, the order of the flanking terms has no impact on the truth condition of identity statements. Hence, the metaphysical messages conveyed through *x*=*y* and *y*=*x* are the same message. Moreover, I think it is pretty obvious that we do not gain extra knowledge from the swapping. 'Hesperus = Phosphorus' has the same epistemic value as 'Phosphorus = Hesperus'. Neither statement has cognitive significance over the other. In other words, it is trivial that the two statements are interchangeable. This is because the relation of *being the same thing as* is a *symmetrical* relation. Unlike examples in which substitutions of flanking terms occur (e.g. from 'Hesperus = Phosphorus' to 'Hesperus = Venus'), the symmetry move does not involve the addition of new intension to the original flanking terms. Instead it merely swaps their order –from the left to the right, and from the right to the left (e.g. from 'Hesperus = Phosphorus' to 'Phosphorus = Hesperus'). Thus, the new statement has introduced nothing epistemically or metaphysically new.

[161] Strictly speaking, there are two forms here – identity statement in the cases of (I), (II), and (III), and non-identity statement in the case of (IV).

In general, those relations that are symmetrical are defined as follows:

Symmetry of Relations

Definition. A Relation *R* is symmetrical iff for any two objects (or terms) *a and b*, *a* bears *R* to *b* iff *b* bears *R* to *a.*

Identity, as we discussed, is a symmetrical relation. Formally put:

Symmetry of Identity

Definition. If $x=y$, then $y=x$.

$x=y \vdash y=x$

- Transitivity

Another feature of identity relations is transitivity. The following are three true statements:

(HP). Hesperus = Phosphorus.

(PV). Phosphorus = Venus.

(HV). Hesperus = Venus.

Suppose one only knows (HP) and (PV), one's knowledge of the two statements will guarantee that (HV) is true, because the fact that Hesperus is the same thing as Phosphorus and Phosphorus is the same thing as Venus guarantees that Hesperus is the same thing as Venus. Relations are said to be transitive if they obtain such a property.

Transitivity of Relations

Definition. A Relation *R* is transitive iff for any three objects (or terms) *a, b, and c*, if *a* bears *R* to *b*, and *b* bears *R* to *c*, then *a* bears *R* to *c.*

Identity, as illustrated, is a transitive relation. Formally put:

Transitivity of Identity

Definition. If $x=y$, and $y=z$, then $x=z$.

$x=y, y=z \vdash x=z$

Identity statements are transitive and symmetrical. But what do symmetry and transitivity have to do with Putnam's multiple realisability argument against identity theory then? Let us recall the aforementioned tetrad:

(I) $P_h = P_o$

(II) $P_h = C$

(III) $P_o = J$

(IV) $C \neq J$

Since (I), (II), and (III) are identity statements and we just learnt that identity is symmetrical and transitive, we can now apply these two rules to the tetrad and see what happens. Start with (I) and (II):

(I) $P_h = P_o$

(II) $P_h = C$

From symmetry of identity and (1), we get

(V) $P_o = P_h$

We then pair (V) with (II):

(V) $P_o = P_h$

(II) $P_h = C$

From transitivity of identity, this pair gives

(VI) $P_o = C$

Next, pair (VI) with (III)

(VI) $P_o = C$

(III) $P_o = J$

And repeating the inference processes again with symmetry and transitivity, we eventually get

(VII) $C = J$

Put (VII) alongside (IV), we will have a contradiction:

(VII) $C = J$

(IV) $C \neq J$

Formally put:

1. $P_h = P_o$ Assumption

2. $P_h = C$ Assumption

3. $P_o = J$ Assumption

4. $C \neq J$ Assumption

5. $P_o = P_h$ 1, symmetry of identity

6. $P_o = C$ 5, 2, transitivity of identity

7. $C = P_o$ 6, symmetry of identity

8. $C = J$ 7, 3, transitivity of identity

9. Absurd 4, 8, introduction of absurdity

Of course, this is not the exact argument Putnam had in mind, for its conclusion does not mention identity theory at all. But we have now learnt the way to highlight transitivity of identity and symmetry of identity as the crucial rules of inference in his argument. We can therefore reconstruct Putnam's argument as follows:

Argument P-spelt-out

1. If identity theory is correct, then $(P_h = C)$ & $(P_o = J)$.

2. $P_h = P_o$

3. $C \neq J$

4. Identity theory is correct Assumed for reductio

5. $(P_h = C)$ & $(P_o = J)$ 1, 4, *modus ponens*

6. $P_h = C$ 5, &E

7. $P_o = J$ 5. &E

8. $C = P_h$ Symmetry of identity, 6

9. $C = P_o$ Transitivity of identity, 8, 2

10. $C = J$ Transitivity of identity, 9, 7

11. $(C = J)$ & $(C \neq J)$ 3, 10, &I

12. Identity theory is not correct. 4, 11, RAA

Or in short:

Argument P

1. If identity theory is correct, then $(P_h = C)$ & $(P_o = J)$.

2. $P_h = P_o$

3. If identity theory is correct, then $C = J$. 1, 2, Transitivity, symmetry

4. $C \neq J$

5. Identity theory is not correct. 3, 4, *modus tollens*

The new solution that I am going to introduce targets the validity of Argument P. Unlike traditional treatments of the apparent tension between (I) - (IV) where philosophers try to remedy the situation by rejecting one or more of (I) - (IV) (i.e. rejecting one or more of Putnam's premises), my approach does not focus on the truth-values of the four statements (and subsequently the truth-values of Putnam's premises) at all. Instead, I deny that they are genuinely inconsistent in the first place by attacking the main inference rule that the inconsistency is based upon, namely transitivity of identity. In particular, I consider the possibility that this rule fails under certain circumstances, including when the identities involve vague flanking terms. I argue that when identity statements involve vague terms, the resulting identities are indeterminate, and that the transitivity of identity is not a sound rule of inference where such indeterminate identities are involved. In consequence, step 3 of Argument P (or step 9 and 10 in the spelt-out version) is blocked. The *reductio* (or the *modus tollens* in the spelt-out version) therefore does not go through, and Putnam's argument is invalid.

5. 2. Is 'pain' vague?

Formally put, my reply to Putnam's multiple realisability argument against identity theory amounts to the following:

Argument V

V1. 'Human pain' and 'octopus pain' are vague terms.

V2. Vagueness of the flanking terms results in indeterminacy of the identity statements.

V3. Transitivity of identity fails for indeterminate identity.

Sub-conclusion V4. Transitivity of identity fails for identity statements involving 'human pain' and 'octopus pain'.

Conclusion V5. Putnam's argument is invalid.

In the following sections, I will explain my three premises one by one, beginning with clarifications of the two key elements in my repertoire – namely, *vagueness* and *indeterminacy.*

5. 2. 1. Theories of vagueness

Consider the following argument:

1 grain of wheat does not make a heap.
If 1 grain of wheat does not make a heap, then 2 grains of wheat do not.
If 2 grains of wheat do not make a heap, then 3 grains do not.
…
If 9,999 grains of wheat do not make a heap then 10,000 do not.

10,000 grains of wheat do not make a heap.

How about "short"?

A person of 205 cm is not short.
If a person of 205 cm is not short, then a person of 204.99 cm is not short.
If a person of 204.99 cm is not short, then a person of 204.98 cm is not short.
…
If a person of 100.01 cm is not short, then a person of 100 cm is not short.

A person of 100 cm is not short.

As these familiar considerations show, there are predicates like 'heap' and 'short' where there is no point at which we can discern that the properties these predicates denote ceases and their absences begin. This is said to be the vagueness phenomenon. R. M. Sainsbury, in his short and lucid textbook introduces vagueness via the concept of tolerance. A property is tolerant because it is possible to move from possession of *P* to lack of *P* via a series of distinct stages that are not notably different with respect to their *P*-hood.[162] No doubt, tolerant properties (and thereby the vague predicates denoting them) are very easy to find. Common terms like 'heap', 'short', 'bald' … etc. are all classic examples.

Tolerant properties provoke the so-called 'Sorites' paradoxes, as exemplified by the above 'heap' and 'short' examples. The conclusions that the Sorites reasoning give rise to are clearly problematic: it is counterintuitive that 10,000 grains of wheat do not make a heap, and a person of 100 cm is not short. But it is also true that these conclusions are derived from seemingly valid inferences. Take the 'heap' case as an example; Sorites reasoning has the following general form:

Categorical premise: A certain number of grains, say *n*, do not make a heap.

[162] Sainsbury, R. M. (2009). *Paradoxes*, p.41.

Conditional premises: If *n*-grains of wheat do not make a heap, then *n+1*-grains of wheat do not make a heap.

If *n+1*-grains of wheat do not make a heap, then *n+2*-grains of wheat do not make a heap.

…

Conclusion: *M*-grains of wheat do not make a heap.

We start with a categorical premise that seems to be factually correct, then add a conditional premise that reflects the tolerance principle – a single grain of wheat cannot be the difference between a heap and a non-heap. Repeat the second step as many times as you wish, then the paradoxical results arise when we conclude that *m*-grains of wheat do not make a heap where *m* is a very large number.

Given the form of its reasoning, possible solutions to the paradox are: 1) to reject the categorical premise; 2) to reject the conditional premise; or 3) to reject the validity of the argument. These options have evolved into several different views on the nature of vagueness, which I shall now briefly discuss.

The epistemic view on the nature of vagueness is advocated by Timothy Williamson and is best described and argued in his 1994 book *Vagueness*.[163] Epistemological varieties hold that there are sharp boundaries determining the applicability of expressions, but that we are ignorant of them. The argument for this approach insists that one of the premises must be false, namely, the conditional premise. If we hold that there are both heaps and non-heaps, defined by a distinct boundary, then once we knew the relevant information the paradox would disappear. The problem, as the epistemicists reveal, is that we cannot, even in principle, know where the boundary lies. This conclusion is counterintuitive, however: what reason do we have to think that such boundaries exist if we cannot come to know where they lie? Moreover, why can we not, even in principle, come to know these boundaries? As summarised by Tye, these questions indicate that the advocates of the epistemic account must believe in a peculiar precise boundary between heap and non-heap – a boundary that is *inaccessible to competent language users*.[164]

[163] Williamson, T. (1994), *Vagueness.*

[164] Tye, M. (1990), 'Vague Objects', in *Mind*, Vol. 99 p. 542.

Another way of rejecting the conditional premise is to take the *supervaluationist* proposal, which avoids the problems of the epistemic account. Kit Fine's 1975 article 'Vagueness, Truth, and Logic' states clearly by what criteria a term is vague or precise:[165]

To show what would make vague terms more precise, supervaluationists employ a technique called sharpening, which involves the following conditions:

- If *w* is definitely true of something, then *S*(*w*) is true of it.
- If w is definitely false of something, then *S*(*w*) is definitely false of it.
- For each object, *S*(*w*) is either true of it or false of it.
- *S*(*w*) respects the underlying ordering (if there is one). For example, if *S*('tall') is true of someone that is 188 cm tall, then it is also true of someone 189 cm tall.

Where these conditions disagree (i.e. some hold true while others are false), the supervaluationist believes there is vagueness; where they are jointly true, on the other hand, there is not. Moreover, the supervaluationist derives truth or falsity from the unanimous agreement or disagreement of these conditions. That is to say, where they all hold true, the supervaluationist identifies truth, and where they all hold false the supervaluationist identifies falsehood. Conversely, where only some agree and others disagree, the supervaluationist holds the vague predicate to be neither true nor false. Failure of the truth of conditionals which include such predicates as the borderline cases in the Sorites arguments, can be used as a case to reject the premises. Importantly, we do not need to specify exactly which premise is not true (just that there is at least one) in order to dissolve the paradox.

This account of vagueness thus preserves standard logic but is not immune to criticism. Firstly, it takes vagueness to be nothing more than incompleteness of meaning. We may consider, however, a predicate that is incomplete of meaning but not vague. For example, imagine that the meaning of 'adult' included clauses such as 'anyone under the age of 17 is not an adult' and 'anyone over the age of 18 is an adult'. The clauses are precise (i.e. not vague), but the meaning is incomplete, as it fails to speak of those between the ages of 17 and 18. Intuitively, there seems to be a distinction here between

[165] Fine, K. (1975), 'Vagueness, Truth, and Logic', in *Syntheses*, Vol. 30 (3-4), pp. 265-300.

incompleteness of meaning and vagueness; there is a blank where there should be a rule.[166]

Tye considers a further problem with the supervaluationist approach. The process of sharpening implies that there are things that need to be sharpened. Since '*nothing that is already precise can be made precise*', it follows that there are things that are vague – a notion that supervaluationism aims to refute.[167] This brings us to the realist account of vagueness.

Ever since the discussion by Michael Dummett, the realist view about vagueness has provoked heated debate.[168] According to the realist approach, vague terms denote ontic vagueness. That is, the vagueness in words results from a fundamental characteristic of reality – that some parts of the physical world lack precision. According to Merricks' categorisation, the ontic view of vagueness holds that vagueness is a property of the object itself. That is to say, there is no determinate fact of the matter whether an object exemplifies a particular property. The corresponding conditional will have no truth-value; thus, the Sorites paradox is dissolved.[169]

Sainsbury has a number of concerns regarding the ontic account. Firstly, what reason do we have to think that objects themselves can be vague? Take the standard cloud example: there appears to be no clear answer as to where the boundary of a cloud lies. In order to prove the existence of vague objects, however, we need to do more than point out the existence of borderline cases, as these are unanimously accepted by proponents of other theories of vagueness, too. Friends of ontic vagueness may instead draw upon an analogy to the question of whether there is necessity in the world. If the world contains some objects which must necessarily be a particular way, then the world contains necessity. Similarly, if it is vague whether an object is a particular way, then the world contains vagueness. This analogy is problematic, however. It does not give a sufficient condition for locating necessity in the world rather than in language.[170] A serious question therefore emerges: does vagueness in objects explain vagueness in language or the other

[166] This problem has been discussed by Sainsbury (1995, 2009) and Tye (1994).

[167] Tye (1990), p. 541.

[168] Dummett, M. (1975), 'Wang's paradox' in *Synthesis*, Vol. 30 (3-4), pp. 301-32.

[169] Merricks, T. (2001), 'Varieties of Vagueness' in *Philosophy and Phenomenological Research*, 62, pp. 145-57.

[170] Sainsbury (1995), pp. 68-73.

way around? While the ontic account opts for the latter answer, the semantic account supports the former.

Contrary to the ontic view, one can subscribe to the belief that vagueness is merely a semantic phenomenon. This view is sometimes called the semantic theory or the linguistic theory of vagueness. Holders of this view believe that linguistic items such as words can be vague, but objects cannot be. In the 'short' example, the predicate 'is short' is vague because the term 'short' is, not because the property that this term refers to is. Vagueness, thus, is a property of terms, not a property of the objects that terms refer to. Lewis, one of the first advocates of the semantic view, sums up the view with a reference to the term 'outback':

> The only intelligible account of vagueness locates it in our thought and language. The reason it's vague where the outback begins is not that there's this thing, the outback, with imprecise borders; rather there are many things, with different borders, and nobody has been fool enough to try to enforce a choice of one of them as the official referent of the word 'outback'. Vagueness is semantic indecision.[171]

In summary, there is reason to think vagueness is epistemic; there is a distinct boundary between heaps and non-heaps, we just do not know where it lies because there is a margin of error that exists in all observations. This margin of error infests the symbols we use to represent the world, meaning that all symbols are vague. Traditional logic, however, assumes these symbols to be precise, hence the Sorites paradox. Supervaluationists, on the other hand, argue that all vagueness is linguistic: the property of an incomplete or defective language. They employ the notion of sharpening to determine the meaning of a predicate and its truth-value when applied to an object. The result is that vague propositions can be neither true nor false, so any conditional involving them fails to be true and Sorites reasoning comes to a halt. There are two problems for supervaluationism, however. Firstly, it seems vagueness is not limited to language. Secondly, the theory may collapse into either metaphysical or epistemological vagueness. We also cannot rule out vagueness being metaphysical - a property of the object itself. In this case, there is no determinate fact of the matter whether an object exemplifies a property, so any conditional involving the object will fail to be true.

Despite the obvious distinctions, what do these aforementioned solutions to the Sorites paradox have in common? At least one thing is certain: these different options all seem to

[171] Lewis (1986), *On the Plurality of Worlds*, p.212.

agree that the occurrence of vague terms mean that vagueness exists somewhere in our world and is causing us a noticeable philosophical problem. Nevertheless, as I briefly noted in the above section, different treatments of the Sorites paradox will in turn reflect different answers to the question of where exactly is vagueness situated. In elaboration, is the vagueness in words an evidence for the *ontic, realist* view – that the actual world contains objects that are intrinsically lacking precise boundaries in them? Or, is vagueness an *epistemic* issue only – nothing out there in the world and external to us is vague, and the occurrence of vague words only underlies our inability to know the precise, sharp boundaries? Or, could it be the case that vagueness only comes about at the *semantic* level – that vague words indicate nothing but, well, vagueness in words, and no further conclusions about the external world should we made from this semantic feature? As I will elaborate in 6.3, I take an agnostic position about the nature of vagueness. To lay the foundation of my argument against Putnam, all it requires is to treat vagueness as a predicate of terms. In this regard, my approach resembles the semantic account. Nevertheless, I refrain from answering whether or not vagueness is *merely* a predicate of terms as all the abovementioned accounts tend to argue.

5. 2. 2. Vagueness and phenomenal terms

Once we understand what it means for a term to be vague, we are in a position to discuss an important question: are phenomenal terms vague? Anthony Everett (1996) provides powerful reasoning to the effect that phenomenal concepts *are* vague, as part of his argument against the existence of qualia[172]:

<u>Argument E</u>

Premise 1. Phenomenal properties are tolerant properties.

Premise 2. Tolerant properties are vague.

Premise 3. Vague properties, by Gareth Evans' argument, do not exist.

Conclusion. Phenomenal properties do not exist.

[172] Everett, A. (1996), 'Qualia and Vagueness' in *Syntheses*, Vol. 106, No. 2, pp. 205-26.

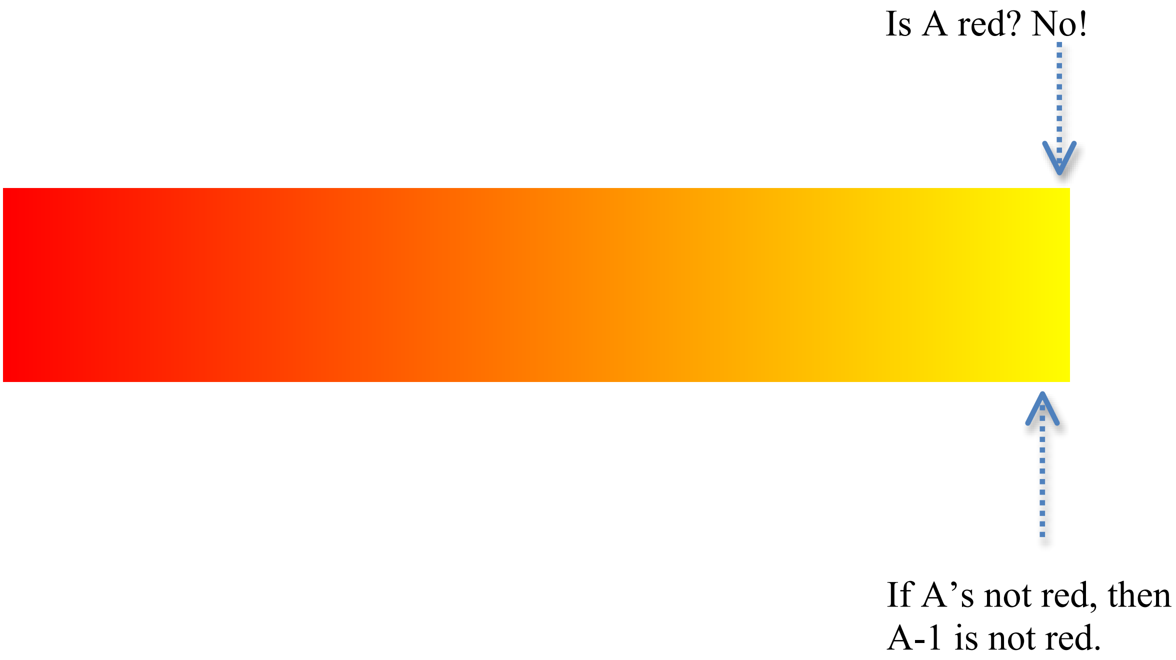
The soundness of this argument is of no interest here. Our only concern is the truth of Premise 1 and Premise 2, which together entail an affirmative answer to the question of whether phenomenal terms are vague:

Argument EH

Premise 1.	Phenomenal properties are tolerant properties.
Premise 2.	Tolerant properties are vague.
Conclusion.	Phenomenal properties are vague.

By the Soritical reasoning explained above, we have already defined vagueness in terms of tolerance, so Premise 2 is true by definition. Everett argues for Premise 1 by having us consider a spectrum with yellow at its right end and red at its left:

Figure 5-a



This allows the following argument to be constructed:

The colour at the right end is not red.
If the colour at the right end is not red, then its immediate left-hand side neighbour (L_1) is not red.

If L_1 is not red, then the immediate left-hand side neighbour of L_1 (L_2) is not red.
…
If L_n is not red, then L_{n+1} is not red.

L_n is the immediate right-hand side neighbour of the left end.

The colour at the left end is not red.

The argument shows that the term "red" is tolerant: if one colour appears to us to be red, so will its immediate neighbours, and if one colour appears not to be red, so will its immediate neighbours. There will be no point at which we can discern that redness ceases and its absence begins. We know that if a concept can be used to produce such a Soritical sequence (i.e. tolerant), then it is 'vague'. Hence at least one phenomenal concept is vague—namely, the concept of redness.

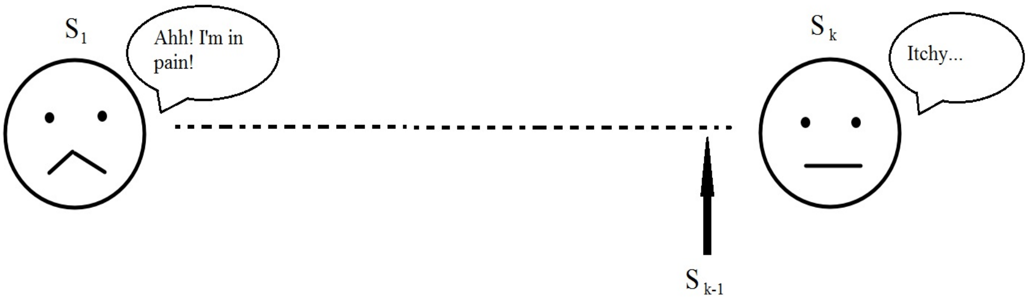
The above Soritical sequence about "red" readily generalizes. It applies not only to redness and other colour concepts (terms), but also to other phenomenal concepts, like pain, as similar reasoning shows that other phenomenal concepts are also vague. Everett goes so far as to suggest that all phenomenal properties are vague, saying:

> For the sorts of phenomenal features that we have been talking about; redness, heat, loudness, bitterness, and so on, are all notoriously tolerant [i.e. 'it is possible to move from possession of a phenomenal property *P* to lack of *P* via a series of distinct stages that are not notably different with respect to their *P*-hood'] … We can construct an analogous argument with respect to any other purported phenomenal property, taking an appropriate well ordering of objects and letting that quality stand in place of redness.[173]

To clarify this point, imagine Fred is in pain and swallows a painkiller. As the chemical takes effect, the painful sensation changes over a minute or two into an annoying but no longer painful itch. This transition forms a Soritical sequence of a familiar sort. Let us call the original phenomenal state (immediately before the analgesic is taken) S_1, and the final itchy-but-not-painful phenomenal state S_k. By means of the usual Soritical reasoning, it can be shown that S_k is painful, since: (i) S_1 is painful, and (ii) for all n, $1 \leq n \leq k$, if S_n is painful then S_{n-1} is painful, where the steps are sufficiently small:

[173] Everett (1996), pp. 208-9

Figure 5-b



The phenomenal state at the right end, S_k, is not pain.

If the phenomenal state at the right end, S_k, is not pain, then its immediate left-hand side neighbour, S_{k-1}, is not pain.

If S_{k-1} is not pain, then the immediate left-hand side neighbour of S_{k-1}, S_{k-2}, is not pain.
…
If S_n is not pain, then S_{n+1} is not pain.

The phenomenal state at left end, S_1, is not pain.

Again, the important philosophical issue here is that if one phenomenal state on the spectrum feels to us to be pain, so will its immediate neighbours, and if one phenomenal state feels not to be pain, so will its immediate neighbours. Thus, there will be no point at which we can discern that pain ceases and its absence begins. In a nutshell, ‘pain’ is tolerant and can be used to produce a genuine Soritical sequence. By Argument EH and the instantiation rule, ‘pain’ is vague. Next, I will discuss the implications for this idea to Putnam’s argument by first introducing the notation for vagueness.

- **Notations**

Let ‘V’ mean “is vague” and V(*t*) mean that term *t* is vague. For example, V(P_h) denotes the claim that the term ‘human pain’ is vague. On the flip side (pun intended), let us abbreviate ‘¬V’ as ‘Λ’, which means ‘is not vague’ (or ‘is crisp’, as I will often put) and Λ(*t*) as meaning that the term *t* is not vague (or is crisp). For example, Λ(C) denotes the claim that the term ‘C-fibre firing’ is crisp. Given the definitions, we know that the following equivalence principle of vagueness and crispness is sound:

(VP1) $V(t) \dashv\vdash \neg\Lambda(t)$

(VP1) in turn implies the fact that no terms are both vague and crisp. In other words, *t* is either vague or crisp:

(VP2) $\quad V(t) \vee \Lambda(t)$

With the help of these new notations, and on the basis of above the reasoning, I conclude that (1) and (2) are true:

(1) $\quad V(P_h)$

(2) $\quad V(P_o)$

In other words, 'human pain' and 'octopus pain' are vague terms. In conclusion, the first premise of my Argument V is true:

(V1) $\quad V(P_h) \& V(P_o)$

5. 3. Vagueness to indeterminacy

Having established that phenomenal terms are vague, I now examine what consequences this has viz-a-vis the truth-values of identity statements involving vague terms such as phenomenal terms P_h and P_o.

5. 3. 1. Evans on vagueness and indeterminacy

In his classic 1978 paper, Gareth Evans wrote:

> It is sometimes said that the world might *be* vague. Rather than vagueness being a deficiency in our mode of describing the world, it would then be a necessary feature of any true description of it. It is also said that amongst the statements which may not have a determinate truth-value as a result of their vagueness are identity statements. Combining these two views we would arrive at the idea that the world might contain certain objects about which it is a *fact* that they have fuzzy boundaries. But is this idea coherent?[174]

As carefully examined by Copeland,[175] Evans' target in this opening passage was a view of Dummett's—that the world contains vague objects:[176]

[174] Evans, G. (1978), 'Can There Be Vague Objects?', *Analysis*, Vol. 38, No.4, p. 208.

[175] Copeland, B. J. (1997), 'Vague Identity and Fuzzy Logic' in *Journal of Philosophy*, Vol. 94 (10), pp. 514-34. Copeland discovered Evans had acknowledged in a letter to David Lewis that he

(DV) $(\exists x)\mathrm{V}x$

In arguing agianst DV, Evans attempts to prove that identity statements cannot have an indeterminate truth value. Hence, he argues against the following claim, (ID), where "∇" is Evans' indeterminacy operator (see page 142).

(ID) $\nabla(a = b)$[177]

Evans' argument against (ID) is via a *reductio*:[178]

(EV1) $\nabla(a = b)$ Assumed for *reductio*

(EV2) $\lambda x(\nabla(x = a))b$[179] EV1, property abstraction

(EV3) $\neg\nabla(a = a)$ Axiom

(EV4) $\neg\lambda x(\nabla(x = a))a$ EV3, property abstraction

(EV5) $\neg(a = b)$ EV2, EV4, Leibniz's Law

The reasoning goes: suppose for *reductio* that *a* is indeterminately identical to *b*. By property abstraction, *b* has the property of *being indeterminately identical to a*. But it is undisputable that self-identity is always determinate, hence it is not the case that *a* is indeterminately identical to *a*. It immediately follows that *a* does not have the property of *being indeterminately identical to a*. There is, therefore, a property that *b* has and *a* doesn't, by Leibniz's Law, *a* is not identical to *b*, which negates the first assumption that *a* is indeterminately identical to *b*.

Before I assess this argument and its implication for my critique of Putnam, it is important to note that, as several commentaries soon pointed out, there is a discrepancy between two views – the one Evans appears to be attacking in his opening passage (and referenced in the article's title), which is (DV), and the one he proceeded to attack in the

was questioning Dummett; and Evans' reluctance to name Dummett in the paper was due to the lack of clear reference in Dummett's publications at the time.

[176] Dummett (1975), pp. 314-24.

[177] The upside-down triangle "∇" will be defined on pg. 142.

[178] Advocates of the same argument are Salmon (1981) and Wiggins (1986).

[179] Evans used "û" instead of "λ*x*" to notate property abstraction.

rest of the paper, which is (ID).[180] This discrepancy, or lack of apparent flow in Evans' paper has subsequently instigated a discussion about his precise aim. Is the paper trying to disprove (ID) or (DV) or both? As the body of literature on this topic continues to grow, the answer to this question remains, at best, unclear. For my purposes, however, I can set aside this distracting issue about who the 'real' target of Evans' paper is. The more interesting question, I think, is the logical relation between (ID) and (DV). We know Evans at least appears to be objecting to both claims in the two halves of his paper. What logical bearing, if any, in spite of Evans' intention, does the truth or falsity of (ID) have on the truth or falsity of (DV)? By contraposition, what would Evans' purported opposition, namely Dummett, have to add to (DV) should he wish to derive (ID)? Recalling Evans' opening paragraph, we can spot a vital clue: 'It is also said that amongst the statements which may not have a determinate truth-value as a result of their vagueness are identity statements.'[181] Thus, as Evans acknowledged, the Dummetian reasoning gap is filled by a '*vagueness to indeterminacy'* theorem that attributes the indeterminacy of an identity statement to the vagueness of one or both of its flanking terms:

(VTI) $\mathrm{V}(x), \mathrm{V}(y) \vdash \nabla(x = y)$

Let us try to derive (ID) from (DV) and (VTI) to see if (VTI) really serves as a Dummettian assumption:

Derivation D

(DV) $(\exists x)\mathrm{V}x$

(VTI) $(\mathrm{V}(x) \vee \mathrm{V}(y)) \vdash \nabla(x = y)$

(ID) $\nabla(a = b)$

Filling in some steps of propositional calculus we can see that the derivation runs through, free from any apparent logical flaws. Consequently, against this Dummettian derivation, an Evansian argument[182] can thereby be represented as

Derivation E

[180] These include Burgess (1989), Broome (1984), Copeland (1994, 1997, 2000), Garrett (1988, 1991), Lewis (1988), Noonan (1982, 1984, 1990, 1991), Parsons (1987, 2000), Parsons & Woodruff (1995), Pelletier (1989), Thomason (1982), and Wiggins (1986).

[181] Evans, (1978), p. 208.

[182] An Evansian argument is not necessarily Evans' argument in his (1978).

(NID) $\neg\nabla(a = b)$

(VTI) $(\mathrm{V}(x) \vee \mathrm{V}(y)) \vdash \nabla(x = y)$

(NDV) $\neg(\exists x)\mathrm{V}x$

which is also valid.

Having stated both sides' reasoning explicitly above, one thing needs to be emphasized immediately: this Dummettian assumption (VTI), or something close to it, is a vital step for both arguments. (VTI) is both a Dummettian and Evansian assumption, should either side wish to derive their respective conclusion. The crucial role of (VTI) appears to have been touched on by not only Evans himself, but also discussed by several others, as evidenced by Harold Noonan's (1990) succinct summary:

> Evans's target is rather the view that there can be identity statements which are indeterminate in truth-value not because of any semantic indeterminacy but rather because of indeterminacy *in the world*, that is, *for no other reason* than that one or both of the objects *determinately* denoted by the singular terms flanking the identity sign is a vague object.[183]

David Wiggins (1986) also emphasises the importance of (VTI) in the following passage where he interprets and attempts to rectify Evans' argument:

> It is important to see that (iv) [$(a = b) \rightarrow \Delta\ (a = b)$] does not entail that every true identity sentence will remain true when prefixed with 'definitely'. For identity sentences may contain descriptions, vagueness [of the sentence] may result from a vagueness of these descriptions. Consider the not implausible claim 'The greatest ruler was the wisest ruler'. Very likely there is indeterminacy in this claim.[184]

Echoing the same line of thought, B. J. Garrett (1988) makes the stronger claim that someone who believes (DV) must also believe (VTI):

> The thesis that there can be vague objects is the thesis that there can be identity statements which are indeterminate in truth-value (i.e. neither true nor false) as a result of vagueness (as opposed e.g. to reference-failure), the singular terms of which do not have their referents fixed by vague descriptive means. (If this is *not* what is meant by the thesis that there can be vague objects, it is not clear what *is* meant by it).[185]

[183] Noonan (1990), pp.157-8. The point was also stated in Noonan (1991), pp. 183.

[184] Wiggins (1986), p. 174.

[185] Garrett (1988), p. 130.

Referring to Wiggins' 'ruler' example, Garrett even claimed that this "vagueness to indeterminacy" move is 'uncontentious':

> It seems uncontentious that there can be vague identity statements the vagueness of which is a consequence of the vagueness of their component singular terms – e.g. 'the greatest ruler was the wisest ruler'.[186]

Despite the consensus that (VTI), or something similar to it, is needed in both Dummettian and Evansian reasoning, the popular impression of (VTI), as Garrett's comment showed, seems to be that it is either proven or doesn't need to be proved. So far, no explanation has been given by anyone as to why this is so. For my purpose, it will be very helpful if we can develop a proof for this "vagueness to indeterminacy" theorem, or something similar, because it justifies V2 – the second premise of my Argument V against Putnam—and allows us to subsequently derive (by adding V1) that some or all of (I)-(IV) are indeterminate identity statements. But before I proceed to introduce the proof, a few words about indeterminacy are required.

5. 3. 2. Introducing the 'vagueness to indeterminacy' theorem

To express indeterminacy we use Evans' delta-operators: "$\nabla(S)$" reads '*S* is indeterminate', and this is true when *S* is not determinately true or not determinately false. Since the delta-operators are duals, "$\Delta(S)$" reads 'statement *S* is determinate', and this is true when *S* is determinately true or determinately false.[187] Thus, the following equivalence principles are sound:

(EP1) $\Delta(S) \dashv\vdash \neg\nabla(S) \vee \neg\nabla(\neg S)$

(EP2) $\nabla(S) \dashv\vdash \neg\Delta(S) \vee \neg\Delta(\neg S)$

Since $\Delta(S)$ is true when *S* is either determinately true or determinately false, and similarly $\nabla(S)$ is true when either *S* or $\neg S$ is indeterminate, we also have:

(EP3) $\Delta(S) \dashv\vdash \Delta(\neg S)$

(EP4) $\nabla(S) \dashv\vdash \nabla(\neg S)$

[186] Ibid. Footnote 1.

[187] According to Copeland (1997), Evans acknowledged in a letter to David Lewis that $\Delta(S)$ is true when *S* is false.

Given EP3 and (EP4), we can thereby shorten (EP1) and (EP2):

(EP1S) $\Delta(S) \dashv\vdash \neg\nabla(S)$

(EP2S) $\nabla(S) \dashv\vdash \neg\Delta(S)$

From EP1S and EP2S, we derive that S cannot be both determinate and indeterminate. In other words, S is either determinate or indeterminate:

(EP5) $\Delta(S) \vee \nabla(S)$[188]

Now we are in a position to discuss the "vagueness to indeterminacy" theorem:

(VTI) $\mathrm{V}(x), \mathrm{V}(y) \vdash \nabla(x = y)$

Or in English:

(VTI) If one or both of the flanking terms of an identity statement are vague, then the identity statement is indeterminate.

By instantiation of the variables, we know VTI entails

$$\mathrm{V}(x), \mathrm{V}(x) \vdash \nabla(x = x)$$

However, as Evans plausibly maintains (EV3 as abovementioned), $\neg\nabla(x = x)$ is always true, because self-identity is always determinately true:

$$\Delta(x = x)$$

From the equivalence principle EP1, it follows that $\nabla(x = x)$ is never true. So by modus tollens, (VTI) is not true. Thus, for reasons specified by Evans, in place of (VTI), we should use something else that does not imply $\Delta(x = x)$ and still justifies V2:

V2. Vagueness of the flanking terms results in indeterminacy of the identity statements.

I thereby suggest the weakened but more promising

[188] Strictly speaking, EP5 does not showcase an equivalence relation. Rather, it announces a disjunctive relation between $\mathrm{V}(S)$ and $\Lambda(S)$. Nevertheless, EP5 is derivable from other equivalence principles, as described. For convenience, I am retaining the acronym "EP" for this disjunctive statement.

(VTIW) $\mathrm{V}(x), \Lambda(y) \vdash \nabla(x = y)$

And its dual

(VTIWD) $\mathrm{V}(y), \Lambda(x) \vdash \nabla(x = y)$

which in English reads

(VTIW) If one flanking term of an identity statement is vague, *and the other flanking term is crisp,* then the identity statement is indeterminate.

The upshot is quite clear. VTIW improves upon the strong version, VTI, by elucidating an ambiguity voiced in Evans' quotation and the interpretations ensued from it (best typified by Noonan's interpretation quoted above): Does the vagueness of *one* or *both* flanking terms result in the indeterminacy of the identity statements? As shown, VTI remains uncertain on this question and thereby allows the possibility of having two vague flanking terms. Its weakened modification, VTIW, to the contrary, specifies the number of vague flanking terms to be just one, and it is immune to Evans' attack as a result. Precisely, VTIW does not entail $\nabla(x = x)$, the crucial step in Evans' proof, because if x is vague then we can't substitute x for y in $\mathrm{V}(x)$ & $\Lambda(y)$. At the same time, in spite of the fact that they are weakened, VTIW and VTIWD together suffice to formalize the Dummettian view previously expressed in VTI. This can be seen by recalling the Dummettian and Evansian derivations and replacing VTI with VTIW:

Derivation D

(DV) $(\exists x)\mathrm{V}x$

(VTI) $\mathrm{V}(x), \mathrm{V}(y) \vdash \nabla(x = y)$

(ID) $\nabla(a = b)$

Derivation D – weakened (using VTIW)

(DV) $(\exists x)\mathrm{V}x$

(VTIW) $\mathrm{V}(x), \Lambda(y) \vdash \nabla(x = y)$

(ID) $\nabla(a = b)$

Derivation E

(NID) $\neg\nabla(a = b)$

(VTI) $\mathrm{V}(x), \mathrm{V}(y) \vdash \nabla(x = y)$

(NDV) $(\exists x)\mathrm{V}x$

Derivation E – weakened (using VTIW)

(NID) $\neg\nabla(a = b)$

(VTIW) $\mathrm{V}(x)\,,\, \Lambda(y) \vdash \nabla(x = y)$

(NDV) $(\exists x)\mathrm{V}x$

In summary, the weakened version, I claim, is a much-improved formalization of the ‘vagueness to indeterminacy’ theorem. It preserves the key notion expressed by Evans’ quotation: ‘It is also said that amongst the statements which may not have a determinate truth-value as a result of their vagueness are identity statements.’ Evans’ proof, which targets the strong interpretation of this idea, namely (VTI), does not undermine the weakened version (VTIW). Evans and subsequent commentators of his paper focused only on (VTI) and overlooked the possibility of (VTIW). Furthermore, spelling out the correct formalization of the ‘vagueness to indeterminacy’ move is essential for my argument against Putnam as it explains V2. Having established that (VTIW) suffices for the task, I can rewrite V2 as follows:

V2. If *only one* flanking term of an identity statement is vague, then the identity statement is indeterminate.

I have just explained how (VTIW) escapes Evans’ attack based on the determinacy of self-identity, and thereby improves on (VTI). This does not answer why (VTIW) and (VTIWD) are true per se. I have not yet offered a proper proof of (VTIW). I will do this in Chapter 7. In the meantime, I am going to assume it is true. The consequence of it being true, with regard to Putnam’s multiple realisability argument, will now be examined.

How would this weakened version of the vagueness-to-indeterminacy principle affect Putnam’s argument? Recall the two psychophysical identity claims that Putnam aims to demolish:

(II) $\mathrm{P_h} = \mathrm{C}$

(III) $\mathrm{P_o} = \mathrm{J}$

We have established in the last section that ‘pain’ (i.e. ‘$\mathrm{P_h}$’ and ‘$\mathrm{P_o}$’) is a vague term, hence $\mathrm{V}(\mathrm{P_h})$ and $\mathrm{V}(\mathrm{P_o})$. Let us assume that, unlike ‘pain’, ‘C-fibre firings’ (i.e. ‘C’) is crisp. (I will argue for this in the next Chapter, section 6.2.). That’s to say, we assume the following:

$$\Lambda(\mathrm{C})$$

For the same reason, we can attribute crispness to ‘Jelly-firings’, thus

$$\Lambda(\mathrm{J})$$

Now, we can instantiate the variables in (VTIW) with ‘$\mathrm{P_h}$’ and ‘C’, and ‘$\mathrm{P_o}$’ and ‘J’ and get (VTIWP) and (VTIWO):

(VTIW)	$\mathrm{V}(x), \Lambda(y) \vdash \nabla(x = y)$
(VTIWP)	$\mathrm{V(P_h)}, \Lambda(\mathrm{C}) \vdash \nabla(\mathrm{P_h} = \mathrm{C})$
(VTIWO)	$\mathrm{V(P_o)}, \Lambda(\mathrm{J}) \vdash \nabla(\mathrm{P_o} = \mathrm{J})$

Thus, due to the vagueness of ‘$\mathrm{P_h}$’ and the crispness of ‘C’, the identity between the two terms is indeterminate identity by *modus ponens*:

1.	$\mathrm{V(P_h)}$	Assumption
2.	$\Lambda(\mathrm{C})$	Assumption
3.	$\mathrm{V(x)}, \Lambda(\mathrm{y}) \vdash \nabla(\mathrm{x} = \mathrm{y})$	Theorem (VTIW)
4.	$\mathrm{V(P_h)}\ \&\ \Lambda(\mathrm{C}) \rightarrow \nabla(\mathrm{P_h} = \mathrm{C})$	3, instantiation
5.	$\nabla(\mathrm{P_h} = \mathrm{C})$	1, 2, 4, *modus ponens*

The same goes for the identity between ‘octopus pain’, ‘$\mathrm{P_o}$’, and its physical-chemical correlate, ‘J’:

1.	$\mathrm{V(P_o)}$	Assumption
2.	$\Lambda(\mathrm{J})$	Assumption
3.	$\mathrm{V(x)}, \Lambda(\mathrm{y}) \vdash \nabla(\mathrm{x} = \mathrm{y})$	Theorem (VTIW)
4.	$\mathrm{V(P_o)}, \Lambda(\mathrm{J}) \vdash \nabla(\mathrm{P_o} = \mathrm{J})$	3, instantiation
5.	$\nabla(\mathrm{P_o} = \mathrm{J})$	1, 2, 4, *modus ponens*

Hence, we can rewrite (II) and (III) to express the indeterminacy therein:[189]

(II) $\nabla(P_h = C)$

(III) $\nabla(P_o = J)$

5. 4. Indeterminacy and transitivity

In previous two sections, it was established that identity statements (II) and (III) in Putnam's tetrad should be modified by the indeterminacy operator. The tetrad shall therefore be stated as

(I) $P_h = P_o$

(II) $\nabla(P_h = C)$

(III) $\nabla(P_o = J)$

(IV) $C \neq J$

5. 4. 1. Determinacy status of identities and terms

Here, I shall clarify a side issue first. That is, should we also modify (I) and (IV)? We now know from the weakened vagueness to indeterminacy theorem (VTIW) that the identity between a vague term and a crisp term is indeterminate identity. We have also learnt from Evans' proof (part of it, to be precise), barring possible objections,[190] that the identity between two vague terms is not always indeterminate identity. Specifically, Evans stressed the determinacy between self-identical terms. But what about the identity between two terms that are a) not self-identical, and b) vague, like P_h and P_o? In other words, is (I) determinate or indeterminate? And what about the identity between two crisp terms as shown in (IV)? As the following table demonstrates, neither of these questions can be answered by (VTIW).

[189] (II) and (III) are specifically formulated version of mind-brain identity claims. They add that the identity in question is indeterminate, not determinate.

[190] See 6.3.

Table 5-c

	Λa	$\mathrm{V}a$
Λb	?	$\nabla(a = b)$
$\mathrm{V}b$	$\nabla(a = b)$	? (except for the case of $a = a$, which is Δ)

To find out the determinacy of identity statements outside of (VTIW)'s scope, we will need to find a proper axiom-schema that includes all three possible combinations of vague and/or crisp flanking terms, since (VTIW) only determines 2):

1) Λa & Λb
2) Λa & $\mathrm{V}b$
3) $\mathrm{V}a$ & $\mathrm{V}b$

That leaves us with 1) and 3): identity between two crisp terms and identity between two vague terms. Situation 3) is the most problematic kind, as exposed by Evans' proof. On the other hand, 1) seems to be relatively easy to deal with. It is a situation in which no vagueness is involved. We shall therefore treat the truth function of these identity statements as we usually do - determinately, that is. After all, this is how we normally take identity statements to be - either determinately true or determinately false, until we learn about the concepts of vagueness and indeterminacy. But since we have learnt that at least in some cases, for instance, situation 2), identity statements are not either determinately true or determinately false, the scope of determinate identity statements is restricted. To specify this scope, we need to specify the necessary and sufficient conditions for determinate identities. We are now in a position to describe one of them:

(CTD) $$\Lambda(x), \Lambda(y) \vdash \Delta(x = y)$$

which in English reads

(CTD) If both flanking terms of an identity statement are not vague (crisp), then the identity statement is determinate.

I am hereby proposing a 'crispness to determinacy' theorem, which gives a sufficient condition for determinate identity statements. I believe little explanation is needed for this theorem, as it is so simple and intuitive. For solidification, we can prove that CTD is always true via the same crucial step as in Evans' proof against $\nabla(a = b)$:

By substitution,

$$\Lambda(x), \Lambda(y) \vdash \Delta(x = y)$$

entails

$$\Lambda(x), \Lambda(x) \vdash \Delta(x = x)$$

which is always true, because self-identity is always determinate, as Evans emphasised.

Now we can replace a question mark in the Table 5-c and get the following Table 5-d:

Table 5-d

	Λa	$\mathrm{V}a$
Λb	$\Delta(a = b)$ – explained by (CTD)	$\nabla(a = b)$ – explained by (VTIW)
$\mathrm{V}b$	$\nabla(a = b)$ – explained by (VTIW)	? (except for the case of $a = a$, which is Δ)

So, although further work is required to envisage an axiom-schema of determinacy and indeterminacy that accommodates all possible combinations of vague/crisp flanking terms in identity statements, this work are beyond the scope of my PhD thesis. I have shown that we can decide whether an identity statement is determinate or indeterminate in two combinations of vague/crisp flanking terms. Accordingly, we can answer the question put forward at the beginning of this section: are (I) and (IV) determinate or indeterminate? I will conclude that (IV) is determinate, whereas I will refrain from the determinacy or indeterminacy of (I). Putnam's tetrad, therefore, can be further modified as

(I) $\quad P_h = P_o$

(II) $\quad \nabla(P_h = C)$

(III) $\quad \nabla(P_o = J)$

(IV) $\quad \Delta(C \neq J)$

As I will soon explain, we do not need to be informed about whether (I) is determinate or indeterminate in order to attack Putnam's argument against identity theory.

5. 4. 2. Transitivity fails for indeterminate identity

Without further ado, the final premise of my argument against Putnam deals with transitivity of identity and asks whether it works for indeterminate identities. I argue it doesn't, and that this in turn erodes the foundation of Putnam's. argument As stated in 5.1.2, transitivity of identity is the following principle:

(ToI) $((x = y) \& (y = z)) \rightarrow (x = z)$

If the principle also works for indeterminate identities then:

(IToI) $(\nabla(x = y) \& \nabla(y = z)) \rightarrow \nabla(x = z)$

Or alternatively:

(IToI) $\neg\ (\nabla(x = y) \& \nabla(y = z) \& \neg\nabla(x = z))$

Thus, transitivity fails for indeterminate identities if we can find instances in which (TT) is true:

(TT) $\nabla(x = y) \& \nabla(y = z) \& \neg\nabla(x = z)$

What kind of instantiations of *x*, *y*, and *z* can make (TT) true? Let us first try with the terms in Putnam's argument. Recall that Putnam's argument runs as follows:

1.	$P_h = P_o$	Assumption
2.	$P_h = C$	Assumption
3.	$P_o = J$	Assumption
4.	$C \neq J$	Assumption
5.	$P_o = P_h$	1, symmetry of identity
6.	$P_o = C$	5, 2, transitivity of identity
7.	$C = P_o$	6, symmetry of identity
8.	$C = J$	7, 3, transitivity of identity
9.	Absurd	4, 8, AbsI

Let us focus on line 7, line 3 and line 8. It seems we can derive 8 from 7 and 3 by transitivity:

$$\begin{array}{ll} C = P_o & \\ P_o = J & \\ \hline C = J & \text{Transitivity of identity} \end{array}$$

It was established above that phenomenal terms like P_o are vague; in contrast, C and J are crisp terms. So from VTIW we get the ∇-modified version of the first two lines:

$$\nabla(C = P_o)$$
$$\nabla(P_o = J)$$

Now if transitivity works for indeterminate identity, we would have

$$\begin{array}{ll} \nabla(C = P_o) & \\ \nabla(P_o = J) & \\ \hline \nabla(C = J) & \text{Transitivity of identity} \end{array}$$

However, the inferred statement is false, since both 'C' and 'J' are crisp terms, (CTD) entails that C = J is a determinate identity statement. That is, $\Delta(C = J)$. But this is determinately false! C-fibre firings are entirely different from Jelly firings – a fact Putnam emphasised himself and uses to support his own argument. So instead of

$$\nabla(C = J),$$

we should have

$$\Delta(C \neq J),$$

which, by the equivalence principle (EP1), is equivalent to

$$\neg\nabla(C = J)$$

Thus, in place of

$$\begin{array}{ll} \nabla(C = P_o) & \\ \nabla(P_o = J) & \\ \hline \nabla(C = J) & \text{Transitivity of identity} \end{array}$$

in which transitivity works as the essential rule of inference, we in fact have the following triad of identity statements:

$$\nabla(C = P_o)$$

$$\nabla(P_o = J)$$
$$\neg\nabla(C = J)$$

which is an instance of TT. Or in other words, transitivity fails to infer the third statement from the first two statements. Then by *modus tollens*, transitivity fails to be a valid inference for indeterminate identity. Notice that the argument does not just show that (TT) is true and thus (ItoI) is false. It also shows that Putnam's argument is invalid. Without (IToI), Putnam is entitled neither to the move from line 5 and line 2 to line 6, nor from line 7 and line 3 to line 8.

Unfortunately, there is a possible drawback in the above reasoning.[191] My *modus tollens* might be seen as Putnam's *modus ponens.* I have explained that if both $\nabla(C = P_o)$ and $\nabla(P_o = J)$ are true, $\nabla(C = J)$ cannot be true, providing an obvious counterexample to (IToI). But Putnam would presumably respond by simply insisting that (IToI) is true and instead concluding that $\nabla(C = P_o)$ and $\nabla(P_o = J)$ cannot both be true. To be precise, the potential pitfall of mine is at the very beginning where I extracted lines 3, 7, and 8 from the argument for an inconsistent tetrad and arranged them into an independent argument:

$$\begin{array}{ll} C = P_o & \\ P_o = J & \\ \text{--------------} & \\ C = J & \qquad\qquad \text{Transitivity of identity} \end{array}$$

Putnam would say this set-up is flawed in the first place because he would not approve the two premises. In particular, $P_o = J$ is the very target he tries to reject! My argument against him is therefore in trouble. So what kind of instantiations of *x*, *y*, and *z* can make TT true if we cannot directly use the terms in Putnam's argument? The message from this apparent mistake of mine suggests that we should look for some identity statements that Putnam would absolutely assent to, and we need another example in order to refute (IToI).

Therefore, I recommend looking no further than the tallest thing on Earth – Mt. Everest![192] Consider the boundary of this giant mountain. Most ordinary people cannot pinpoint where exactly the mountain starts and ends. Although we have a general idea of which part of the Earth counts as Mt. Everest, we would have difficulty deciding whether

[191] I am grateful to Douglas Campbell for discovering this glitch.
[192] The Everest example was used by Tye (1990, 1994, 2000) and Zemach (1991) in their discussions of vagueness.

a particular rock lying near the base of the incline counts as part of Mt. Everest or not. We can construct a Soritical sequence involving the boundary of this magnificent piece of land:

Rock 1 (R_1) is part of Mt. Everest.

If R_1 is part of Mt. Everest, then its immediate outside adjacent rock (R_2) is also part of Mt. Everest.

If R_2 is part of Mt. Everest, so is its immediate outside adjacent rock (R_3).

…

If R_n is part of Mt. Everest, then so is R_{n+1}.

R_{n+1} lies on a beach in Mumbai and is 1700 kilometres southwest of Kathmandu.

A rock in Mumbai is part of Mt. Everest.[193]

[193] Some might prefer the opposite description of the same Soritical difference:

Rock 1 (R_1) is not part of Mt. Everest.

If R_1 is not a part of Mt. Everest, then its immediate inside adjacent rock (R_2) is not a part of Mt. Everest either.

If R_2 is not a part of Mt. Everest, so is its immediate inside adjacent rock (R_3).

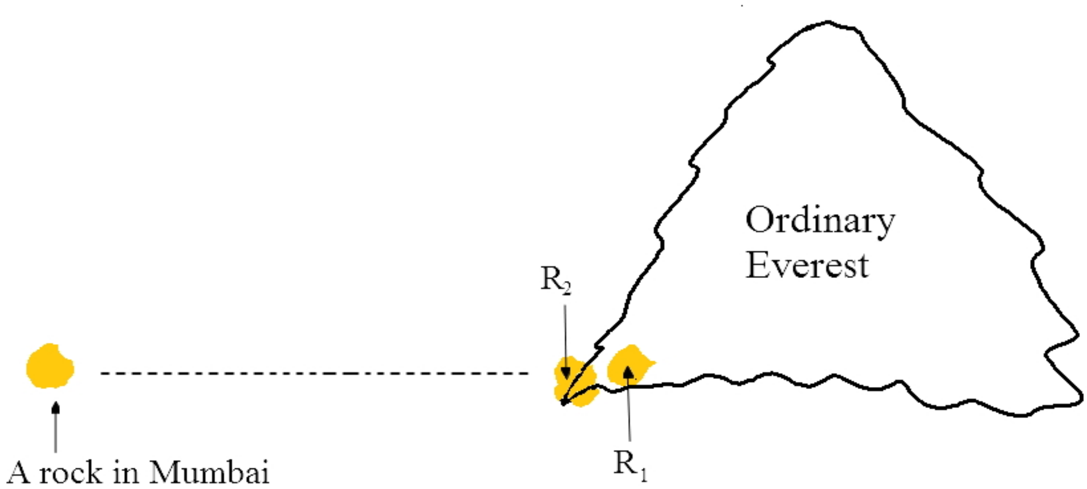
…

If R_n is part of Mt. Everest, then so is R_{n+1}.

R_{n+1} lies on the peak of Mt. Everest.

A rock on the peak of Mt. Everest is not part of Mt. Everest.

Figure 5-e

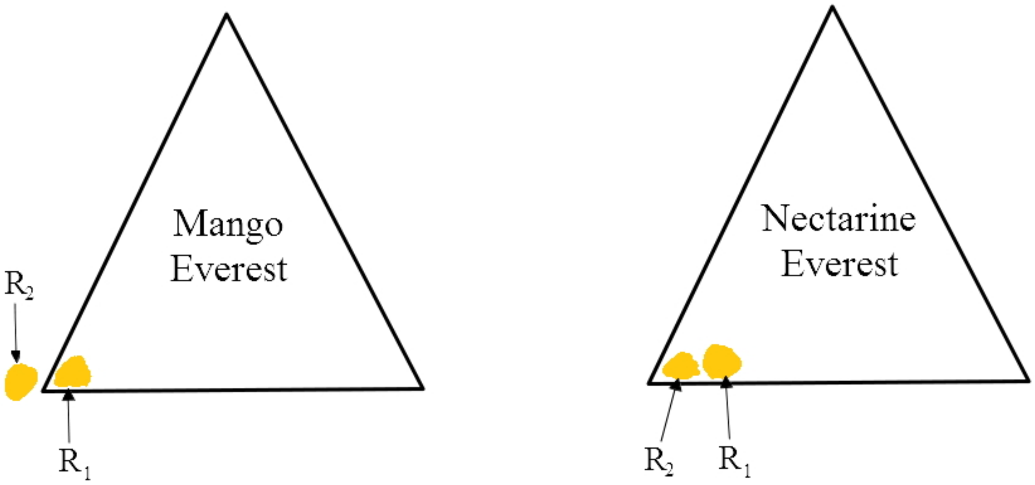


In short, ‘the boundary of Mt. Everest’ is a vague term for ordinary people. Letting OE stand for the ‘ordinary’ Everest concept, we have

$$V(OE)$$

Meanwhile, let there be two surveyors, Surveyor Mango and Surveyor Nectarine, who have taken independent measurements of the mountain base and thereby have their own, different ideas of what ‘the boundary of Mt. Everest’ refers to. For Surveyor Mango, the boundary runs through R_1, whereas according to Surveyor Nectarine the boundary runs through the adjacent rock, R_2. Let their respective Everest concepts be denoted, ME and NE:

Figure 5-f



Both ME and NE are crisp terms, for the surveyors have both pegged out exact boundaries for the mountain, and so they can't be used to generate a Soritical series. Thus, we have:

$$\Lambda(\mathrm{ME})$$

$$\Lambda(\mathrm{NE})$$

Since the two surveyors put the boundaries of the mountain in different places, we also have:

$$\mathrm{ME} \neq \mathrm{NE}$$

Since there is nothing vague in the two flanking terms, the non-identity asserted here is a determinate one, thus:

(0) $$\Delta\neg(\mathrm{ME} = \mathrm{NE})$$

This move is explained by the following 'crispness to determinacy of non-identity' principle:

(CTDN) $$\Lambda(x), \Lambda(y), x \neq y \vdash \Delta(x \neq y)$$

which in English reads

(CTDN) If both flanking terms of an identity statement are not vague (crisp), then the non-identity statement is determinate.

(CTDN) is provable from (CTD) and the equivalence principle (EP3) via sequent calculus:

Proof. $\Lambda(x), \Lambda(y) \vdash \Delta(x \neq y)$ from $\Lambda(x), \Lambda(y) \vdash \Delta(x = y)$

1. $\Lambda(x), \Lambda(y) \vdash \Delta(x = y)$ CTD
2. $\Delta(S) \dashv\vdash \Delta\neg(S)$ EP3
3. $\Lambda(x), \Lambda(y), x \neq y \vdash \Delta(x \neq y)$ 1, 2, cut

Instantiating the variables with ME and NE, we can see that the crispness of ME and NE yields the determinacy of ME ≠ NE, hence confirming the truth of (0). Next, we can use the equivalence principle (EP1S) to get (1) from (0):

(1) $\neg\nabla\neg(\mathrm{ME} = \mathrm{NE})$

OE is a vague concept, which leaves it open as to exactly where Everest's boundaries lie. Let us suppose that for all OE says about where Mt. Everest's borders lie, they might lie where ME puts them, or where NE puts them. This being so, it will be indeterminately true that Everest's boundaries, as picked out by OE, match Everest's boundaries as picked out by ME, and likewise for NE. The following pair of indeterminate identities will therefore obtain:

(2) $\nabla(\mathrm{ME} = \mathrm{OE})$

(3) $\nabla(\mathrm{NE} = \mathrm{OE})$

From (3) and symmetry of identity, we get (4):[194]

(4) $\nabla(\mathrm{OE} = \mathrm{NE})$

(2), (4) and (1) together entail (5):

(5) $\nabla(\mathrm{ME} = \mathrm{OE})\ \&\ \nabla(\mathrm{OE} = \mathrm{NE})\ \&\ \neg\nabla(\mathrm{ME} = \mathrm{NE})$

(5) is an instantiation of (TT), which is to say, it is a counterexample to (IToI). More importantly, (5) is a conjunction of three identity and non-identity statements that Putnam

[194] Here, the symmetry of identity for indeterminate identity is assumed. This assumption appears to be undeniable whereas it is highly doubtful whether transitivity works for indeterminate identity.

cannot deny, and therefore serves to prove that transitivity of identity fails for indeterminate identities. Hence, the third premise of my Argument V is correct, as are the first two:

Argument V

V1. 'Human pain' and 'octopus pain' are vague terms. 'C-fibre firings' and 'jelly firings' are crisp terms.

V2. If one flanking term of an identity statement is vague, and the other flanking term is crisp, then the identity statement is indeterminate.

V3. Transitivity of identity fails for indeterminate identity.

Sub-conclusion V4. Transitivity of identity fails for identity statements 'Human pain = C-fibre firings' and 'Octopus pain = jelly firings'.

Conclusion V5. Putnam's argument is invalid.

Finally, having established that all my premises are correct, sub-conclusion V4 follows logically, and this causes trouble for Putnam. Let us recall Putnam's argument and see how so:

Argument P-spelt-out

1. If identity theory is correct, then $(P_h = C) \& (P_o = J)$.
2. $P_h = P_o$
3. $C \neq J$
4. Identity theory is correct — Assumed for *reductio*
5. $(P_h = C) \& (P_o = J)$
6. $P_h = C$
7. $P_o = J$
8. $C = P_h$ — Symmetry of identity, 6
9. $C = P_o$ — Transitivity of identity, 8, 2
10. $C = J$ — Transitivity of identity, 9, 7
11. $(C = J) \& (C \neq J)$
12. Identity theory is not correct. — 4, 11, RAA

According to reasons described so far in this Chapter, the following derivation for line 10 does not go through:

$$\begin{array}{ll}\nabla(C = P_o) & \\ \nabla(P_o = J) & \\ \hline \nabla(C = J) & \text{Transitivity of identity}\end{array}$$

Argument P is therefore invalid.

I must concede that I do not know if the same thing can be said for line 9. As described earlier, we have not yet developed a theorem to stipulate the determinacy of identity statements with both flanking terms being vague, so I am uncertain whether $P_h = P_o$ is determinate or not. I therefore have honestly no idea if line 9 is a valid inference from lines 2 and 8. Nevertheless, either way suffices to subvert Argument P: it is invalid due to the failure of either line 9, or line 10 in which case line 9 is a valid inference.

To sum up, in this Chapter, I have analysed the determinacy status of phenomenal terms, and argued that 'P_h' and 'P_o' are vague, while 'C-fibre firings' and 'Jelly firings' are crisp. It follows, via the 'vagueness to indeterminacy' theorem, that '$P_h = C$' is an indeterminate identity. From the Mt. Everest example, I have also concluded that transitivity—the vital inference rule in Putnam's argument—does not work for indeterminate identity. It follows that Putnam's multiple realisability argument against identity theory is invalid. Apart from providing a novel way to reply to a famous objection to identity theory, the ideas in this Chapter also suggest a revamped version of identity theory, under which, mind-brain identities are held to be indeterminate identities. Envisaging the fruition of this 'new' identity theory, I foresee many criticisms and questions. In the next Chapter of my thesis, I will aim to clear these potential obstructions on our way towards indeterminate mind-brain identity.

Chapter 6. Towards indeterminate mind-brain identity (a) - questions and replies

6. 0. Outline of Chapter 6

Chapter 5 introduces a new way of replying to Putnam’s multiple realisability argument against mind-brain identity theory, namely, the indeterminacy argument (Argument V). Unlike orthodox criticisms that predominantly target Putnam’s argument’s soundness by rejecting one or more of its premises (as explained in 4.2), my Argument V, if sound, shows Putnam’s argument to be invalid. The present Chapter considers possible objections to my argument. In order to consolidate the plausibility of my reply to Putnam, this Chapter will scrutinise some serious rebuttals that might potentially render my argument unconvincing. The Chapter is divided into three sections: 6.1 will answer a question about Argument V’s validity; 6.2 explains why ‘C-fibre firing’ is crisp and thereby fully vindicates premise V1; and 6.3 defends V2 by offering putative reasons to repudiate Evans’ argument against indeterminate identities in general.

6. 1. Multiple realisability *qua* identity?

As discussed in 5.1, the indeterminacy argument targets a reconstructed version of Putnam’s argument:

<u>Argument P</u>

1. If identity theory is correct, then ($P_h = C$) & ($P_o = J$).
2. $P_h = P_o$
3. If identity theory is correct, then $C = J$. 1, 2, Transitivity, Symmetry
4. $C \neq J$
5. Identity theory is not correct. 3, 4, *modus tollens*

Recall from Chapter 5 that Argument P is motivated by the following seemingly inconsistent tetrad:

(I) $P_h = P_o$

(II) $P_h = C$

(III) $P_o = J$

(IV) $C \neq J$

According to my reconstruction, Putnam concludes on the basis of (I)-(IV) being inconsistent that (I) - the multiple realisability claim and (IV) - the non-identity between C-fibre firings and Jelly firings—are undisputed facts. He then concludes that (II) and (III)—namely, the central tenets of identity theory—must be false. My reply to Putnam was that the tetrad is not inconsistent, because (II) and (III) are indeterminate identities upon which transitivity does not work. As a result, line 3 of Argument P is not warranted.

Here is a question that could seriously undermine my entire strategy at the first place: what if Argument P is a misconstruction of Putnam's multiple realisability argument? What if proponents of the multiple realisability objection do not take (I) to be true, and hence are not motivated by the above tetrad in the first place? More precisely, what if multiple realisability, on Putnam's view, is not to be understood as identity, in contrast with what I have proposed. I will now block this possible objection with two counterarguments, either one of which is sufficient by itself to defeat the objection. Hence, I will address two questions: firstly, is the multiple-realisability-*qua*-identity interpretation of Putnam's argument a correct one? Secondly, is the alternative interpretation immune to my indeterminacy attack? For the first question, I will argue that the notion of multiple realisability as it appears in Putnam's argument must be a claim of identity. In support of this view, textual evidence from Putnam's own writings and other commentaries will be provided. For the second question, I will assume, for the purpose of the argument, that (I) is not to be construed in terms of identity (contrary to what will have just been shown). I will introduce a new kind of transitivity of identity which I contend the alternative interpretation is committed to. I will explain that if the normal kind of transitivity fails for indeterminate identities, so does the new kind, and therefore, the alternative interpretation, even if correct, would not restore validity to Putnam's argument.

6. 1. 1. Putnam on multiple realisability *qua* identity

It would be unfair to Putnam if we did not revisit his own words on this matter. Putnam's disapproval of the classical mind-brain identity theory can be found in more than one place. In 'Psychophysical Predicates', he announces:

> Similarly, the purpose of saying that pains are brain states is precisely to exclude from empirical meaningfulness the questions 'What is the pain, then, if it isn't the same as the brain state?' If there are grounds to suggest that these questions represent, so to speak, the wrong way to look at the matter, then those are grounds for a theoretical identification of pains with brain states.[195]

He then puts forward what he, at the time, believed to be the correct solution to the mind-body problem, namely, functionalism:

> I shall, in short, argue that pain is not a brain state, in the sense of a physical-chemical state of the brain (or even the whole nervous system), but another *kind* of state entirely. I propose the hypothesis that pain, or the state of being in pain, is a functional state of a whole organism.[196]

According to Putnam, the kind of identity relation that the mind-brain identity theorist is claiming is what Putnam calls *theoretical identification*. As I explained in Chapter 1, this is a term on which Lewis gives extensive exposition in his 1972 article 'Psychological and Theoretical Identifications'.[197] For Lewis, theoretical identification obtains if and only if T-terms - the terms a theory is going to implicitly imply, can be uniquely defined by sentences involving only O-terms - terms we have already understood independently of the theory. However, it must be clarified that although Putnam and Lewis use the same terminology, Lewis' intention differs drastically from Putnam's. Lewis utilises the concept of theoretical identification to endorse a version of analytic/conceptual materialism. For Lewis, mental state-terms are the T-terms, and the folk-psychological analysis of mental states is the O-terms.[198]

The bearers of theoretical identification that Putnam refers to in 'Psychological Predicates' are quite different. For Putnam, psychological predicates like pain are the T-terms, and physical-chemical correlates such as C-fibre firings are the O-terms. Putnam elucidates this in 'Minds and Machines'. In section 4 of this article, Putnam emphasises that his target is classical identity theory by saying:

[195] Putnam (1967b), p. 40.
[196] Ibid. P. 41. Author's italic.
[197] Lewis (1972).
[198] Ibid. Pp. 253-7.

> At the beginning of this paper, I pointed out that the *synthetic* character of the statement ('I am in pain if, and only if, my C-fibers are stimulated' has been used as an argument for the view that the 'properties' (or 'events' or 'states') 'having C-fibers stimulated' and 'being in pain' cannot be the same.[199]

Next, he gives a general account of what he means by theoretical identification, which is visibly the same as the Lewisian account:

> In order to do this, it is necessary to talk about one important kind of 'is' – the '*is*' *of theoretical identification*. The use of 'is' in question is exemplified in the following sentences:
>
> (2) Light is electromagnetic radiation (of such-and-such wavelengths).
>
> (3) Water is H_2O.
>
> What was involved in the scientific acceptance of, for instance, (2) was very roughly this: prior to the identification there were two distinct bodies of theory – optical theory … and electromagnetic theory.[200]

And, he recognises that the classical identity theory is a form of theoretical identification:

> Now let us try to envisage the circumstances under which a theoretical identification of mental states with physiological states might be in accordance with good scientific procedure.[201]

Seeing how Putnam expounds his target theory, at least one observation is unmistakable: Putnam is committed to at least two lines of Argument P. One is the first premise that says if identity theory is correct, then ($P_h = C$) & ($P_o = J$), and the other is the conclusion that says identity theory is incorrect. To see whether he is committed to the other premises, in particular the second one that states multiple realisability as identity, it will be necessary to revisit his 'Psychological Predicates'. The following words of his, are, in my view, a succinct expression of the multiple realisability objection:

> Finally, the [brain-state] hypothesis becomes still more ambitious when we realize that the brain-state theorist is not just saying that *pain* is a brain state; he is, of course, concerned to maintain that *every* psychological state is a brain state. Thus if we can find even one psychological predicate which can clearly be applied to both a mammal and an octopus (say 'hungry'), but whose physical-chemical 'correlate' is different in the two cases, the brain-state theory has collapsed. It seems to me overwhelmingly probable that we can do this.[202]

Once we unpack this short passage it will not be difficult to see that Putnam is explicitly making five assertions.

[199] Putnam (1975b), p. 374.
[200] Ibid. P. 379. Author's italic.
[201] Ibid. P. 380.
[202] Putnam (1967b), p. 44-5.

By saying ‘if we can find even one psychological predicate which can clearly be applied to both a mammal and an octopus (say “hungry”)’ and ‘It seems to me overwhelmingly probable that we can do this’, Putnam is in effect claiming that there is a psychological predicate that can be assigned to both human and octopus. Let ‘H’ denote this predicate, Putnam’s claim can then be expressed as follows:

(I*) H_h , H_o

To maintain neutrality at this point, I use a comma to notate the logical connective between H_h and H_o. This is to refrain from making the assumption that Putnam assigns identity to the pair. I will come back to this matter shortly and discuss whether we can replace the comma with some other connective.

Next, the phrase ‘but whose physical-chemical ‘correlate’ is different in the two cases’ is conveying three messages. The first two messages are a) that H_h has a physical-chemical ‘correlate’, and b) that H_o has a physical-chemical ‘correlate’. Since Putnam is, beyond any doubt, attacking the identity theory in this passage, we can rephrase a) and b) in terms of identity. Thus, a) says if the identity theory is correct, then H_h is identical to its physical-chemical ‘correlate’; b) says if the identity theory is correct, then H_o is identical to its physical-chemical ‘correlate’. Let ‘L’ be a predicate that stands for the physical-chemical ‘correlate’ of hunger. We can notate a) and b) as (II*) and (III*):

(II*) If identity theory is correct, then $H_h = L_h$

(III*) If identity theory is correct, then $H_o = L_o$

The phrase also makes a third assertion, which is that the two physical-chemical ‘correlates’ are *different*. The key word here is ‘different’. To me, and hopefully to everyone who understands the meaning of the word ‘different’, Putnam is clearly assigning non-identity to L_h and L_o. I have honestly no idea how it might be understood otherwise. Thus, Putnam must be saying:

(IV*) $L_h \neq L_o$

The fifth and final assertion Putnam makes is ‘the brain-state theory has collapsed’, which, without any fancy symbolisation, can be put plainly as:

(V) Identity theory is incorrect.

These five assertions make an argument, with (I*), (II*), (III*), and (IV*) as the premises and (V) as the conclusion:

<u>Argument P*</u>

(I*) H_h , H_o

(II*) If identity theory is correct, then $H_h = L_h$

(III*) If identity theory is correct, then $H_o = L_o$

(III*) $L_h \neq L_o$

(V) Identity theory is incorrect.

Now we can go back to the initial question: what should be the logical connective between H_h and H_o in (I*)? In other words, is Putnam committed to multiple realisability *qua* identity? The answer couldn't be more obvious: to make the argument valid, (I*) needs to be stating $H_h = H_o$. As a result, Argument P* will be essentially the same as my reconstruction of Putnam's multiple realisability objection, namely, Argument P. The only difference is that Argument P*—an extraction of Putnam's own words—uses 'hunger' as the example of psychological predicate, whereas Argument P uses 'pain'. Thus, since my indeterminacy attack works against Argument P, it also works against Argument P*. If all the premises of my Argument V are true, the conclusion that Putnam's argument is invalid will be true. Thus, Argument V is valid.

As shown, it is crystal clear that Putnam's multiple realisability argument must include a premise that asserts the multiple realisability of mental states *qua* identity. As a result, it should not be contentious that one of the key inference rules in Putnam's argument is transitivity of identity. I have been very surprised to discover how rarely this important point has been made explicit in the literature. I have found only two works that explicitly pinpoint the transitivity of identity as Putnam's inference rule and thereby recognise Putnam's commitment to multiple realisability *qua* identity. In their book *Philosophy of Mind and Cognition*, Braddon-Mitchell and Jackson discuss the argument:

> Thus, a functionalist approach to the mind leads to the stronger variety of mind-brain identity theory, a type-type one. Or so it seems to us; but we should emphasize that many philosophers of mind draw the opposite conclusion. They hold that functionalism with its lesson about the possibility of multiple realizability shows that the type-type identity theory is false.[203]

[203] Braddon-Mitchell, D. & Jackson, F. (2007), *The Philosophy of Mind and Cognition*, p. 102.

Note that Putnam's name isn't mentioned here and the authors only credit the argument to 'many philosophers of mind'. However, Jackson has confirmed to me via personal communication that they indeed had Putnam in mind when they wrote this book section.[204] Swapping octopuses with dolphins, Braddon-Mitchell and Jackson summarise the argument as follows:

> The problem is that different types of state might occupy, say, the pain role in different creatures … it is *C* fibres firing in humans but *D* fibres firing in dolphins. But dolphins with their *D* fibres firing would then be just as much in pain as we are when our *C* fibres are firing … But the identity theorist cannot allow both that pain = *C* fibres firing, and that pain = *D* fibres firing. That would, by the transitivity of identity, lead to the false contention that *C* fibres firing = *D* fibres firing.[205]

Jackson, Pargetter, & Prior note the same point in their 1982 'Functionalism and Type-Type Identity Theory',[206] when they summarise the multiple realisability objection:

> The view that they [Armstrong and Lewis] ought not be derives from the (in itself perfectly correct) point that the way a functional state is defined allows for the possibility, and perhaps likelihood, that different (kinds of) states occupy the same functional role in different organisms, or even in the same organism at different times. For instance, suppose that *H*-fibres' firing plays the pain-functional role in humans, but that *D*-fibres' firing plays this role in dolphins; then it cannot be the case that pain is both *H*-fibres' firing and *D*-fibres' firing, by transitivity.[207]

Despite not explicitly stating 'human pain = dolphin pain' in their synopses, the multiple realisability *qua* identity interpretation of Putnam's argument is highly visible in those words. Unpacking these premise by premise, it is easy to see that the structure of Braddon-Mitchell & Jackson and Jackson et al's reconstructed arguments are identical to mine (Argument P), and in both places the role of transitivity of identity as a crucial inference rule is highlighted:

Argument J (P_d = dolphin pain, D = D fibres)

1. If identity theory is correct, then ($P_h = C$) & ($P_d = D$).
2. $P_h = P_d$

[204] In our email correspondence, Jackson cited two reasons for not making a reference to Putnam. One, he thought the objection was so famous that a reference was not needed. Two, he in fact came to know of the objection from sources other than Putnam.

[205] Braddon-Mitchell & Jackson (2007), p. 102. Authors' italics.

[206] Jackson, F. Pargetter, R., & Prior, E. (1982), 'Functionalism and Type-Type Identity Theories' in *Philosophical Studies*, Vol. 42, pp. 209-25.

[207] Ibid. Pp. 209-10.

3. If identity theory is correct, then C = D. 1, 2, Transitivity, Symmetry

4. $C \neq D$

5. Identity theory is not correct. 3, 4, *modus tollens*

To sum up, Putnam's own presentation of the multiple realisability objection in 'Psychological Predicates' makes it clear that the argument has the form of Argument P*, which in turn shares the exact same form as Argument P. With the exception of Braddon-Mitchell &Jackson and Jackson et al, this point has not received due attention in the putative literature. However, is Argument P the right way of reconstructing Putnam's multiple realisability objection against identity theory? I believe we have enough analytic and textual evidence to confirm that the answer is obviously affirmative.

6. 1. 2. An alternative interpretation and *super-transitivity*

Strong reasons have just been given for thinking that Putnam's multiple realisability argument has the notion of identity. However, by way of carefully covering all bases, I now consider an alternative interpretation. One might argue that Argument P* (and consequently, Argument P) is not the right interpretation of Putnam, on the basis that multiple realisability is not identity and thereby the logical connective between H_h and H_o is not '='. But in order for this alternative interpretation of Putnam to work, one needs to supply new premises to form an entirely different argument. Since (III*) and the conclusion (V) is absolutely Putnam's assertion, the only possible points of error in my (and Braddon-Mitchell & Jackson and Jackson et al's) interpretation are (II*) and (III*). What might Putnam's argument look like if H_h holds a relation to H_o that is not identity? Let this obscure relation be 'R^', I think the alternative interpretation would look like the following:

Argument P^ - general

(I^) $R^\wedge(H_h , H_o)$

(II^) If identity theory is correct, then $R^\wedge(H_h , H_o) \rightarrow R^\wedge(L_h , L_o)$.

(III^) $R^\wedge(L_h , L_o) \rightarrow (L_h = L_o)$

(IV^) $L_h \neq L_o$

(V) Identity theory is incorrect.

According to this interpretation, Putnam would be arguing that if identity theory is true, then if R^ holds for human hunger and octopus hunger, then R^ also holds for their respective physical-chemical correlates. If R^ also holds for their physical-chemical correlates, then these two physical-chemical correlates must be identical. Since they are not identical, and R^ holds for human hunger and octopus hunger, identity theory is false.

Now I will examine this foreseeable way of reconstructing Putnam's in detail and draw the conclusion that even this reading still falls prey to my indeterminacy attack. To begin with, let us find out what this obscure R^ might be. One possible instantiation of R^ is evident in the following illustration. Suppose Fred is a perfectly normal human being, and Greg is an octopus. If identity theory is correct, then if Fred and Greg have *indiscernible* mental states, then they have *indiscernible* physical states. Now, Putnam's crucial point against identity theory might be that the pain of Fred the person is *indiscernible* from the pain of Greg the octopus. It follows that identity theory would then be forced to accept that Fred's physical-chemical correlate of pain is *indiscernible* from Greg's physical-chemical correlate of pain – an obviously false conjecture. In other words, under this reading, the multiple realisability argument would be as follows:

Argument P^ - specific

(I^) P_h is indiscernible from P_o

(II^) If identity theory is correct, then (P_h is indiscernible from P_o) → (C is indiscernible from J).

(III^) ¬(C is indiscernible from J).

(V) Identity theory is incorrect.

One weakness of this argument arises immediately. What exactly does it mean to say *a* is indiscernible from *b*? Does it mean the same notion of indiscernibility as in Leibniz's Law? If so, then what it really says is just that *a* and *b* share all their properties - $\forall F(Fa \leftrightarrow Fb)$. Someone who adheres to *the identity of indiscernibles* might argue on my behalf that Argument P^ is no different to Argument P, for indiscernibility entails identity. Due to the unsettledness of *the identity of indiscernibles*, I am not going to argue in this way. For the sake of this argument and my analysis thereof, I will refrain from discussing what might be the entailments of indiscernibility. I will take the word 'indiscernible' at face value, to say that *a* is indiscernible from *b* is just to say *a* is not distinguishable from *b*.

Whether or not this means that the two are identical, I don't know and I don't need to know. I am going to show that even if it doesn't (i.e. Argument P^ is not the same as Argument P), Argument P^ is flawed.

Is human pain indistinguishable from octopus pain? This is a difficult question, but let us assume the answer to be affirmative for now. Hence, let (I^) be true. Next, is C-fibre firing distinguishable from Jelly-firing? We have established that the two are not identical, but it does not follow that they are discernible unless we adopt the contentious *identity of indiscernibles*. Nevertheless, I think there are far less contentious independent reasons to believe that C-fibre firing and Jelly-firing are noticeably different. So let (III^) be true as well. The pitfall of Argument P^, I contend, resides in (II^), which has the form P → (Q → R). It is logically equivalent to (P & Q) → R. Thus, another way of putting (II^) is to say *if identity theory is correct and P_h is indiscernible from P_o, then C is indiscernible from J.* Since according to Putnam the central tenets of identity theory amount to $P_h = C$ and $P_o = J$, we can expand the premise as:

(II^)* $((P_h = C) \,\&\, (P_o = J) \,\&\, (P_h \text{ is indiscernible from } P_o)) \rightarrow (C \text{ is indiscernible from } J)$

I am now going to prove that (II^)* is false. Since (II^)* is logically equivalent to the second premise of Argument P^, it follows that my proof, if correct, shows Argument P^ to be unsound.

The main logical connective in (II^)* is '→', the material conditional. Standard logic tells us that a material conditional is false only when its antecedent is true and its consequent is false. In this case, the antecedent is a triadic conjunction of which two components are identity statements. As announced above, I am happy to grant the truth of 'P_h is indiscernible from P_o' and the falsity of the consequent—'C is indiscernible from J'. The proponent of Argument P^ will be in full agreement with me here. Her position is to claim the truth of all propositions in (II^)* except the first two conjuncts in the antecedent, namely $P_h = C$ and $P_o = J$. According to her, the conditional is true since there are false conjuncts in the antecedent that in turn render the whole antecedent false. It looks like my only chance of rejecting (II^)* is by claiming the truth of $P_h = C$ & $P_o = J$. But in doing so I would be claiming the truth of identity theory. *Petitio principii*!

The message is simple: I cannot falsify (II^)* under standard logic without committing circularity. Luckily, accepting the indeterminacy of the two identity statements in question can circumvent this. From what I have argued in the last Chapter, $P_h = C$ and P_o

= J are indeterminate identities. Will this revelation influence the overall truth-value of (II^)*? Given the framework of fuzzy logic developed by Zadeh, it will.[208] Within Zadeh's framework, an indeterminate statement *S* has a value of larger than 0 and smaller than 1; determinate statements, on the other hand, have integral values, i.e. either 0 or 1:

$$\text{Value }(\Delta(S)) = 1 \leftrightarrow ((\text{Value }(S) = 0) \vee (\text{Value }(S) = 1))$$

$$\text{Value }(\nabla(S)) = 1 \leftrightarrow 0 < \text{Value }(S) < 1$$

Now consider the values of the four component statements of (II^)*. Due to the indeterminacy therein, the value of $P_h = C$ is between 0 and 1. This is also the case for $P_o = J$. By stipulation, 'C is indiscernible from J' is determinately false, and therefore takes the value of 0. 'P_h is indiscernible from P_o' is also true by stipulation, but its determinacy status is unknown. As a result, we do not know whether its value should be 0~1 (in which case it is indeterminate) or 1 (in which case it is determinate). Either way, it is safe to say that its value is larger than 0 and smaller or equal to 1: $0 <$ Value (P_h is indiscernible from P_o) ≤ 1. Compiling these values yields a fuzzy truth-value table for all component statements of (II^)*, shown in Table 6-a:

Table 6-a

$P_h = C$	$P_o = J$	P_h is indiscernible from P_o	C is indiscernible from J
0 ~1	0 ~ 1	0 ~ 1 or 1	0

As shown, the three conjuncts that constitute the antecedent of (II^)* all have higher-than-zero values. This means that the conjunction they form also has a higher-than-zero value. Contra wise, the consequent of (II^)* is zero-valued. A conditional with a zero-valued consequent and a higher-than-zero-valued antecedent is false. Hence (II^)* and its logically equivalent proposition (II^)—a premise in Argument P^—are thus false. Argument P^ is thus unsound.

Furthermore, not only is the premise that states ((P_h = C) & (P_o = J) & (P_h is indiscernible from P_o)) → (C is indiscernible from J) false, a general diagnosis can also be given – that any conditional statements of such form are false. It should be evident that

[208] Zadeh (1975).

proposition (II^)*, which has just been shown to be false, is an instantiation of what I will call *super-transitivity of identity* (SToI):

(SToI) $(x = w)$ & $(y = z)$ & $R(x, y) \rightarrow R(w_, z)$

For any four terms w, x, y, and z, and any relation R, if $x = w$, $y = z$, and x bears a relation R to y, then w bears R to z.

Super-transitivity is a sound principle for 'normal' identities in which determinacy is usually assumed. In the case of determinate identities, from $x = w$ & $y = z$ and transitivity, $w = z$ can be inferred. It follows that the notion of super-transitivity is really just a variant of Leibniz's Law: identities share all their properties, including relational properties. However, the lesson to be learnt from the fact that (II^) is false is that (SToI) fails for indeterminate identities involving vague and crisp terms.

To see this point, let us recall the Mt. Everest example. Mango Everest (ME) and Nectarine Everest (NE) are crisp terms that refer to the giant mountain. Ordinary Everest (OE) is a vague term denoting the same mountain. It is reasonable to think that for average people who do not have knowledge of the precise boundary of Mt. Everest, their so-called OE concepts can vary. For example, my vague conception of the mountain could be very different from Jack Copeland's. Thus, it is reasonable to conjure up another vague term whose referent is not identical but very similar to OE. Call this new vague term 'Common Everest' (CE). Just like OE, CE is indeterminately identical to ME, thus:

$$\nabla(\mathrm{CE} = \mathrm{ME})$$

As explained in the earlier version of this example, OE is indeterminately identical to NE:

$$\nabla(\mathrm{OE} = \mathrm{NE})$$

We can also easily think of a relation that CE bears to OE but ME doesn't bear to NE, say the relation of *having no more than one extra rock than*. So CE has no more than one rock more than OE, and ME has two more rocks than NE. Let R be such a relation, we then have:

$$R(\mathrm{CE}, \mathrm{OE})$$
$$\neg R(\mathrm{ME}, \mathrm{NE})$$

Applying SToI to CE, OE, ME, NE, and relation R becomes a worrying move, for it is not the case that:

$$\nabla(\mathrm{CE} = \mathrm{ME}) \ \& \ \nabla(\mathrm{OE} = \mathrm{NE}) \ \& \ R(\mathrm{CE}, \mathrm{OE}) \rightarrow R(\mathrm{ME}, \mathrm{NE})$$

Another way of addressing the inapplicability of SToI here is to recognise the determinacy status of *R*(CE, OE) and *R*(ME, NE). In terms of the latter, there is no question that the determinacy status thereof is determinate, for the crispness of both ME and NE. As for the former, just as in the case of R^(P_h, P_o), the determinacy of *R*(CE, OE) is something that I am not sure of, due to the vagueness of both terms. But this hardly matters for my purpose. It is either determinately true, in which case its truth-value would be 1, or indeterminately true in which case its truth–value would be 0~1. Since the antecedent consists of a triadic conjunction in which two of them have the values of 0~1, no matter which determinacy status *R*(CE, OE) has, the overall value of the conjunction will be 'trumped' by the value of 0~1, making the value of the antecedent to be 0~1. In other words, when a conjunction is made of some determinately true conjuncts and some indeterminate ones, the determinacy status of the conjunction will be indeterminate. What we have here is therefore a case of a conditional with an indeterminate antecedent and determinately false consequent, which certainly does not seem to be a correct type of conditional.

To summarise, if one is reluctant to translate Putnam's first premise as an identity claim between human pain and octopus pain, then one has to cook up a different relation, R^, and assume Putnam is asserting R^(P_h, P_o). The second premise of Putnam's argument, as so reconstructed, would then be as follows:

$$((P_h = C) \,\&\, (P_o = J) \,\&\, R\hat{}(P_h, P_o)) \rightarrow R\hat{}(C_, J)$$

This is an instantiation of SToI. If mind-brain identity is an indeterminate identity, as I contend, then SToI fails to apply, and the assumed premise of the alternative interpretation is false. Moreover, I have also mentioned that SToI is a variant of Leibniz's Law. Due to this, my analysis in this section has a major implication: Leibniz's Law is inapplicable for indeterminate identities. This is a major feature of indeterminate identity at large. In 6.3.4, I will offer a detailed discussion of this feature and explain how it undermines Evans' argument against indeterminate identity.

As for my overall argument against Putnam, it can already be seen that my indeterminacy attack has considerable force. Not only does the multiple-realisability-*qua*-identity reconstruction (i.e. Argument P) fall prey to it, so does the alternative reconstruction (i.e. Argument P^). To be more precise, Argument P is invalid and

Argument P^ is unsound. Either way, the indeterminate mind-brain identity approach has its target right on Putnam.

At this point, one might voice a reservation about the dialectics so far. Putnam makes it clear that his objection is aimed at identity theory. Has any identity theorist claimed indeterminate identity? The answer is evidently no. So there seems to be an alarming mismatch between my purported opponent's view and Putnam's view. In reply to this worry, I concede that identity theory does not distinguish between determinate and indeterminate identity; but it should have done so, because $P_h = C$ if true, must be an indeterminate identity, as I argue. So, identity theory, charitably reconstructed, is that $\nabla(P_h = C)$. But then, since indeterminate identity is neither transitive nor super-transitive, Putnam's multiple realisability objection fails to damage identity theory.

6. 2. Is 'C-fibre firing' crisp?

Having discussed and removed the worry about my reconstruction of Putnam's argument, we can reaffirm the validity of Argument V. As for its soundness, V3 is explained in 5.4, V2 will be explained in 6.3. V1 has two statements: 1) 'Human pain' and 'octopus pain' are vague terms, 2) 'C-fibre firing' and 'jelly firing' are crisp terms. The first statement is argued in 5.2. The current section aims to vindicate statement 2).

Argument V

V1. 'Human pain' and 'octopus pain' are vague terms. *'C-fibre firing' and 'jelly firing' are crisp terms.*

V2. If one flanking term of an identity statement is vague, and the other flanking term is crisp, then the identity statement is indeterminate.

V3. Transitivity of identity fails for indeterminate identity.

Sub-conclusion V4. Transitivity of identity fails for identity statements 'Human pain = C-fibre firings' and 'Octopus pain = jelly firings'.

Conclusion V5. Putnam's argument is invalid.

It is worthwhile to revisit the dialectic so far. I am proposing a refutation of Putnam that is based on indeterminate identity, since transitivity of identity—the vital inference rule Putnam relies upon—fails for indeterminate identities. The subsequent question of interest is thus: what makes an identity indeterminate? By postulating the 'weakened vagueness to indeterminacy theorem', I suggest that the determinacy status of flanking terms (i.e. whether a term is vague or crisp) sufficiently determines the determinacy status of the identity (i.e. whether the statement is determinate or indeterminate) that holds between these flanking terms. In 5.4, I discussed three possible combinations of the determinacy status of flanking terms:

1) Λa & Λb
2) Λa & Vb
3) Va & Vb

I concluded that the determinacy status of their corresponding identity statement would be as follows:

Table 5-d

	Λa	Va
Λb	$\Delta(a = b)$ – explained by (CTD)	$\nabla(a = b)$ – explained by (VTIW)
Vb	$\nabla(a = b)$ – explained by (VTIW)	? (except for the case of a = a, which is Δ)

Here, an important fact is visible: indeterminate identity occurs when there is a difference in the determinacy status of terms – that is, when one flanking term is vague and the other flanking term is crisp. However, as the table shows, there is one unsolved mystery: it remains inconclusive whether indeterminate identity *only* occurs when the terms each have a different determinacy status. To be precise, the uncertainty is about the vague-vague combination, barring the special self-identity cases where 'Va = Va' is always determinate. As I announced in 5.4, I do not wish to pursue the truth on this matter. It follows that, to accommodate my agnosticism about the determinacy status of 'Va =Vb', and to maintain the soundness of my argument against Putnam, I need to establish that 'P_h' and 'C' each have a different determinacy status so that they can form indeterminate identity $\nabla(P_h = C)$. Since I have argued that 'pain' is vague in 5.2, the remaining piece of the puzzle is therefore to show that 'C-fibre firing' is crisp. Note that the issue of whether or not C is a vague or crisp term is not relevant to either identity theory or the functionalist's objection to identity theory. I may assume for the present purpose that C is crisp. Nevertheless, it will be best to not only assume but to make the case for it.

One clarification needs to be made immediately though. As I indicated at the end of 5.3.2, 'C-fibre firing' is a traditional term that many philosophers use to refer to whatever the neurophysiological state that pain is purportedly identical to.[209] Contra to the view advocated by the likes of Puccetti[210] and Hardcastle,[211] I think whether or not pain is in fact C-fibre firing is of no philosophical importance at all. We shall, instead, inquire of whether or not pain is the neurophysiological state(s) that the best science says it is, for 'C-fibre firing' is just a folk term that generalises these neurophysiological state(s). Following this reasoning, it is easy to see that the real question that needs to be investigated in this section is: is the neurophysiological state(s) that pain is purportedly identical to crisp? To reveal the answer, I will firstly provide a brief overview of the neurophysiology of pain. Secondly, I will show evidence that these neural-physical states that pain corresponds to have the predicate of being crisp.

6. 2. 1. Neurophysiology of pain

First of all, the abovementioned view pushed by Puccetti calls for the abandonment of the use of 'C-fibre firing' or 'C-fibre stimulation' as the candidate neural process for pain. This view is right about one thing: pain, as current science has revealed, is in fact a highly complex neurophysiological phenomenon. Despite the fact that even state-of-the-art understanding of pain is still fragmentary, one consensus is that there is more than one pathway leading to the activation of pain. According to Kandel, Schwartz, and Jessell, three types of neuron subsystems are collectively considered to be the neurophysiological system of pain:[212]

- *Thermal nociceptors* – these are responsible for pain caused by extreme temperature (higher than 45°C or lower than 5°C), and are subserved by small-diameter, thinly myelinated fibres known as Aδ-fibres. Because they have a relatively small diameter,

[209] The trend started from Rorty, R. (1965) 'Mind-body Identity, Privacy, and Categories' in *The Review of Metaphysics*, Vol. 19, pp.24-54.

[210] Puccetti, R. (1977), 'The Great C-Fiber Myth' in *Philosophy of Science*, Vol. 44, pp. 303-5.

[211] Hardcastle, V. (2001), 'The Nature of Pain' in W. Bechtel et al. (eds.) *Philosophy and the Neuroscience*, pp. 295-311.

[212] Kandel, E. Schwartz, J. and Jessell, T. (eds.) (2000). *Principles of Neural Science*, pp. 472-5.

information travels fast along these fibres, at about 5-30 m/s. They are therefore responsible for rapid pain.

- *Mechanical nociceptors* – these are responsible for pains caused by intensive pressure applied to the skin, such as being punched by someone. Like *thermal nociceptors*, the corresponding neurons are Aδ-fibres. As a result of their short diameter, the pain you get from being punched in the face is also sharp and fast.
- *Polymodal nociceptors* – this subsystem is activated by multi-modal stimuli including mechanical, chemical, and thermal. It consists of the infamous C-fibres, which contrary to Aδ-fibres, are non-mylinated. Due to this, information travels relatively slowly along C-fibres, at a speed of less than 1.0 m/s. As a result, C-fibres are responsible for pains that are slow and dull.

From the above, we can see the so-called 'philosophy's error'[213] in referring to pain being C-fibre firing, for pain is in fact the firing of C-fibres and/or Aδ-fibres.[214] Furthermore, the 'philosophy's error' is even more severe when we consider the location of these pain-responsible neural subsystems. The cell bodies of these neurons are in fact located in the *dorsal root ganglia* and *trigeminal ganglia*. In other words, these neural firings are not exclusively brain states at all, but brain and/or spinal states. Nevertheless, this shouldn't cause any great harm to the identity theorists or anyone who supports the idea that 'pain = a certain type of brain state', for the purported identity claim can be easily modified to 'pain = a certain type of neurophysiological state', without altering the basic tenets of identity theory.

Having learnt who and where the somatosensory neurons are, I now describe how the firing of these neurons works. To say that somatosensory neurons, as well as other neurons, activate or 'fire' is another way of saying that they send electro-chemical signals. As Zupanc summarises, a neuron is at rest (that is to say when it is not sending these signals) when there is a *differential distribution* of the sodium, potassium, and calcium ions residing inside and outside the membrane that surrounds the nerve cell.[215] At this resting stage, known as *resting potential*, there are more sodium ions outside the nerve cell and more potassium ions inside. As a result, the inside of the nerve cell is negatively charged, relative to the outside. To be precise, the inside is about 60 to 80 mV

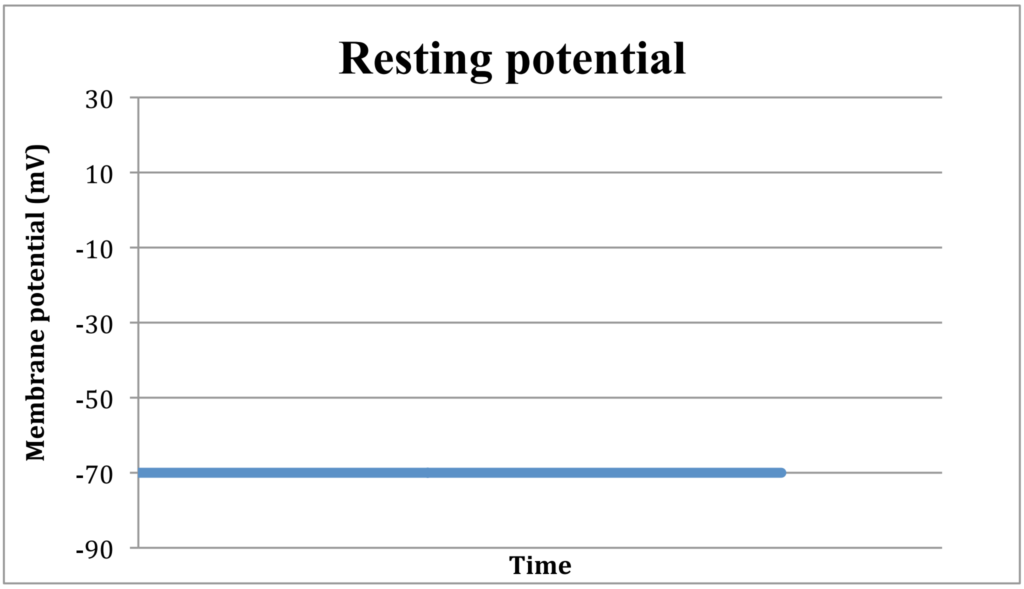
[213] A term dubbed by Hardcastle in her (2001), p. 300.

[214] For simplification, I have omitted the Aβ-fibres, which is a group of large-diameter fibres that are activated only occasionally. This omission makes no negative impact on the identity theory.

[215] Zupanc, G. (2010), *Behavioural Neurobiology*, pp.17-22.

(millivolt) less than the outside, creating a potential difference of -60 ~ -80mV. Thus, the membrane of the nerve cell is *polarized.*

Chart 6-b



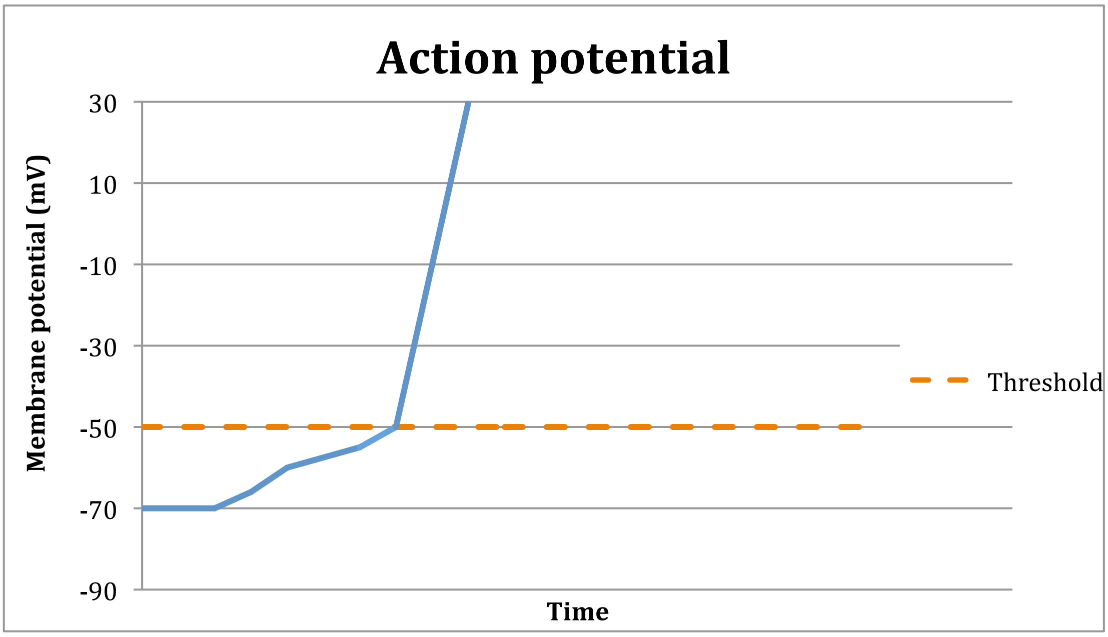
Once the *polarizing* distribution of sodium and potassium ions is tipped, *depolarization* occurs. That is, some stimuli (e.g. pinching of the skin) cause the ions to exchange across the inside and outside of the membrane. In most cases, this will be an influx of potassium ions from the inside to the outside, which leads to a net increase of the membrane potential towards 0mV. As the depolarization continues, this increase of voltage will be linear with the increase in current. But once it reaches a critical value – about -50mV, it triggers '*an explosive, all-or-nothing event*' known as the *action potential.*[216] This type of event is distinctive in that there is a critical threshold and once reached the neuron will always fire. On the other hand, if it is not reached, then no action potential will fire. More importantly, this distinctive feature generalises: no matter how intensive the stimuli (how large the currents are), the size of the action potential is always the same. In short, there are no big or small neural firings; neurons either fire or do not fire.

6. 2. 2. Neural activation function as hybrid function

The following chart illustrates how action potentials occur.

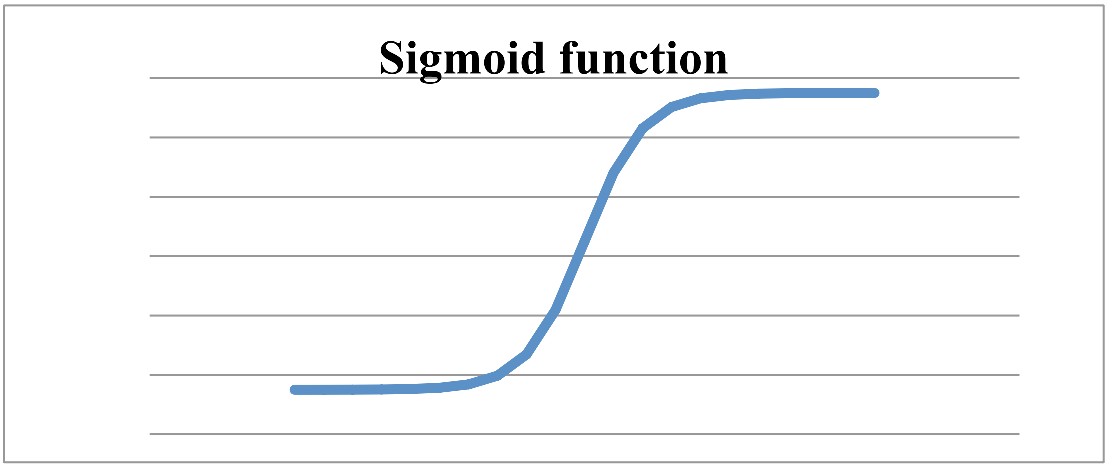
[216] Nicholls et al. (2001), *From Neuron to Brain*, p.14.

Chart 6-c



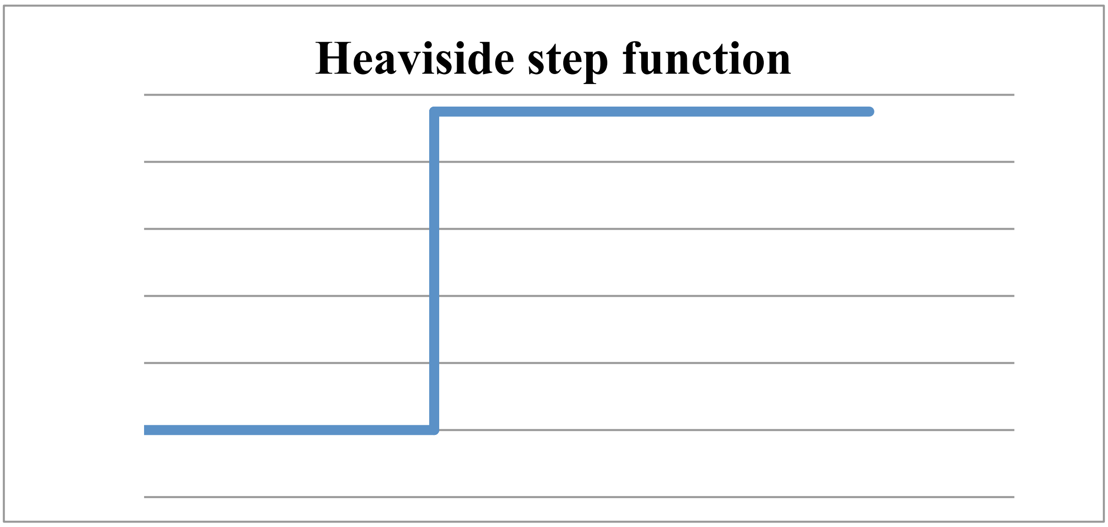
Based on Chart 6-c, we can scale up and define the activation function of neural firing. Usually, the approximation of a Chart 6-c-shaped function will be a *sigmoid function* where α is the point at which the stimuli kicks in and the membrane potential starts to increase towards 0m, and β is the point where action potential is reached and the neuron starts to fire:

Chart 6-d



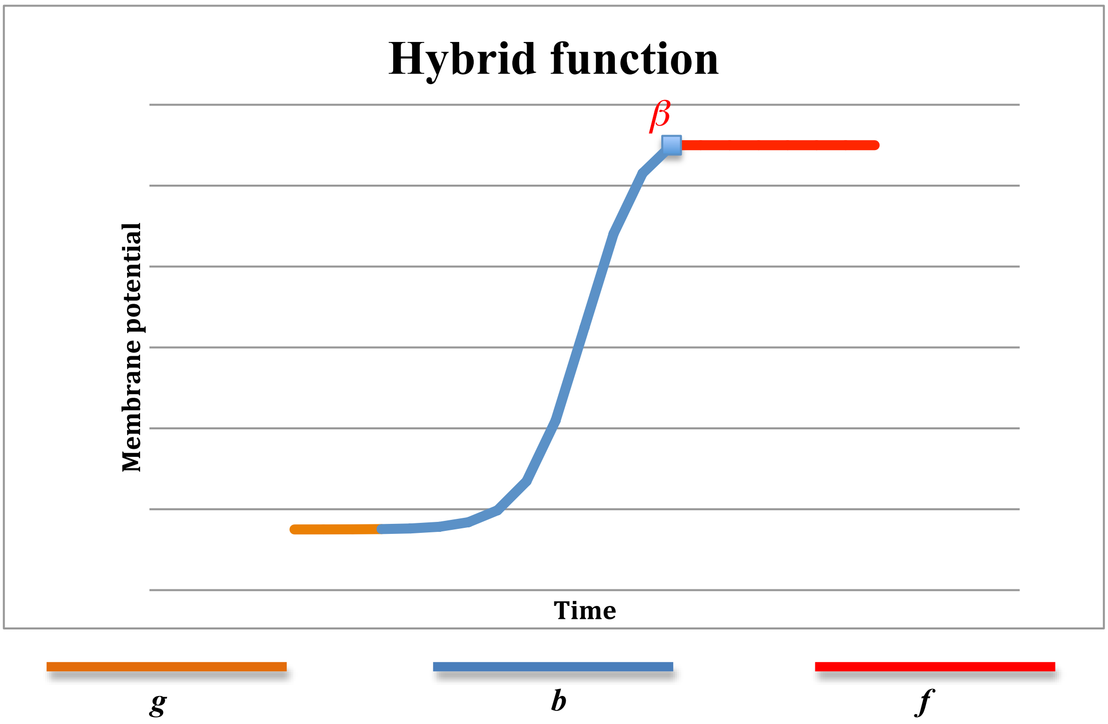
However, due to the all-or-nothing nature of neural firing, the neural activation function seems more akin to the *threshold function* or the *Heaviside step* function where a sharp point differentiates the firing stage from the non-firing state:

Chart 6-e



But the Heaviside step interpretation of neural firing is only half correct since it misrepresents the build-up period. Instead of sigmoid and Heaviside step functions, I suggest the more accurate way to define the neural activation function is a hybrid shape where the first half (that is when the value of y is below β) is a curve, and the second half (that is when the value of y is equal to or larger than β) resembles a step:

Chart 6-f



As Chart 6-f shows, let *g* be the ground function where resting potential is obtained, and *b* is the build-up function where voltage gradually increases as a consequence of stimuli, and f is the firing function where β, the critical threshold, is reached. Presented this way, it can be seen that neurons, including somatosensory ones, either do not reach the threshold of β and therefore do not fire, or do reach β and fire. We might not be able to pinpoint the exact value of β, and it could vary from one type of neuron to another. What can be concluded is that there is a sharp point at which neural firing in general and—for our purpose, pain-responsible neural firing—happens or not. A crisp definition of C-fibre firing can be defined based on Chart 6-f:

$$f = 0 \leftrightarrow y < \beta$$
$$f = 1 \leftrightarrow y \geq \beta$$

There is simply nothing uncertain about whether or not firing occurs when membrane potential is at β - 0.00000001 or β + 0.00000001. In short, there is no room for neural firings to be tolerant. Terms that denote all kinds of neural firing, including 'C-fibre firing' and 'jelly firing' are therefore crisp terms.

Returning to the big picture, i.e. my argument against Putnam, we can now affirm the truth of V1:

V1. 'Human pain' and 'octopus pain' are vague terms. *'C-fibre firing' and 'jelly firing' are crisp terms.*

6. 3. Can there be indeterminate identities?

Having confirmed the plausibility of V1, I now move on to examine the plausibility of V2:

Argument V

V1. 'Human pain' and 'octopus pain' are vague terms. 'C-fibre firings' and 'jelly firings' are crisp terms.

V2. If one flanking term of an identity statement is vague, and the other flanking term is crisp, then the identity statement is indeterminate.

V3. Transitivity of identity fails for indeterminate identity.

Sub-conclusion V4. Transitivity of identity fails for identity statements 'Human pain = C-fibre firings' and 'Octopus pain = jelly firings'.

Conclusion V5. Putnam's argument is invalid.

As noted above, V2 amounts to the 'weakened vagueness to indeterminacy' theorem (VTIW) that I introduced in 5.3. Thus far we have been proceeding on the assumption that VTIW is true. In demonstrating that repudiating the transitivity of indeterminate identity (IToI) can undermine the multiple realisability argument, I have so far only explained why VTIW is an improvement on VTI. The current section and Chapter 7 aim to finally establish the truth of V2 and the soundness of my argument against Putnam. The current section is going to scrutinise a famous objection to V2. Chapter 7 will attempt to explain why V2 is plausible via a proper proof of an updated version of the vagueness to indeterminacy theorem.

In a nutshell, I am going to present two independent arguments in establishing the truth of V2:

Negative argument – I will explain what is wrong with Evans's argument for the general absurdity of indeterminate identities.

Positive argument – I will explain why and how vagueness of term(s) leads to indeterminate identity that holds between these terms.

6. 3. 1. The problem

To begin with, let us revisit Evans' proof as described in 5.3.1:

(EV1)	$\nabla(a = b)$	Assumed for *reductio*
(EV2)	$\lambda x(\nabla(x = a))b$	EV1, property abstraction
(EV3)	$\neg\nabla(a = a)$	Axiom
(EV4)	$\neg\lambda x(\nabla(x = a))a$	EV3, property abstraction
(EV5)	$\neg(a = b)$	EV2, EV4, Leibniz's Law
(EV6)	$\neg\nabla(a = b)$	EV1, EV5, *reductio*

As shown earlier, philosophers have different views regarding which position Evans' purported proof was really against. Broadly, there are two views that can be drawn:

Interpretation 1. Evans intends to prove that there cannot be indeterminate identities as a result of flanking terms denoting vagueness.

This view is overwhelmingly popular. For example, citing Lewis, Noonan wrote:

> As David Lewis has stressed in his [1988], Evans is not against the idea that there can be identity *statements* which are indeterminate in truth-value… Evans's target is rather the view that there can be identity statements which are indeterminate in truth-value not because of any semantic indeterminacy but rather because of indeterminacy *in the world*, that is, *for no other reason* than that one or both of the objects *determinately* denoted by the singular terms flanking the identity sign is a vague object.[217]

Along a similar line of thought,[218] Garrett wrote:

> Evans is concerned, not with the question of whether there can be vague identity statements, but with the question of whether there can be vague objects, i.e. vague identity statements the singular terms of which do not have their references fixed by vague descriptive means. Evans's proof – on this interpretation – purports to demonstrate that it cannot be indeterminate whether *a* is *b* if neither '*a*' nor '*b*' have their references fixed by descriptive means.[219]

Garrett (1988, 1991), Lewis (1988), Noonan (1982, 1984, 1990, 1991) have all advocated and maintained this interpretation of Evans.

In contrast, one might take up a stronger interpretation:

Interpretation 2. Evans intends to prove that there cannot be indeterminate identities.

As Copeland elegantly puts: 'In a five-line derivation, Gareth Evans reduced to absurdity the assumption that some identity - any identity - is indeterminate, or so he claimed.'[220]

[217] Noonan (1990), pp. 157-8. The point was also stated in Noonan, (1991), p. 183.

[218] There is a significant difference between Noonan's and Garrett's understanding of vagueness. The former thinks vagueness is located in the objective world, whereas the latter believes vagueness resides in language. Noonan noted the distinction explicitly by saying:

> It should be noted then that Garrett is wrong when he writes in his [1988] 'The thesis that there can be vague objects is the thesis that there can be identity statements which are indeterminate in truth-value as a result of vagueness *the singular terms of which do not have their reference fixed by vague descriptive means*', for the existence of vague objects could be at most a necessary condition of the existence of identity statements which are indeterminate in truth-value in the way he characterizes. (1990, p. 160)

Here it is worthwhile to point out that this distinction makes no difference to Interpretation 1.

[219] Garrett (1988), p. 131.

[220] Copeland (1997), p. 514.

My interest here is not to determine which of the two reflects the real intention of Evans'. Rather, I wish to bring attention to the fact that under both interpretations Evans' proof would cast doubt over my premises and could thereby make my argument against Putnam unsound.

Argument V

V1. 'Human pain' and 'octopus pain' are vague terms. 'C-fibre firings' and 'jelly firings' are crisp terms.

V2. If one flanking term of an identity statement is vague, and the other flanking term is crisp, and they are not determinately non-identical, then the identity statement is indeterminate.

V3. Transitivity of identity fails for indeterminate identity.

Sub-conclusion V4. Transitivity of identity fails for identity statements 'Human pain = C-fibre firings' and 'Octopus pain = jelly firings'.

Conclusion V5. Putnam's argument is invalid.

Evans' proof makes trouble for my second premise, V2, and this is regardless of how you interpret his intention. Under Interpretation 1, Evans would be opposing (VTI) and its weakened forms (VTIW) and (VTIWD) - which is exactly what V2 amounts to. Under Interpretation 2, Evans would be opposing the possibility of all indeterminate identities including $\nabla(P_h = C)$ and $\nabla(P_o = J)$ – which are the hidden sub-conclusions of my argument. Therefore, by way of defending my argument I must find a flaw in Evans' purported proof. To me, there are four types of reply to Evans that can potentially render it unconvincing. But before we get to those, let us survey some of the early attempts against Evans' proof.

6. 3. 2. Honourable mentions

Richmond Thomason (1982) provides a criticism to Evans that draws attention to the transition from EV3 to EV4:

> The fallacy is analogous to the modal one that would have been committed had ∇ been interpreted as 'neither necessarily true nor necessarily false'. In the modal case $\nabla\ [a = a]$ is equivalent to $\lambda x \nabla\ [x = a]\ (a)$ only if a is a "rigid designator"; so to assume the equivalence in arguing for $a = b \rightarrow \Box\ [a = b]$ is to beg the question. In the case of vague singular terms, $\neg \nabla\ [a = a]$ is equivalent

> to $\neg \lambda x \nabla [x = a] (a)$ only if a is a "precise designator", and when Evans infers the second from the first he is assuming what he is trying to prove.[221]

The same 'fallacy' was also discussed and elaborated in Lewis (1988):

> The operator 'it is vague whether…' is analogous to an operator of contingency, and means 'it is true on some but not all of the precifications that…'. [Evans' transition from EV3 to EV4] is analogous to the fallacious modal equivalence between 'It is contingent whether the number of planets is nine' (true) and 'The number of planets is such that it is contingent whether it is nine' (false), or between 'It is contingent whether the number of planets is the number of planets' (false) and 'The number of planets is such that it is contingent whether it is the number of planets' (true).[222]

This reply accents the fact that the following property abstraction move with reference to modality is fallacious unless *a* is a rigid designator:

(TL1) $\Diamond(a = b)$

(TL2) $\lambda x(\Diamond(x = b))a$ TL1, λ-abstraction

Analogously, as Thomason claims, Evans' EV3 to EV4 is fallacious unless the determinacy status of *a* is presupposed to be crisp (or precise, as they both prefer to say):

(EV3) $\neg\nabla(a = a)$ Axiom

(EV4) $\neg\lambda x(\nabla(x = a))a$ EV3, λ-abstraction

While it is true that TL1 can only entail TL2 when *a* is rigid, why is it the case that EV3 can only entail EV4 if *a* is crisp? Neither Thomason nor Lewis has offered an adequate answer. It is almost customary to read $\neg\lambda x(\nabla(x = a))a$ as saying '*a* does not have the property of being indeterminately identical *to a*', which is fine. However, it is equally unproblematic to read it as '*a* does not have the property of being indeterminately identical *to itself*'. If the second reading is taken the Thomason reply is easily rebuffed, because the property abstraction would go through regardless of *a*'s determinacy status. Interestingly, this second reading of $\neg\lambda x(\nabla(x = a))a$ also motivates the most powerful objection against Evans' proof, which I will examine at the end of this section.

[221] Thomason (1982), p. 331. Author's italics. Instead of "x^", I have change the symbol for property abstraction to "λx".

[222] Lewis (1988), p. 129. It is worth clarifying that the number of planets was discovered to be eight in 2005. Furthermore, unlike Thomason, Lewis does not think this objection against Evans works.

Another honourable mention goes to John Broome's (1984) in which he describes '*a good prima facie example of indefinite identity*'.[223] Adding slight changes, we can summarise Broome's example as follows:

Suppose a knitting club was founded in year 1975 with fifty members. They had knitted together on every Tuesday night for five years, and then they stopped doing so in year 1980 without signalling the club's official demise. Since then, there were no club activities until year 1990. In 1990 some original members but not all, say twenty, met up and decided to re-engage in Tuesday night knitting again. This continues for another twenty-five years. In 2015 all original members died and there have been no more club activities since.

Now, consider the following identity statement:

(KC) The knitting club in 1975 = the knitting club in 1990

Is KC determinate (i.e. determinately true or determinately false) so we can add to it the determinacy operator 'Δ'?

(KC1) Δ (The knitting club in 1975 = the knitting club in 1990)

Or is it indeterminate so we shall add to it the indeterminacy operator '∇':

(KC2) ∇ (The knitting club in 1975 = the knitting club in 1990)

Broome suggests KC1 is false, for if it were true then there must be enough detail of the club's constitution to which any concerns about the club's identity can be decided unanimously. But this requirement of preciseness in the club's constitution is unlikely to obtain, since it is rarely the case in reality. Instead, Broome thinks KC2 is true. He believes the example showcases indeterminate identity for the lack of '*sharp division between*'[224] the reference denoted by the flanking terms. Furthermore, he makes it clear that the indeterminateness is not a result of epistemic matter because no additional information would alter the indeterminateness of the statement. The indeterminacy, therefore, must reside in the identity.[225]

[223] Broome (1984), p. 7.
[224] Ibid.
[225] Ibid. Pp. 6-7.

Several texts have regarded Broome's club as an unsuccessful counterexample to Evans' proof. Tye (2000) thinks members' intention is a key factor in deciding the club's identity.[226] If the twenty original members simply met again with no intention to resurrect the old club then their meeting-again should be considered as creating a new club. If this scenario is true, then KC is determinately false, hence KC1 is true. On the other hand, the more natural scenario is that the twenty members intend to resume their knitting club after a 10-year hiatus, therefore KC is true, and KC1 is also true. However, for Tye's reply to stand, the club's constitution has to include a precise rule regarding members' intention, which as Broome explains, is quite impractical and unlikely.

A more promising criticism of Broome's club is due to Noonan's (1984), where he sights an unpalatable consequence of accepting the truth of KC2.[227] Consider the following two predicates: 'lasted for at most five years' (L5) and 'lasted for at least twenty-five years' (L25). Usually, we are very much inclined to say the former is determinately true of 'the knitting club in 1975', and the latter is determinately true of 'the knitting club in 1990':

(KC3) Δ(L5)(the knitting club in 1975) & (L5)(the knitting club in 1975)

(KC4) Δ(L25)(the knitting club in 1990) & (L25)(the knitting club in 1990)

In virtue of meaning, we also know for sure that L5 and L25 are not co-extensive, thus L5 and L25 are conjointly true of nothing. It follows that L25 is determinately false of 'the knitting club in 1975', and L5 is determinately false of 'the knitting club in 1990':

(KC5) $\Delta\neg$(L25)(the knitting club in 1975) & $\neg$(L25)(the knitting club in 1975)

(KC6) $\Delta\neg$(L5)(the knitting club in 1990) & $\neg$(L5)(the knitting club in 1990)

However, according to Noonan, the truth of KC2 would tarnish the apparent joint truth of KC3-KC6.

(KC2) ∇(The knitting club in 1975 = the knitting club in 1990)

For if the two terms are identical, then by Leibniz's Law, KC3 and KC6 cannot be jointly true, as cannot KC4 and KC5. Furthermore, as Noonan explains, we are then forced to

[226] Tye (2000), p. 206.

[227] Noonan (1984), pp. 119-120.

say that ‘the knitting club in 1975’ determinately satisfies neither L5 nor L25, and the same can be said for ‘the knitting club in 1990’. Hence, the following conjunctions would be true as a result of KC2 being true:

(KC7) $\neg(\Delta(L5)(\text{the knitting club in 1975}) \ \& \ (L5)(\text{the knitting club in 1975})) \ \& \ \neg(\Delta(L25)(\text{the knitting club in 1975}) \ \& \ (L25)(\text{the knitting club in 1975}))$

(KC8) $\neg(\Delta(L5)(\text{the knitting club in 1990}) \ \& \ (L5)(\text{the knitting club in 1990})) \ \& \ \neg(\Delta(L25)(\text{the knitting club in 1990}) \ \& \ (L25)(\text{the knitting club in 1990}))$

However, a disjunctive predicate featuring L5 and L25 must be determinately true of both flanking terms, so:

(KC9) $\Delta(L5 \vee L25)(\text{the knitting club in 1975})) \ \& \ (L5 \vee L25)(\text{the knitting club in 1975}))$

(KC10) $\Delta(L5 \vee L25)(\text{the knitting club in 1990})) \ \& \ (L5 \vee L25)(\text{the knitting club in 1990}))$

Accepting the joint truth of KC7 and KC9 (likewise, KC8 and KC10) is to accept that a disjunctive predicate can be determinately true of *x* while neither disjunct is determinately true of *x*. Noonan illustrates the logical haziness here with his own example of a person who is determinately either a child or elderly while being neither determinately a child nor determinately elderly. To Noonan, such a consequence is ‘hard to understand’[228] and the remedy therefore is to abandon KC2.

I find Noonan’s reply to Broome unsatisfactory for two reasons. Firstly, Noonan claims that the triad of KC3, KC6, and KC2 (and KC4, KC5, and KC2) is inconsistent. The underlying assumption here is a result of applying Leibniz’s Law to KC2 – an indeterminate identity statement. Unfortunately for Noonan, the application of Leibniz’s Law in indeterminate identities equates to conceding the ‘super-transitivity’ of identity, which for reasons I specified in 6.1.2, is false. Secondly, Noonan’s argument correctly identifies that the tetrad of KC3-KC6 is incompatible with KC2, but incorrectly concludes from there the inevitable denial of KC2. I suggest that one need not reject KC2 had he

[228] Noonan (1984), p. 120.

known the *principle of difference*, which I will discuss in full detail in 7.1. Instead, it is the joint truth of KC3-KC6 that needs to go.

The principle of difference says there is no difference in the determinacy status of terms without a difference in the determinacy status of statements, and the difference in the determinacy status of terms must manifest at the level of predicates:

(PoD) $\mathrm{V}x, \Lambda y \vdash \exists F\ (\nabla Fx\ \&\ \Delta Fy\ \&\ Fy)$

That is, in the case of indeterminate identity as a result of a vague flanking term and a crisp flanking term, there must be a predicate which is indeterminately true of the vague term and determinately true of the crisp term. Now let us recall KC2 – an indeterminate identity statement. If it falls into the category of having both a vague and a crisp flanking term, then by PoD there must be a predicate true of both flanking terms separating the determinacy status. Suppose 'the knitting club in 1990' is vague and 'the knitting club in 1975' is crisp, we then have:

$\exists F$ (∇F(the knitting club in 1990) & ΔF(the knitting club in 1975) & F(the knitting club in 1975))

This provides a reason for rejecting the joint truth of KC3-KC6. A predicate such as L5 can be said to be the differentiating predicate, hence yielding:

∇(L5)(the knitting club in 1990) & Δ(L5)(the knitting club in 1975) & (L5)(the knitting club in 1975)

which falsifies KC6 from the equivalence principle. Once his premise that KC3-KC6 are jointly true gives away, Noonan's conclusion that KC2 is false will collapse.

Critics might say that my attempt to rebut Noonan's argument is entirely *ad hoc*. it is true that my argument relies on two assumptions: 1) that the determinacy status of the flanking terms in KC2 must vary (i.e. one vague, one crisp); and 2) that predicates describing temporal duration are suitable candidates for F – the differentiating predicate in PoD. To vindicate my rebuttal, I need to offer independent reasons for both 1) and 2). In this regard, I concede that my rebuttal is conditional. Thus, one shall not take my argument as claiming that KC3-KC6 *are not* jointly true while KC2 *is true.* Instead, I am merely detailing how the joint truth of KC3-KC6 can be rejected, *given certain conditions*. In short, I am claiming that it is *possible* to deem KC3-KC6 false and KC2 true. Noonan's criticism of Broome's, on the contrary, relies on the exclusion of this possibility.

For aforementioned reasons, I believe Broome's club describes a genuine *possibility* of indeterminate identity. The specified scenario need not *guarantee* the indeterminacy of KC. This mere possibility of KC2 being true suffices to counter Evans' conclusion that no identities are indeterminate. This, I think, is the intent and also the upshot of Broome's club example. However, the club example only targets Evans' conclusion and tries to refute it by giving a counterexample. As far as trying to expose the flaws in Evans' attempted proof, Broome's example is an innocuous reply. In order to attack Evans' logic, we need to investigate the form of his attempted proof.

6. 3. 3. Deducing the conclusion

In what follows, I will discuss the four major objections to the logic of Evans' attempted proof in what I believe to be the ascending order of strength. The first objection targets the conclusion EV5 and asks how it refutes EV1. It focuses on the lack of apparent contradiction between the two lines. Evans' proof is supposed to falsify EV1 via *reductio* and should therefore end with a negation of the initial assumption:

(EV1)	$\nabla(a = b)$	Assumed for *reductio*
(EV2)	$\lambda x(\nabla(x = a))b$	EV1, property abstraction
(EV3)	$\neg\nabla(a = a)$	Axiom
(EV4)	$\neg\lambda x(\nabla(x = a))a$	EV3, property abstraction
(EV5)	$\neg(a = b)$	EV2, EV4, Leibniz's Law
(EV6)	$\neg\nabla(a = b)$	

The proof fails to be a *reductio* without the addition of EV6.[229] Nonetheless, adding this extra conclusion isn't going to be a problem for Evans if he can justify the transition from EV5 to EV6. He seems to justify it in the following passage:

[229] There is another, milder slip of Evans'. Namely, the transition from (EV1) to (EV2) relies on symmetry of identity. But as reported by Parsons (2000), this missing step is of no significant importance.

> If ‘Indefinitely’ and its dual, ‘Definitely’ (‘Δ’), generate a modal logic as strong as S5, (1)-(4) and, presumably, Leibniz’s Law, may each be strengthened with a ‘Definitely’ prefix, enabling us to derive:
>
> (5′) $\Delta \sim (a = b)$
>
> which is straightforwardly inconsistent with (1).[230]

Here, Evans’ reasoning appeals to a parity between the delta-operators ‘Δ’ and ‘∇’, and the modal operators ‘$\Box$’ and ‘$\Diamond$’.[231] In modal system S5, the following necessitation rule is permissible:

$$S \rightarrow \Box(S)$$

By the parity of reasoning assumed by Evans, a delta version of this necessitation rule is also permissible:

$$S \rightarrow \Delta(S)$$

Thus EV5’ (which Evans called (5’)) can be entailed by EV5 via the following rule:

(IoD) $\neg(a = b) \rightarrow \Delta\neg(a = b)$

Call this the ‘introduction of determinacy’ rule (IoD for short),[232] the last two steps of Evans’ proof become:

(EV5) $\neg(a = b)$

(EV5′) $\Delta\neg(a = b)$ *EV5, Introduction of determinacy*

(EV6) $\neg\nabla(a = b)$

So far so good. Now the gap that Evans needs to fill in is to explain the transition from EV1 ~ EV5’ to EV6. This is done via the equivalence principles EP2:

(EP2) $\nabla(S) \leftrightarrow \neg\Delta(S) \vee \neg\Delta(\neg S)$

We can deduce EV5″ from EV1 and EP2:

(EV1) $\nabla(a = b)$

[230] Evans (1978), p. 208.

[231] This is also evidenced by the letter to David Lewis, in which Evans stated that the delta operators were intended as modal operators, as confirmed by Pelletier (1989) and Garrett (1991).

[232] This is equivalent to the ‘Evans’ axiom’ coined by Copeland (1994).

(EV5″)	$\neg\Delta\neg(a = b)$	*EV1, EP2*

Since EV5″ is a formal contradiction of EV5’, the negation of EV1 is yielded. Hence, filling out the details, Evans’ proof has a proper form of *reductio*:

(EV1)	$\nabla(a = b)$	Assumed for *reductio*
(EV2)	$\lambda x(\nabla(x = a))b$	EV1, property abstraction
(EV3)	$\neg\nabla(a = a)$	Axiom
(EV4)	$\neg\lambda x(\nabla(x = a))a$	EV3, property abstraction
(EV5)	$\neg(a = b)$	EV2, EV4, Leibniz’s Law
(EV5')	$\Delta\neg(a = b)$	EV5, Introduction of determinacy
(EV5″)	$\neg\Delta\neg(a = b)$	EV1, Equivalence principle
(EV6)	$\neg\nabla(a = b)$	EV1, EV5′, EV5″, RAA

The addition of the last steps explains how the negation of his *reductio* assumption is derived, but it has generated concerns against Evans. In particular, the deduction of (EV5′), which relies solely on the ‘introduction of determinacy’ rule - $\neg(a = b) \rightarrow \Delta\neg(a = b)$, is not guaranteed.

In his (1994, 1997) thorough examinations, Copeland brilliantly observes that by contraposition and duality, $\neg(a = b) \rightarrow \Delta\neg(a = b)$ is equivalent to

(EoI) $\nabla(a = b) \rightarrow (a = b)$

which I shall call the ‘elimination of indeterminacy’ rule (EoI for short).[233] Why should we believe in this rule and its equivalence $\neg(a = b) \rightarrow \Delta\neg(a = b)$? Copeland suggests that given the framework of fuzzy logic developed by Zadeh, we shouldn’t. As mentioned, within this framework, the value of $\Delta(S)$ is 1 iff the value of S is 0 or 1, and the value of $\nabla(S)$ is 1 iff the value of S is in the range of 0 to 1:

$$\text{Value}\ (\Delta(S)) = 1 \leftrightarrow ((\text{Value}\ (S) = 0) \vee (\text{Value}\ (S) = 1))$$

$$\text{Value}\ (\nabla(S)) = 1 \leftrightarrow 0 < \text{Value}\ (S) < 1$$

[233] Copeland calls it ‘Evans’ axiom’.

The second line can be represented in the following truth table:

Table 6-g

∇(*S*)	*S*	¬*S*
1	0 ~ 1	0 ~ 1

We also know from the second equivalence principle EP2 that when ∇(*S*) takes the value of 1 so do ¬Δ(*S*) and ¬Δ(¬*S*):

Table 6-h

∇(*S*)	¬Δ(*S*)	¬Δ(¬*S*)
1	1	1

Now, consider the values of the negation of ¬Δ(*S*) and ¬Δ(¬*S*). Since Δ(*S*) negates ¬Δ(*S*), the pair have opposite integral values, that is, when ¬Δ(*S*) takes the value of 1, Δ(*S*) takes the value of 0, and vice versa. This is also the case for ¬Δ(¬*S*) and Δ(¬*S*):

Table 6-i

¬Δ(*S*)	Δ(*S*)
1	0
0	1

Table 6-j

¬Δ(¬*S*)	Δ(¬*S*)
1	0
0	1

Selectively merging columns from Table 6-g, 6-h, 6-i and 6-j, we get Table 6-k:

Table 6-k

$\nabla(S)$	S	$\neg S$	$\Delta(\neg S)$
1	0 ~ 1	0 ~ 1	0

Instantiating *S* with $a = b$, we have Table 6-l:

Table 6-l

$\nabla(a = b)$	$a = b$	$\neg(a = b)$	$\Delta\neg(a = b)$
1	0 ~ 1	0 ~ 1	0

Table 6-l shows the set of truth values for the four component propositions in EoI and IoD:

(EoI) $\nabla(a = b) \rightarrow (a = b)$

$\quad 1 \qquad 0 \sim 1$

(IoD) $\neg(a = b) \rightarrow \Delta\neg(a = b)$

$\quad 0 \sim 1 \qquad 0$

Thus, it is visible that when the antecedent in EoI takes the value of 1, the antecedent in IoD would take the value of 0 ~ 1, and the consequence in IoD would take the value of 0. *Modus ponens* with a zero-valued consequent and a higher-than-zero-valued antecedent does not go through.[234] In short, when $\nabla(a = b)$ is set up as an assumption in a deduction, $\neg(a = b) \rightarrow \Delta\neg(a = b)$ cannot be used as a rule of inference in subsequent lines. The deduction of EV5' is blocked due to this particular reason:

(EV1) $\nabla(a = b)$ Assumed for *reductio*

…

(EV5) $\neg(a = b)$

(EV5') $\Delta\neg(a = b)$ EV5, Introduction of determinacy

[234] Copeland (1994), pp. 85-7.

Evans' proof, as a whole, is therefore invalid. However, Tye (2000) offers a rebuttal to this reply, claiming that the deduction of EV5 seems to fall prey to the same logical problem:

> For, as just noted, within fuzzy logic, where $\nabla(a = b)$ takes the value 1, the value of $\neg(a = b)$ is less than 1. But, given the very strong, intuitive plausibility of the reasoning from (1) and (3) to (5), once it is properly elucidated, the natural conclusion to draw is that any fuzzy logic that classifies the reasoning as invalid is unsatisfactory.[235]

Here, Tye correctly spotted that EoI might also fall short of being a *modus tollens* given the fuzzy framework we described, since its consequence would have a lower value when its antecedent has value 1. Copeland also affirms this by saying:

> The difficulty is that in the present setting the inference from (5) to (5′) is invalid, since the derived formula has a lower value than the formula from which it is derived (for under the assumption that [value] (∇ (a = b)) = 1, [value] (a $\neq$ b) is non-integral and so [value] (Δ (a $\neq$ b)) = 0).[236]

In consequence, the deduction of EV5 is in trouble, according to the fuzzy logic reply. But why does refuting the deduction of EV5 go against 'strong intuition'? In fact for reasons unrelated to this fuzzy logic reply I really do think EV5 is invalidly infered in Evans' proof. I have no idea what 'strong, intuitive plausibility of the reasoning from (1) and (3) to (5)′ Tye has in mind that he thinks is the unquestionable justification for EV1 – EV5. Maybe it is the 'determinacy of self-identity' axiom, or Leibniz's Law, or both. Nevertheless, if I can show that the deduction of EV5 is flawed, Tye's rebuttal collapses. Lastly, Tye is absolutely right about one thing – some elements of the reasoning from EV1 to EV5 do need to be 'properly elucidated'. Once this is done, I will show that the 'natural conclusion', contrary to Tye's suggestion, is to 'classify the reasoning as invalid'.

6. 3. 4. Inapplicability of Leibniz's Law

This brings us to the next and more compelling reply to Evans' proof. As hinted above, this reply focuses on the deduction of EV5:

(EV2) $\lambda x(\nabla(x = a))b$

(EV4) $\neg\lambda x(\nabla(x = a))a$

[235] Tye, (2000), p. 205.
[236] Copeland, (1994), p. 87.

(EV5) $\neg(a = b)$ *EV2, EV4, Leibniz's Law*

which supposedly relies on Leibniz's Law:

(LL) $(\forall x)(\forall y)(\forall F)((\mathrm{F}x \leftrightarrow \mathrm{F}y) \leftrightarrow (x = y))$[237]

The Evansian reasoning is as follows: *b* has a property that *a* lacks, namely, the property of being indeterminately identical to *a*. So *a* and *b* are discernible and therefore not identical. However, is Leibniz's Law really the rule of inference here? It does not take long to realise the answer is straightforwardly no. The underpinning rule of inference here is in fact the contrapositive of Leibniz's Law:

(LLcp) $(\forall x)(\forall y)(\forall F)(\neg(Fx \leftrightarrow Fy) \rightarrow (x \neq y))$

The deducibility of EV5 would then be questioned by anyone who is sceptical of LLcp.[238] To discharge this unnecessary worry, Evans could replace Leibniz's Law (or its contrapositive version) with another principle. What would suffice to derive EV5 from EV2 and EV4 then? In his (1990), Noonan suggests the role can be filled by what he calls the Principle of the Diversity of the Definite Dissimilar (DDD):

(DDD) $(\forall x)(\forall y)(\Delta Fx \,\&\, \Delta Fy) \rightarrow ((Fx \,\&\, \neg Fy) \rightarrow (x \neq y))$[239]

In English:

(DDD) For all *x* and all *y*, if *x* determinately has a property that *y* determinately lacks, then *x* is not identical to *y*.

Employing DDD rather than LL in Evans' proof, we can see prima facie why DDD suffices to generate EV5 from EV2 and EV4:

(EV2) $\lambda x(\nabla(x = a))b$

(EV4) $\neg\lambda x(\nabla(x = a))a$

(EV5) $\neg(a = b)$ EV2, EV4*, DDD*

[237] Here, I take Leibniz's Law to include both the identity of indiscernibles and the indiscernibility of identicals. Common criticisms against the plausibility of the latter principle are not relevant to the present discussion.

[238] Parsons (2000) maintains such a view. He argues that (LLcp) only holds for the predication of genuine properties and $\lambda x(\nabla(x = a)$ is not a genuine property, and concludes that (EV5) crumbles as a result. The argument is also elaborated upon in Parsons & Woodruff (1995).

[239] Noonan (1990), p. 160. I have added two sets of brackets to Noonan's original formula.

The *dissimilar* property in DDD, F, is instantiated by $\lambda x(\nabla(x = a)$ – the property of being indeterminately identical to a, and thereby rendering a and b to be non-identical. This seems like good news for Evans and the foes of indeterminate identity, since Evans' proof is immune to the attack on Leibniz's Law when we reformulate it with DDD. On the other side, the new question lurking in the minds of friends of indeterminate identities is the applicability of DDD. Noonan maintains that DDD is undeniable by anticipating a possible objection that appeals to relative identity and then rebuffs its relevance to Evans' proof:

> [H]ow can (DDD) (or (LLnv)[240]) be regarded as objectionable? For all it says is that it is a sufficient condition of the distinctness of objects a and b that they be *definitely dissimilar*, and what can be the objection to that? … [W]e can see a last remaining bolthole for the proponent of indefinite identity. For he may say that indefinite identity is a kind of *relative identity* – a relation which ensures the indiscernibility of terms in some, but not in all, respects, and in particular, not in respects expressible only by predicates containing '∇', or synonyms of such predicates… But neither Evans's original argument, nor the slightly more detailed version above, demonstrates any such thing.[241]

I am in full agreement with Noonan here. No doubt, the talk of relative identity is entirely irrelevant to the present discussion. To equivocate indeterminate identity with relative identity is an injudicious categorical mistake. Despite this, Noonan fails to detect that DDD's role in Evans' proof is doubtful in a way that no appeal is made to relative identity.

The pitfall, according to Copeland (1997, 2000), is to do with higher-order indeterminacy.[242] Up to this point, it has been assumed that an indeterminate statement is determinately indeterminate:

$$\nabla(S) \rightarrow \Delta(\nabla(S))$$

When considered, higher-order indeterminacy allows the possibility of an indeterminate statement being indeterminately indeterminate:[243]

$$\nabla(S) \rightarrow \nabla(\nabla(S))$$

[240] (LLnv), according to Noonan, refers to the equivalent of (DDD), which was formulated and rejected by Johnsen (1989).

[241] Noonan, (1990), 160-1.

[242] Copeland (1997), pp. 533-4 and (2000), pp. 16-7.

[243] The consideration of higher-order indeterminacy also breaks the parity between modal operators in S5 and delta-operators – Evans' assumption upon which the 'introduction of determinacy' is based. For $\Diamond(S)$ entails $\Box\Diamond(S)$ in S5 but $\nabla(S)$ does not simply entail $\Delta(\nabla(S))$.

This in turn tarnishes the deduction of EV5 via DDD, because given the plausibility of $\nabla(a = b) \rightarrow \lambda x(\nabla(x = a))b$, both antecedent and consequence of the conditional might still be indeterminate themselves, which makes it uncertain whether '*x determinately* has a property that *y* determinately lacks' – the sufficient condition specified in DDD.

To remedy this, Copeland advocates a modified version of DDD, which he calls the principle of the Definite Diversity of the Definite Dissimilar (DDDD):

(DDDD) $(\forall x)(\forall y)(\Delta Fx \,\&\, \Delta Fy) \rightarrow ((Fx \,\&\, \neg Fy) \rightarrow ((x \neq y) \,\&\, \Delta(x \neq y)))$[244]

Having DDDD in Evans's proof produces two stunning results, from an Evansian viewpoint. Firstly, EV5 and its puzzling derivation are no longer needed in the first place, because DDDD enables EV2 and EV4 to directly deduce EV5'. Secondly, the annoying 'introduction of determinacy' rule that EV5' is originally derived upon is bypassed:

(EV1)	$\nabla(a = b)$	Assumed for *reductio*
(EV2)	$\lambda x(\nabla(x = a))b$	EV1, property abstraction
(EV3)	$\neg\nabla(a = a)$	Axiom
(EV4)	$\neg\lambda x(\nabla(x = a))a$	EV3, property abstraction
(EV5')	$\Delta\neg(a = b)$	EV2, EV4, DDDD
(EV5″)	$\neg\Delta\neg(\mathrm{a} = \mathrm{b})$	EV1, Equivalence principle
(EV6)	$\neg\nabla(a = b)$	EV1, EV5′, EV5″, RAA

From what I have discussed so far it is clear that employing DDDD delivers major improvement to Evans' proof, but even this DDDD-refined version cannot save the proof from two ruinous obstacles.

6. 3. 5. Indeterminate identity *qua* property

The second-to-last reply to Evans' proof that I am going to introduce is due to Terence Parsons and Peter Woodruff (1995)[245] and explained in further detail in Parsons (2000).[246] Their objection unfolds by focusing on the two property abstraction steps in the proof:

[244] Copeland (2000), p.16. In line with the presentation of DDD I choose to write DDDD as a conditional, which differs from Copeland's turnstile presentation:
$\neg\varnothing(\mathrm{a}), \varnothing(\mathrm{b}), \Delta(\varnothing(\mathrm{a})), \Delta(\varnothing(\mathrm{b})) \vdash \mathrm{a}\neq\mathrm{b} \,\&\, \Delta(\mathrm{a}\neq\mathrm{b})$.

(EV1) $\nabla(a = b)$ Assumed for *reductio*

(EV2) $\lambda x(\nabla(x = a))b$ EV1, property abstraction

(EV3) $\neg\nabla(a = a)$ Axiom

(EV4) $\neg\lambda x(\nabla(x = a))a$ EV3, property abstraction

Parsons and Woodruff argue that the two inferences are fallacious because ‘being indeterminately identical to *a*’ is not a genuine property, so $\lambda x(\nabla(x = a))b$ does not denote a property of *b* (as in EV2), and $\lambda x(\nabla(x = a))a$ does not denote a property that *a* lacks (as in EV4). It follows that Evans cannot appeal to a violation of Leibniz’s Law or its variants[247] in subsequent lines of the proof.

Why should this be true? Parsons and Woodruff attempt to prove this with a *reductio* argument of their own:

Suppose *a* and *b* are indeterminately identical:

(PW1) $\nabla(a = b)$ Assumption

From the indiscernibility of identicals, $a = b$ is interchangeable with $\forall F(Fa \leftrightarrow Fb)$:

(PW2) $\nabla(\forall F(Fa \leftrightarrow Fb))$ PW1, Leibniz’s Law

Now assume for *reductio* that ‘being indeterminately identical to *a*’ is a property of *b*, then the following λ-abstract stands for a property. Here, instead of abstracting this property from PW1, which yields $\lambda x(\nabla(\mathrm{x} = a))b$, we instead do it with PW2:

(PW3) ‘Being indeterminately identical to *a*’ is a property of *b.*

(PW4) $\lambda x(\nabla(\forall F(Fx \leftrightarrow Fa))b$ PW3, λ-abstraction

Next is the insertion of the axiom that Evans himself asserted – that self-identity is not indeterminate:

(PW5) $\neg\nabla(a = a)$ Axiom

Again, this can be written as $\neg\nabla(\forall F(Fa \leftrightarrow Fa))$ due to Leibniz’s Law:

(PW6) $\neg\nabla(\forall F(Fa \leftrightarrow Fa))$ PW5, Leibniz’s Law

By λ-abstraction, PW6 can be stated as:

[245] Parsons & Woodruff (1995), pp. 175-8.

[246] Parsons (2000), pp. 50-5.

[247] These variants include DDD or DDDD or LL_{cp} as described earlier.

(PW7) $\neg\, \lambda x \nabla(\forall F(Fx \leftrightarrow Fa))a$ PW6, λ-abstraction

Conjoining PW4 and PW7 yields:

(PW8) $\lambda x \nabla(\forall F(Fx \leftrightarrow Fa))b \;\&\; \neg\, \lambda x \nabla(\forall F(Fx \leftrightarrow Fa))a$ PW4, PW7, &I

which by existential introduction gives:

(PW9) $\exists F(\neg Fa \;\&\; Fb)$ PW8, ∃I

Since '*the truth of this sentence* [PW9] *contradicts the truth of* [$\nabla(\forall F(Fa \leftrightarrow Fb))$, i.e. PW2]'[248], the *reductio* assumption is proved to be false, giving us PW10:

(PW10) 'Being indeterminately identical to *a*' is not a property of *b*. 2, 9, RAA

Parsons and Woodruff claim that the failure of Evans' reasoning echoes a familiar logical mistake – Russell's paradox. In the Evansian proof, indeterminate identity is analysed in terms of a property, and this property is in turn addressed by a lambda abstract ($\lambda x(\nabla(x = a))b$) which carries that property in its bracket. As highlighted by the PW-proof, it would lead to a problematic result:

> If the language is sufficiently rich then we cannot assume that *any* such abstract refers to a property whose application to objects is perfectly characterized in the usual way by the lambda abstraction, the principle that '$\Phi(a)$' is interchangable with '$\lambda x[\Phi(x)](a)$' in all extensional contexts. Such a powerful assumption leads to paradoxes (like the Russell paradox) whenever the language is sufficiently rich. Such constraints are well-known, and people are used to restricting either the abstraction axiom or quantification over properties to avoid such paradoxes.[249]

Parsons and Woodruff's message is clear: we must restrict the use of lambda abstraction when referring to properties. Some abstraction axioms, like Russell's axiom schema cannot *at the same time* (i) satisfy the principle of abstraction *and* (ii) stand for properties. Evans' argument against indeterminate identity no doubt relies on the assumption that $\lambda x(\nabla(x = a))b$ meets both (i) and (ii), because if $\lambda x(\nabla(x = a))b$ only satisfies (i) but fails (ii), then Evans' proof collapses at EV5.[250] On the other hand, if it satisfies (ii) but fails (i), then the proof collapses at the first two property abstraction steps (EV1 to EV2 and

[248] Parsons (2000), p. 51.

[249] Parsons & Woodruff (1995), p. 175. Authors' italics.

[250] Hence, if the λ-abstract does not stand for a property, then the application of Leibniz's Law fails.

EV3 to EV4).[251] According to Parsons and Woodruff, Evans is here neglecting to bear in mind a major lesson from Russell's paradox. Nevertheless, their critique also concedes that it might turn out to be the case that $\lambda x(\nabla(x = a))b$ does indeed fully satisfy both (i) and (ii). But before someone makes the case for it, no one should construct a proof by simply assuming it, as Evans did.[252]

The first way by which Evans could respond to Parsons and Woodruff's objection is to question the relevance of their point. Suppose there really are no such properties as the property of 'being indeterminately identical to *a*'. So what? Does it automatically imply that no violation of Leibniz's Law or its variants occurred? Convention has it that Leibniz's Law is used with regards to properties, but this need not always be the case. Someone like Quine, who has trouble accepting the idea of properties in the first place, can use Leibniz's Law with regards to predicates only. Then, as long as Parsons and Woodruff admit (which they do) that 'being indeterminately identical to *a*' is a predicate, their criticism would inflict no damage on Evans' proof.

In addition, I think the Parsons-Woodruff objection has two more subtle defects. Firstly, the major force behind their argument comes from the last step where they claim that '*the truth of this sentence* [PW9] *contradicts the truth of* [PW2]'[253]:

(PW2) $\nabla(\forall F(Fa \leftrightarrow Fb))$

(PW9) $\exists F(\neg Fa \,\&\, Fb)$

It is easy to see that the pair would be straightforwardly contradictory if PW2 didn't begin with an indeterminacy operator '∇'. But since '∇' is asserted (by Parsons and Woodruff themselves), I cannot see why their claim is true. By the equivalence principle, $\nabla(S)$ is true iff $\neg\Delta(S)$ is true. So PW2 can be read as saying '$\forall F(Fa \leftrightarrow Fb)$ is not determinately true', which *appears* to be consistent with $\exists F(\neg Fa \,\&\, Fb)$. I am not going to argue here that the pair *is* consistent. Whether that is the case is going to be determined by the choice of logical system (e.g. Zadeh's fuzzy valuation etc.) that we map the delta operators onto. Nonetheless, the *prima facie* result (that the pair is consistent) is enough to bring down Parsons and Woodruff's conclusion, because their conclusion is yielded under the assumption that the pair *must not* have consistent truth values. This additional assumption

[251] Hence, if the principle of abstraction does not hold for the property of 'being indeterminately identical to *x*', then EV2 and EV4 are incorrect inferences.
[252] Parsons & Woodruff (1995), pp. 175-6.
[253] Parsons (2000), p. 51.

must be proved to be true, and in doing so they must notice the impact from "∇", but they addressed none of these.

The second worry is about the transition from PW8 to PW9:

(PW8) $\lambda x\nabla(\forall F(Fx \leftrightarrow Fa))b \;\&\; \neg\lambda x\nabla(\forall F(Fx \leftrightarrow Fa))a$

(PW9) $\exists F(\neg Fa \;\&\; Fb)$ PW8, ∃I

where Parsons and Woodruff treat the two lambda abstracts in $\lambda x\nabla(\forall F(Fx \leftrightarrow Fa))b$ and $\neg\lambda x\nabla(\forall F(Fx \leftrightarrow Fa))a$ as having the form of $\neg Fa \;\&\; Fb$. This is a result of assuming that two syntactically identical lambda expressions must have the same meaning. I will explain shortly that this assumption is mistaken – you can give two syntactically identical lambda expressions different meanings. It follows that PW8 could be in the form of $\neg Fa \;\&\; Gb$, which ultimately invalidates the argument against Evans. Luckily, this error in Parsons' and Woodruff's reasoning sheds light on a nearly identical error in the reasoning of Evans. This brings us to the final reply to Evans.

6. 3. 6. Dissimilar property or disparate properties?

Lastly, supposing 'being indeterminately identical to *a*' is a genuine property of *b* that can be abstracted from $\nabla(a = b)$ (i.e. suppose Parsons and Woodruff's reply fails), is there another way of rendering the DDDD-refined Evans' proof invalid? Yes, there is one final, knockout blow delivered by Copeland (2000).[254] The fallacy, as Copeland sees it, is rooted in the transition from EV2 and EV4 to EV5':

(EV1) $\nabla(a = b)$ Assumed for *reductio*

(EV2) $\lambda x(\nabla(x = a))b$ EV1, property abstraction

(EV3) $\neg\nabla(a = a)$ Axiom

(EV4) $\neg\lambda x(\nabla(x = a))a$ EV3, property abstraction

(EV5') $\Delta\neg(a = b)$ EV2, EV4, DDDD

Together, EV2 and EV4 voice the message that *b* has a property that *a* lacks. Hence, there is *one dissimilar property* responsible for concluding $a \neq b$ˆ, via DDDD. But is this really

[254] Copeland (2000), pp, 17-22.

the case? What if the property *b* possesses, as expressed in EV2, is an entirely different property from the one that *a* doesn't possess, as expressed in EV4? In other words, what if instead of *one dissimilar property* EV2 and EV4 are concerned with *two disparate properties*? In symbolisation, we have so far presumed that EV2 and EV4 express propositions of the following form:

(EV2) $\mathrm{F}b$

(EV4) $\neg\mathrm{F}a$

Let us call this the *One Dissimilar Property*-Reading (ODPR) of EV2 and EV4.

In contrast, the two premises might express:

(EV2) $\mathrm{F}b$

(EV4) $\mathrm{G}a$

Call this the *Two Disparate Properties*-Reading (TDPR).

Under ODPR, the two premises deduce EV5′ via DDDD:

(EV2) $\mathrm{F}b$

(EV4) $\neg\mathrm{F}a$

(DDDD) $(\forall x)(\forall y)(\Delta Fx \,\&\, \Delta Fy) \rightarrow ((Fx \,\&\, \neg Fy) \rightarrow ((x \neq y) \,\&\, \Delta(x \neq y)))$

(EV5') $\Delta\neg(a = b)$

But this is not the case under TDPR. Hence, the following deduction does not go through:

(EV2) $\mathrm{F}b$

(EV4) $\mathrm{G}a$

(DDDD) $(\forall x)(\forall y)(\Delta Fx \,\&\, \Delta Fy) \rightarrow ((Fx \,\&\, \neg Fy) \rightarrow ((x \neq y) \,\&\, \Delta(x \neq y)))$

(EV5') $\Delta\neg(a = b)$

The message is clear: if we can find a way to justify TDPR, then Evans' proof is undermined, for the joint truth of EV2 and EV4 won't entail that *b* has a property that *a* lacks. But how is this possible, the foes of indeterminate identities may ask. Aren't $\lambda x(\nabla(x = a))b$ and $\neg\lambda x(\nabla(x = a))a$ straightforwardly of the form of $\mathrm{F}b$ and $\neg\mathrm{F}a$ since they have the same property '$\lambda x(\nabla(x = a))$' stated in them?

In reply, the friends of indeterminate identity would answer that they are not of the form of Fb and ¬Fa, because the two expressions '$\lambda x(\nabla(x = a))$' in the two formulae have different meanings in EV2 and EV4. In EV2, '$\lambda x(\nabla(x = a))$' means 'the property of being indeterminately identical to a', and as entailed by the *reductio* assumption EV1, b has this property – i.e. $\lambda x(\nabla(x = a))b$. On the other hand, EV4 is deduced from EV3 ($\neg\nabla(a = a)$), the indeterminacy of self-identity axiom. As a result, '$\lambda x(\nabla(x = a))$', as it stands in EV4, should be understood as denoting 'the property of being indeterminately self-identical'. This point can be made clearer by rewriting EV3 as

(EV3) $(\forall x)\neg\nabla(x = x)$

Reformulating this way it gives us a clear indication that '$\lambda x\nabla(x = a)$' in EV4 is the same as '$\lambda x\nabla(x = x)$', namely, 'the property of being indeterminately self-identical', which of course, is not attributable to a. But this property is also not attributable to b! Nothing is indeterminately self-identical to itself! Thus, on this view, to hold that both EV2 and EV4 are true is like saying 'b has the property of "being indeterminately identical to a", and a doesn't have the property of "being indeterminately identical to itself", and b doesn't have the property of "being indeterminately identical to itself". We might symbolise it as:

$$\mathrm{F}b \,\&\, \neg\mathrm{G}a \,\&\, \neg\mathrm{G}b$$

Here, F represents the property $\lambda x(\nabla(x = a))$, and G represents the property $\lambda x\nabla(x = x)$. These two properties are the two different extensions of the formula '$\lambda x(\nabla(x = a))$' expressed in EV2 and EV4. Having explained this distinction, we can conclude that Evans falls short of showing a dissimilar property that DDDD applies to. In other words, it hasn't shown that there is a property b has and a lacks.

This error is fatal for Evans' proof, but subtle. As mentioned in the last section, not only Evans but also some of his opponents have committed this mistake. Recall Parsons' and Woodruff's reply to Evans, in which they aim to deduce $\exists F(\neg Fa \,\&\, Fb)$ from $\lambda x\nabla(\forall F(Fx \leftrightarrow Fa))b \,\&\, \neg\lambda x\nabla(\forall F(Fx \leftrightarrow Fa))a.$ The existential introduction move does not go through once it is recognised that the two lambda expressions have disparate references – the property of 'being indeterminately identical to a' in the left-hand conjunct, and the property of 'being indeterminately identical to itself' in the right-hand conjunct.

It is not hard to understand that the subtlety of this confusion over dissimilar property and disparate properties results from an obvious fact – that *a*'s being self-identical **is** *a*'s being identical to *a*. Hence, the two lambda expressions are co-extensional in *a*. Unfortunately, co-extensionality is a glaring phenomenon in philosophy. Sometimes it might overshadow and keep out of sight the crucial fact that the two expressions do not mean the same thing. The root of this problem can be traced back to what Copeland calls the *'neo-Aristotelian thesis concerning predication'*[255]. Copeland discovered from several commentaries of *Metaphysics*[256] that Aristotle discussed a view about property identity: in Socrates, 'being an animal' *is* 'being a man', yet of course, the two properties do not share the same meaning. What the great ancient Greek philosopher anticipated is what both Evans and Parsons and Woodruff seem to overlook: for *a*, the two lambdas have different meanings despite the fact that 'being self-identical' *is* 'being identical to *a*' in *a*.

In conclusion, I have examined in this section two minor (Thomason's and Broome's) and major (Copeland's, Noonan's, and Parsons' and Woodruff's) objections to Evans' attempted proof against indeterminate identity. To sum up, Evans' argument fails to be a valid one for the reason specified by Copeland. Once we acknowledge that EV2 and EV4 are in the form of $\neg Fa$ & Gb, there is no sufficient rule of inference in the subsequent lines of Evans' proof to derive the non-identity of *a* and *b* from EV2 and EV4. And once we acknowledge the failure of Evans' argument, we can finally erase an uncertainty about V2 of my argument against Putnam. The following Chapter will continue to consolidate the plausibility of V2 by presenting a proper proof of the 'vagueness to indeterminacy' theorem.

[255] Copeland (2000), p. 17.

[256] These are Anscombe (1961), Cresswell (1987), and Kirwan (1993).

Chapter 7. Towards indeterminate mind-brain identity (b) - proving the 'vagueness to indeterminacy' theorem

7. 0. Outline of Chapter 7

In order to establish the plausibility of the second premise of my argument against Putnam, the 'weakened vagueness to indeterminacy theorem' (VTIW)—*If one flanking term of an identity statement is vague, and the other flanking term is crisp, then the identity statement is indeterminate*—I set myself two tasks: to construct a negative argument and a positive argument:

Negative argument - say what is wrong with Evans' argument for the general absurdity of indeterminate identities.

Positive argument - explain why and how vagueness of term(s) leads to the indeterminate identity that holds between these terms.

The first task is accomplished in 6.3, where I surveyed a number of ways to question Evans' argument. This Chapter is about the second task – proving VTIW. Before I give the proof, there is a minor question that needs to be answered. One possible objection to my indeterminacy move is to claim that statements containing a vague term may nevertheless be determinate, and that Putnam's multiple realisability objection can be driven through with respect to those central cases, irrespective of the arguments concerning fringe cases where there is indeterminacy.[257]

I target our actual qualia-concepts, which, as I present, are vague. These concepts are not reformed mental concepts expressed in some precise language, rather like the surveyor's 'Mango/Nectarine Everest'—if this can even be done for qualia. The principal response is that, since only identity statements are at issue, it suffices to prove the weakened 'vagueness to indeterminacy' theorem (VTIW) in order to defeat the objection.

[257] I am grateful to Michael-John Turp for raising this objection.

7. 1. *Principle of difference* and the first proof

To begin with, let us recall the theorem:

(VTIW) $\mathrm{V}(x), \Lambda(y) \vdash \nabla(x = y)$

We might try the following *reductio* proof:

Proof VTIW-(1)

$\mathrm{V}x, \Lambda y \vdash \nabla(x = y)$

1. $\mathrm{V}x$	Assumption
2. Λy	Assumption
3. $\Delta(x = y)$	Assumed for *reductio*
4. $\forall F(Fx \leftrightarrow Fy)$	3, Leibniz's Law
5. $(\mathrm{V}x \mathbin{\&} \Lambda y) \rightarrow (\nabla Fx \mathbin{\&} (Fy \mathbin{\&} \Delta Fy))$	*Principle of difference*
6. $\mathrm{V}x \mathbin{\&} \Lambda y$	1, 2, &I
7. $\nabla F\mathrm{x} \mathbin{\&} (F\mathrm{y} \mathbin{\&} \Delta Fy)$	5, 6, →E
8. ∇Fa	7, &E
9. $\lambda u(\nabla Fu)x$	8, λ-abstraction
10. ΔFy	7, &E
11. $\lambda v(\Delta Fv)y$	10, λ-abstraction
12. $\lambda v(\nabla Fv)y$	4, 9, ↔E
13. $\neg\Delta(x = y)$	3, 11, 12, RAA
14. $\nabla(x = y)$	13, equivalence principle

The reasoning goes: let '*x*' be any vague term and let '*y*' be any crisp term. Then x=y is indeterminate. Since the proof is by *reductio*, it is assumed that $x = y$ is determinate and a contradiction is derived. From this *reductio* assumption, it follows by Leibniz's Law that *y* has all *x*'s properties and vice versa (and moreover determinately so). Call this the 'Leibniz consequence'. The proof now appeals to the following *principle of difference* (which holds for any pair of terms): if '*x*' is vague and '*y*' is crisp, then there is some predicate *F* such that *Fx* is indeterminate while *Fy* is determinate. Since *Fx* is indeterminate, it follows by property abstraction that *x* has the property of being indeterminately *F*. Taking this in conjunction with the Leibniz consequence yields: *y* also has the property of being indeterminately *F*, contradicting the statement that *Fy* is determinate. Therefore $x = y$ is indeterminate.

It is clear that *Proof VTIW-(1)* relies heavily on *principle of difference* (PoD for short) which can be formally put as follows:

(PoD) $\mathrm{V}x \,\&\, \Lambda y \rightarrow \exists F(\nabla Fx \,\&\, ((Fy \,\&\, \Delta Fy) \vee (\neg Fy \,\&\, \Delta\neg Fy)))$[258]

7. 1. 1. The 'solid' delta

Before I proceed to explain this principle, allow me to clarify an important notational issue. So far, we have been following Evans' original way of using the determinacy operator 'Δ' – under which Δ(*S*) means it is determinate whether *S*. The delta symbol stays silent on whether it is determinately *true* that *S* or determinately *false* that *S*. It merely indicates that *S* is determinate. In other words, Δ(*S*) allows both possibilities in which *S* is determinately true or determinately false. For this reason, 'Δ(*S*)' does not suffice to notate '*S* is determinately true'.[259] In order to preserve Evans' usage of 'Δ', we must write '*S* & Δ(*S*)' for '*S* is determinately true' and '¬*S* & Δ¬(*S*)' for '*S* is determinately false'. This is fine, apart from the small disadvantage that sometimes the notation can be quite long as exemplified by the formalisation of PoD above. For convenience, I think it will be handy to have a new symbol that offers us an easier way to say 'it is determinately *true* that *S*'. To this end I introduce the solid delta:

$$\blacktriangle$$

Whereas 'Δ(*S*)' denotes '*S* is indeterminate', let '▲(*S*)' mean '*S* is determinately true', so the following hold:

(EP6) $\blacktriangle(S) \dashv\vdash S \,\&\, \Delta(S)$

(EP7) $\blacktriangle\neg(S) \dashv\vdash \neg S \,\&\, \Delta\neg(S)$

from which we can derive that *S* is determinate when *S* is either determinately true or determinately false:

[258] Here, *x* and *y* are treated as free variables which are presumed to be universally quantified.

[259] For the same reason, 'Δ(*S*)' does not suffice to notate '*S* is determinately false'.

(EP8) $\Delta(S) \dashv\vdash \blacktriangle(S) \vee \blacktriangle\neg(S)$

From this, we can easily prove that the equivalence principles that govern the relations between hollow delta 'Δ' and upside-down delta '∇' as previously mentioned (EP1 and EP2) can be extended to the solid delta.

(EP9) $\blacktriangle(S) \vee \blacktriangle\neg(S) \dashv\vdash \neg\nabla(S)$

By contraposition, *S* is indeterminate when it is not the case that *S* is either not determinately true or not determinately false:

$$\nabla(S) \dashv\vdash \neg(\blacktriangle(S) \vee \blacktriangle\neg(S))$$

By De Morgan's Law, this equates to saying *S* is indeterminate when it is not the case that *S* is determinately true *and* it is not the case that *S* is determinately false:

(EP10) $\nabla(S) \dashv\vdash \neg\blacktriangle(S) \;\&\; \neg\blacktriangle\neg(S))$

These equivalence principles of solid delta will be extremely helpful to our proof of the vagueness to indeterminacy theorem. To begin with, using EP6 and EP7 we can shorten PoD without altering its meaning :

(PoD) $\mathrm{V}x \;\&\; \Lambda y \rightarrow \exists F(\nabla Fx \;\&\; (\blacktriangle Fy \vee \blacktriangle\neg Fy))$[260]

The principle states: for any pair of terms, if one term is vague and the other is crisp, then there must be a property that the vague term possesses indeterminately and the crisp term possesses determinately. The difference at the level of determinacy status of terms must manifest at the level of determinacy status of statements.

Why is PoD true? Let us first focus on the second part of the principle. That is, the consequent of the conditional, it amounts to the following:

(DDSC) $\exists F(\nabla \mathrm{F}x \;\&\; (\blacktriangle Fy \vee \blacktriangle\neg Fy))$

[260] By the definition of solid delta and the equivalence principles, we can shorten the principle even further to $\mathrm{Vx} \;\&\; \Lambda\mathrm{y} \rightarrow \exists\mathrm{F}(\nabla\mathrm{Fx} \;\&\; \Delta\mathrm{Fy})$, due to EP8: $\blacktriangle(S) \vee \blacktriangle\neg(S) \leftrightarrow \Delta(S)$.

This says that there are some *F* such that for all pair of terms *x* and *y*, *x* has *F* indeterminately and *y* either has *F* determinately or does not have *F* determinately. Call this the *difference in determinacy status consequent* (DDSC). Now consider its negation:

(NDDSC) $$\neg\exists F(\nabla Fx \,\&\, (\blacktriangle Fy \vee \blacktriangle\neg Fy))$$

which I am going to call the *no difference in determinacy status consequent* (NDDSC). If NDDSC is never true, then DDSC is never false, and if DDSC is never false, then PoD is always true. The question at hand thereby becomes why NDDSC is never. I suggest the following argument:

Suppose *x* is vague and *y* is crisp, and let us also suppose for *reductio* purpose that there isn't an *F* such that *x* has it indeterminately and *y* either determinately has it or determinately doesn't have it. By the definition of solid delta and the equivalence principles, the *reductio* assumption amounts to saying that for all *F* if *x* has *F* indeterminately then *y* also has *F* indeterminately. The next move is to find an instantiation of such *F*. A rather tricky instantiation of *F* could be *the property of being indeterminately identical to y.*[261] This could be a property of *x*, since *x* might be indeterminately identical to *y*, but it never will be a property of y because *y* is not indeterminately identical to *y*. After all, Evans is right about one thing – self-identity is always determinate! Thus, the *reductio* assumption entails a consequence that contradicts the axiom of determinacy of self-identity, and therefore must be rejected. Therefore, it follows that NDDSC must be false and DDSC must be true, which in turn shows that PoD is never false.

Furthermore, the fatal flaw of NDDSC can be made more clearly without appealing to the self-identity axiom. Admitting NDDSC is to concede $\forall F(\nabla Fx \rightarrow \nabla Fy))$ which says that for any pair of terms, if one possesses a property indeterminately then the other one must also have that very same property indeterminately. The reason for this being false is simple and straightforward. Recall the Mt. Everest example mentioned earlier. The ordinary 'Mt. Everest' might have a property of including a particular rock, say rock 1, indeterminately. If NDDSC were true, it would follow that this *property of indeterminately including rock 1* is shared by *any* other term, such as 'Jelly firings'. 'Jelly

[261] In 6.3.5, I discussed this kind of property in details since it instigates one of the objections to Evans' argument against indeterminate identity.

firings' has the property of indeterminately including a rock in Nepal? What on Earth does this even mean?

Although NDDSC is clearly false, and by the same token PoD is never false, this merely shows that difference in the determinacy status of statements must manifest somewhere, and it hasn't shown why it is an implication of the difference in the determinacy status of the terms. In other words, I have merely shown that the consequent in PoD is always true and the problem is every conditional with DDSC as its consequent would be a true conditional. Any antecedents imply the fact that $\exists F(\nabla Fx \,\&\, (\blacktriangle Fy \vee \blacktriangle \neg Fy))$ – for any pair of terms there is a property such that one has it or lacks it determinately and the other has it indeterminately. In short, $\exists F(\nabla Fx \,\&\, (\blacktriangle Fy \vee \blacktriangle \neg Fy))$ is trivially true, which in turn leads to the trivial truth of PoD. The fact that PoD is trivially true might come as a surprise for some people, but I think they would find the principle intuitive once they had thought about it thoroughly. When you have two terms with different determinacy statuses, i.e. one is vague and the other is crisp, is it more intuitive to say that the difference does manifest in the statements or predications of these terms or that it doesn't? As I showed above, the latter view amounts to accepting that any properties can be indeterminately possessed by any pair of terms, the counter-intuitiveness of which should be obvious to everyone.

Given the plausibility of PoD, it seems that *Proof VTIW-1* goes through unproblematically, thereby providing us justification for VTIW. Unfortunately, there is a drawback in this vagueness to indeterminacy theorem that needs to be addressed. As it currently stands the theorem reads: $\mathrm{V}x, \Lambda y \vdash \nabla(x = y)$. However, a vague x and a crisp y do not necessarily give you indeterminate identity between x and y. Let x stand for the previously introduced ordinary 'Mt. Everest', which is vague, and let y be a precise 'Mt. Cook' pegged by Surveyor Mango, which is crisp. The Asian mountain is by no means identical to the Kiwi peak, not even indeterminately so! Mt. Everest and Mt. Cook are determinately non-identical. The aforementioned example of ordinary 'Mt. Everest' (vague) and 'Jelly firings' (crisp) illuminates this point even more plainly. There are pairs of vague and crisp terms that are just unrelated and not identical, determinately. In a nutshell, the following also holds for a vague x and a crisp y:

$$\mathrm{V}x, \Lambda y \vdash \blacktriangle(x \neq y)$$

In other words, a vague term and a crisp term are either indeterminately identical or determinately not identical:

$$\mathrm{V}x, \Lambda y \vdash \nabla(x{=}y), \blacktriangle(x \neq y)$$

VTIW overlooks this important disjunction and is therefore incorrect. This means there must be some serious problems with its proof that went unnoticed previously. The problems lie in line 12 – the application of the so-called '*Leibniz's Consequence*' (line 4) in the proof:

Proof. VTIW-(1)

…

3. $\Delta(x = y)$	Assumed for *reductio*
4. $\forall F(Fx \leftrightarrow Fy)$	3, Leibniz's Law
…	
8. ∇Fx	7, &E
9. $\lambda u(\nabla Fu)x$	8, λ-abstraction
…	
12. $\lambda v(\nabla Fv)y$	4, 9, ↔E

…

Line 4 does not present any issues. From Leibniz's Law, *x* and *y* would share the same properties given that *x* and *y* are determinately identical to each other (which is the *reductio* assumption). Line 12 cashes in on this idea and aims to deduce that *y* has an indeterminate property of being *F* since *x* has this very property. This is where the trouble kicks in. The reasoning for line 12 assumes that Leibniz's Law and its implication, namely *Leibniz's Consequence* (line 4), work for indeterminate properties as well as determinate ones. As I will explain in the next section,[262] this assumption is highly doubtful. In addition, line 12 is prone to another objection. One might argue that being indeterminately *F* is not a genuine property at all.[263] That is to say even though the lambda abstraction goes through from line 8 to line 9, it does not mean that it is a genuine property abstraction move, i.e. $\lambda x(\nabla Fx)$ is not a genuine property of *x*. It would follow that the elimination of biconditional move on line 12 does not go through.

[262] For details, see 6.3.4 'Inapplicability of Leibniz's Law'.

[263] For details, see 6.3.5 'Indeterminate identity *qua* property'.

7. 1. 2. VTIWW and its impact on Argument V

We are now aware of the fact that VTIW fails to take into account the possibility of determinate non-identity between a vague and a crisp term, and we have also admitted the blunder in *Proof VTIW-(1)*. The upshot is, we have paved the way for a further weakened but more promising 'vagueness to indeterminacy' theorem that captures the idea that a vague term and a crisp term are *either indeterminately identical or determinately not identical*:

$$Vx, \Lambda y \vdash \nabla(x=y), \blacktriangle(x\neq y)$$

By sequent calculus, this can be turned into:

(VTIWW) $$Vx, \Lambda y, \neg\blacktriangle(x\neq y) \vdash \nabla(x=y)$$

which in English reads:

(VTIWW) If one flanking term of an identity statement is vague, and the other flanking term is crisp, *and they are not determinately non-identical*, then the identity statement is indeterminate.

VTIWW is, I believe, the correct formulation of the so-called vagueness to indeterminacy theorem, and I will prove it soon. Before we look at the proof, let us go back to the big picture and revisit my argument against Putnam. VTIWW replaces VTIW and this affects my argument, in particular the second premise. To accommodate the change brought forward by VTIWW, V2 should be rewritten as:

<u>Argument V</u>

V1. 'P_h' and 'P_o' are vague terms. 'C' and 'J' are crisp terms.

V2. If one flanking term of an identity statement is vague, and the other flanking term is crisp, *and they are not determinately non-identical*, then the identity statement is indeterminate.

V3. Transitivity of identity fails for indeterminate identity.

Sub-conclusion V4. Transitivity of identity fails for identity statements 'P_h = C' and 'P_o = J'.

Conclusion V5. Putnam's argument is invalid.

How does the updated version of V2 affect Argument V? I foresee a possible criticism. Due to the extra new condition for indeterminate identity that the flanking terms must not be determinately non-identical, my opponent might ask why this condition is met. I have argued for and explained the vagueness of ‘P_h’ and crispness of ‘C’, but I have not said why the two terms are not determinately non-identical, then given VTIWW how can I conclude that ‘$P_h = C$’ is indeterminate identity? The critic might move on to accuse me of question begging. Note that Putnam’s conclusion is that pain and C-fibre firings are non-identical. My argument that targets Putnam’s validity requires a premise that says it is not the case that ‘P_h’ and ‘C’ are determinately non-identical. It surely seems that I am somehow presuming in my premise the opposite of Putnam’s conclusion. *Petitio Principii*!

Not quite. From VTIW to VTIWW, the extra condition (that indeterminate identity requires not only a pair consisting of one vague and one crisp term but also the non-determinate non-identity between the pair) added in V2 certainly makes a noticeable impact to my argument. But contrary to the critic’s view, I believe the updated V2 does not undermine my argument against Putnam. To elucidate this point, it must be emphasised that the target of my indeterminacy attack is a particular argument of Putnam’s:

Argument P

1. If identity theory is correct, then $(P_h = C)$ & $(P_o = J)$.
2. $P_h = P_o$
3. If identity theory is correct, then C = J. 1, 2, Transitivity, Symmetry
4. $C \neq J$
5. Identity theory is not correct. 3, 4, *modus tollens*

We know V2 states VTIWW and it is equivalent to:

$$Vx, \Lambda y \vdash \nabla(x=y), \blacktriangle(x\neq y)$$

Given the vagueness of ‘pain’ and the crispness of ‘C-fibre firings’, we get:

$$\nabla(P_h = C) \vee \blacktriangle(P_h \neq C)$$

So when ‘pain’ is vague and ‘C-fibre firings’ is crisp, either pain is indeterminately identical to C-fibre firings or they are determinately non-identical. The first conjunct is what I have been pushing for so far in the current Chapter. As detailed in 9.1.4, once we see that '$P_h = C$' is indeterminate identity, Putnam's argument is invalid because line 3 (of Argument P)'s inference rule collapses. Thus, if the first disjunct is taken to be the case, it is mission accomplished for me. The trouble is how to accommodate the second disjunct. In other words, what should I say if the opponent press that VTIWW is correct and 'P_h' and 'C' are respectively vague and crisp but press for $\blacktriangle P_h \neq C$? My answer somewhat resembles the mirroring strategy approach to the conceivability arguments. I contend that if they make this move then Argument P becomes redundant. VTIWW describes three conditions for indeterminate identity: i) x is vague; ii) y is crisp; iii) x and y are not determinately non-identical. Refuting iii) amounts to claiming that $\blacktriangle P_h \neq C$ is true, which resembles Putnam's original conclusion in the first place. Suppose my opponent is able to provide a sound argument for this claim. Let us call it Argument PZ. It would have the following form, where Φ can be substituted with whatever a good justification for $\blacktriangle P_h \neq C$ might be:

Argument PZ

1. Φ
2. $\Phi \rightarrow \blacktriangle P_h \neq C$

----

3. $\blacktriangle P_h \neq C$

Argument PZ would need to be independent of Argument P, since it is being used to buttress Argument P and Argument P cannot be used to buttress itself without circularity. But, like Argument P, it amounts to an argument against identity theory. For this reason, if a proponent of Argument P is in a position to defend it using Argument PZ, then she does not need Argument P in the first place. Therefore, rejecting the third condition in VTIWW renders Putnam's Argument P redundant. On the other hand, if all three conditions are upheld, then $\nabla(P_h = C)$ is derived, in which case Argument P becomes invalid.

My opponent could adopt the same strategy as in the mirroring debate and conclude that the burden of proof is on my side rather than hers. Her reasoning would probably look like this: Argument V and Argument P have the following forms[264]:

<u>Argument V</u>

1. $V(P_h)\ \&\ \Lambda(C)$
2. $Vx, \Lambda y \vdash \nabla(x=y), \blacktriangle(x \neq y)$

----

3. $\nabla(P_h = C) \vee \blacktriangle(P_h \neq C)$
4. $\neg(\blacktriangle P_h \neq C)$

----

5. $\nabla(P_h = C)$
6. Transitivity fails for ∇-statements.

----

7. Argument P is invalid.

<u>Argument P</u>

1. $P_h = P_o$
2. $C \neq J$

----

3. $\blacktriangle(P_h \neq C)$

The critic would call attention to premise 4 of Argument V and highlight the fact that it is a negation of Argument P's conclusion. She then might claim that my argument is either circular, for the reason that I'm assuming in one of my premises the negation of Putnam's conclusion, or redundant, if I happen to have an independent argument to buttress my premise 4 in the following form:

<u>Argument VZ</u>

1. Ψ
2. $\Psi \rightarrow \neg(\blacktriangle P_h \neq C)$

----

[264] Note that the conclusion of Argument P is $\blacktriangle$-modified, which differs from Putnam's original conclusion - $P_h \neq C$. I contend that this is only a syntactical difference. Since Putnam does not anticipate mind-brain identity (or non-identity) as indeterminate identity and, I argue, he cannot do so without rendering his own argument invalid, Putnam would have to agree that his conclusion is in fact $\blacktriangle(Ph \neq C)$.

3. $\neg(\blacktriangle P_h \neq C)$

Her criticism would then mirror mine: since Argument VZ is a simpler, and therefore better argument against Putnam than Argument V, Argument V is redundant. If I attempt to deny the inclusion of premise 4 in my Argument V, my opponent would continue her disproof by showing that without premise 4 I would have to retreat to a weaker position – $\nabla(P_h = C) \vee \blacktriangle(Ph \neq C)$. Since this disjunction is compatible with Putnam's conclusion - $\blacktriangle(Ph \neq C)$, my overall attack on Putnam's argument seems toothless.

At this point, it seems like we are revisiting the 'partners in crime' situation that we encountered in the mirroring campaign against Kripke and Chalmers. It looks like I'm going to concede mutual defeat for both sides' arguments. But this time is different, and I am going to say the criticism is simply misleading and incorrect. Rejecting Putnam's conclusion is never the intention of my argument. By elucidating the possibility of mind-brain identity as indeterminate identity and how transitivity fails for indeterminate identity, I argue that Putnam's argument is invalid. This position is consistent with Putnam's conclusion being true. Therefore, I do not need premise 4 in Argument V and consequently do not need Argument VZ. The critic would be right about her judgement that my overall position is a disjunctive one: $\nabla(P_h = C) \vee \blacktriangle(P_h \neq C)$. Namely, either mind-brain identity is indeterminate or Putnam's conclusion is true. She would also be justified in asserting that $\nabla(P_h = C) \vee \blacktriangle(Ph \neq C)$ is not contradictory to $\blacktriangle(P_h \neq C)$. However, this by no means indicates that my approach is innocuous to Putnam. On the contrary, my strategy is a forceful one, because I present a disjunct – a possibility of mind-brain identity being an indeterminate one – in which Putnam's argument is invalid. To render my attack innocuous is thereby to repudiate this possibility – a task that no one has performed so far[265], including Putnam himself. In this regard, the burden of proof is on my opponent's side.

In conclusion, VTIWW adds to VTIW one extra condition for indeterminate identity between *x* and *y*, which is, the exclusion of determinate non-identity of *x* and *y*. Contrary to the *prima facie* concern that the change might cause Argument V to be question-begging, this alteration does not undermine my argument against Putnam. Instead, the

[265] In 6.3, I examined the argument from Evans against the possibility of indeterminate identity in general. However, Evan's argument is not an attempt to reinforce Putnam's multiple realisability argument.

conceivabilists' dilemma would come back to haunt anyone who wishes to rectify Putnam's argument by questioning the added condition.

7. 1. 3. Proof. Lemma – the first attempt

Now we are clear that VTIWW serves to be the correct formulation of the vagueness to indeterminacy theorem, and consequently the second premise of Argument V. To prove the truth of VTIWW is to infer it from the principles about vagueness and indeterminacy that we have established. So far, we have accepted the following:

Equivalence principles of V and Λ:

(VP1) $\mathrm{V}(t) \dashv\vdash \neg\Lambda(t)$

(VP2) $\mathrm{V}(t) \vee \Lambda(t)$

Equivalence principles of ∇ and Δ:

(EP1) $\Delta(S) \dashv\vdash \neg\nabla(S) \vee \neg\nabla(\neg S)$

(EP2) $\nabla(S) \dashv\vdash \neg\Delta(S) \vee \neg\Delta(\neg S)$

(EP1S) $\Delta(S) \dashv\vdash \neg\nabla(S)$

(EP2S) $\nabla(S) \dashv\vdash \neg\Delta(S)$

(EP3) $\Delta(S) \dashv\vdash \Delta(\neg S)$

(EP4) $\nabla(S) \dashv\vdash \nabla(\neg S)$

(EP5) $\Delta(S) \vee \nabla(S)$

Equivalence principles of ∇ and ▲:

(EP6) $\blacktriangle(S) \dashv\vdash S \,\&\, \Delta(S)$

(EP7) $\blacktriangle\neg(S) \dashv\vdash \neg S \,\&\, \Delta\neg(S)$

(EP8) $\Delta(S) \dashv\vdash \blacktriangle(S) \vee \blacktriangle\neg(S)$

(EP9) $\blacktriangle(S) \vee \blacktriangle\neg(S) \dashv\vdash \neg\nabla(S)$

(EP10) $\nabla(S) \dashv\vdash \neg\blacktriangle(S) \,\&\, \neg\blacktriangle\neg(S)$

In addition, we have also introduced and explained the *principle of difference* and *Leibniz's Consequence* (LLC).

Principle of difference:

(PoD) $Vx \,\&\, \Lambda y \rightarrow \exists F(\nabla Fx \,\&\, (\blacktriangle Fy \vee \blacktriangle \neg Fy))$

Leibniz's Consequence:

(LLC) $\forall F(\blacktriangle(x=y) \rightarrow (Fx \leftrightarrow Fy))$[266]

My task is to use the sixteen principles listed above, plus established rules of logic to derive VTIWW. In what follows, I will present two failed attempts. The first proof is a failure due to a *use-mention* confusion, which I will discuss in depth in 7.3.1. This proof consists of two parts. Part one uses sequent calculus to make visible an important implication of the theorem, that is previously unmentioned – if *x* is vague and *y* is crisp then it is not the case that it is determinately true that *x* is identical to *y*. Let us call this the *vagueness to the denial of determinately true identity* theorem (VDDI for short), which in notation reads:

$$Vx, \Lambda y \vdash \neg \blacktriangle(x=y).$$

Proof. VDDI from VTIWW

1. $Vx, \Lambda y, \neg \blacktriangle(x\neq y) \vdash \nabla(x=y)$
2. $Vx, \Lambda y \vdash \nabla(x=y), \blacktriangle(x\neq y)$ Negation right ($\neg$R)
3. $Vx, \Lambda y \vdash \neg \blacktriangle(x=y)$ EP10

From sequent calculus, we turn the theorem into its familiar variant that states that when one flanking term is vague and the other is crisp they are either indeterminately identical or solid-determinately non-identical. The reasoning then relies on EP10, which specifies the three truth-conditions involving '∇' – statements and '$\blacktriangle$' – statements: *S* is either indeterminately true, or determinately true, or determinately false. It follows that vague and crisp flanking terms lead to rule out the third conjunct. Hence, when the flanking terms of an identity statement are vague and crisp respectively, you have the denial of determinate identity:

$$Vx, \Lambda y \vdash \neg \blacktriangle(x=y)$$

[266] Similar to that of PoD, *x* and *y* in LLC are treated as free variables that are implicitly universally quantified.

Notice that while VDDI is provable from VTIWW, it also works vice versa:

Proof. VTIWW from VDDI

1.	$Vx, \Lambda y \vdash \neg\blacktriangle(x=y)$	
2.	$Vx, \Lambda y \vdash \nabla(x=y), \blacktriangle(x\neq y)$	EP10
3.	$Vx, \Lambda y, \neg\blacktriangle(x\neq y) \vdash \nabla(x=y)$	Negation left (¬L)

In short, the two proofs show that VTIWW is logically equivalent to VDDI. At this point, my strategy is a simple one, since proving the latter suffices to prove the former. I will use VDDI as the lemma for the theorem. This brings us to the second part of my proof – to prove VDDI using the sixteen principles at hand:

Proof. Lemma – (1)

$Vx, \Lambda y \vdash \neg\blacktriangle(x=y)$

1.	Vx	Assumption
2.	Λy	Assumption
3.	$\neg\neg\blacktriangle(x=y)$	Assumed for *reductio*
4.	$\blacktriangle(x=y)$	3, DN
5.	Vy	4, LLC
6.	$\neg\Lambda y$	5, VP1
7.	$\neg\blacktriangle(x=y)$	3, 2, 6, RAA

Let x be any vague term and let y be any crisp term. It is assumed for *reductio* that x=y is determinately true and a contradiction is derived. From this *reductio* assumption, it follows by *Leibniz's Consequence* that y has all x's properties and vice versa. Thus y is vague, because x is vague, which contradicts the assumption that y is vague. Therefore, it is not the case that x=y is determinately true, hence, establishing VDDI.

Combining *Proof. VTIWW's lemma* and *Proof. VDDI from VTIWW*, we have the first complete proof of VTIWW:

Proof. VTIWW - (1)

$Vx, \Lambda y, \neg\blacktriangle(x\neq y) \vdash \nabla(x=y)$

1.	Vx	Assumption

2.	Λy	Assumption
3.	$\neg\blacktriangle(x\neq y)$	Assumption
4.	$\neg\neg\blacktriangle(x=y)$	Assumed for *reductio*
5.	$\blacktriangle(x=y)$	4, DN
6.	Vy	5, LLC
7.	$\neg\Lambda y$	6 ,VP1
8.	$\neg\blacktriangle(x=y)$	4, 3, 7, RAA
9.	$\neg\blacktriangle(x=y)\ \&\ \neg\blacktriangle(x\neq y)$	3, 8, &I
10.	$\nabla(x=y)$	9, EP10

Proof. VTIWW - (1) seems to justify VTIWW, however it might come with a cost. Namely, in proposing this proof, I am treating vagueness and crispness as properties, and using *Leibniz's Consequence* to infer that *x* and *y* must be both vague (or both crisp). The critic could go on to claim that since *Leibniz's Consequence* applies to things, not to names, I therefore must be assuming that vagueness and crispness are properties of things, not of names. This is opposite to my initial stipulation of what 'V' and 'Λ' mean – that they are properties of names, not of things.

The key step of this possible criticism is to argue why *Leibniz's Consequence* only applies to things and not to names. If it applies also to names, then the worry seems toothless to my proof. In order to argue against the applicability of *Leibniz's Consequence* to names, my potential opponent might point to the following scenario:

Albert is known to his workmates as 'Al', but everyone in Albert's family calls him 'Bert'. Albert, the person known as 'Al', is determinately identical to Albert the person known as 'Bert', since it's one and the same person and we learned from Evans that self-identity is always determinate. Thus, every property possessed by Al the person is also possessed by Bert the person. In other words, *Leibniz's Consequence* goes through fine with Al the person and Bert the person:

AL1. $(\blacktriangle(\text{Al the person} = \text{Bert the person})\rightarrow\forall F(F(\text{Al the person})\leftrightarrow F(\text{Bert the person}))$

Can we say the same with 'Al' the name and 'Bert' the name? If someone wants to say yes, she is committed to claiming:

AL2. $(\blacktriangle(\text{'Al' the name} = \text{'Bert' the name})\rightarrow\forall F(F(\text{'Al' the name})\leftrightarrow F(\text{'Bert' the name}))$

At this point, my opponent would say that there are obvious counterexamples to AL2. For instance, 'Al' contains two letters. By *Leibniz's Consequence*, 'Bert' must contain two letters. Absurd! *Leibniz's Consequence*, therefore, does not work for names. Extending this reasoning to *Proof. Lemma - (1)*, the critic would be able to pinpoint the flaw in the proof. 'V' and 'Λ' are predicates of words, not things. Hence we cannot infer $\mathrm{V}y$ from the *reductio* assumption $\blacktriangle(x=y)$ via *Leibniz's Consequence*. In short, line 6 is not a permitted inference.

Although the first attempted proof of VTIWW fails, it is still promising in the sense that it brings attention to a valuable message. The last two steps of the proof lay out the general strategy for proving the theorem:

…

9. $\neg\blacktriangle(x=y)\ \&\ \neg\blacktriangle(x\neq y)$ — 3, 8, &I

10. $\nabla(x=y)$ — 9, EP10

The last two steps tell us that due to EP10, to prove VTIWW equates to proving:

$$\mathrm{V}x,\ \Lambda y,\ \neg\blacktriangle(x\neq y) \vdash \neg\blacktriangle(x=y)\ \&\ \neg\blacktriangle(x\neq y)$$

which by sequent calculus[267] is equivalent to the abovementioned lemma of VTIWW (Lemma for short, henceforth):

Lemma. $\mathrm{V}x,\ \Lambda y \vdash \neg\blacktriangle(x=y)$

Another way of emphasising the transition of desideratum of the proof is to say we need to prove $\neg\blacktriangle(x=y)\ \&\ \neg\blacktriangle(x\neq y)$ from the three assumptions: $\mathrm{V}x$, Λy, and $\neg\blacktriangle(x\neq y)$. And since one half of the conjunction, namely, $\neg\blacktriangle(x\neq y)$, is identical to one of the three assumptions, we only need to prove that the other half of the conjunction, namely, $\neg\blacktriangle(x=y)$, is derivable from the two remaining assumptions – $\mathrm{V}x$ and Λy. Thus, this brings our focus back to finding a proof for VDDI. This task was previously tried by *Proof. VTIWW's lemma*, which suffers the drawback I revealed in the 'Albert' example, namely, that we cannot infer $\mathrm{V}y$ from the *reductio* assumption $\blacktriangle(x=y)$ via *Leibniz's Consequence*. Hence, *Proof. VTIWW's lemma* fails to be a valid one for its line 5 is not a

[267] Proof. From $\mathrm{V}x,\ \Lambda y,\ \neg\blacktriangle(x\neq y) \vdash \neg\blacktriangle(x=y)\ \&\ \neg\blacktriangle(x\neq y)$ to VTIWW's lemma:

1. $\mathrm{V}x,\ \Lambda y,\ \neg\blacktriangle(x\neq y) \vdash \neg\blacktriangle(x=y)\ \&\ \neg\blacktriangle(x\neq y)$
2. $\mathrm{V}x,\ \Lambda y \vdash \neg\blacktriangle(x=y)$ — 1, SC Rule

permitted inference. It is clear that the way forward is to come up with another proof for the Lemma that does not involve the application of *Leibniz's Consequence* within contexts governed by 'V' and 'Λ' on terms, or better yet, a proof that does not appeal to any of Leibniz's reasoning at all.

7. 2. *Principle of harmonious relations* and the second proof

How are we going to achieve this? Looking back at the sixteen laws and principles concerning vagueness and indeterminacy, I discern three levels of determinacy status.

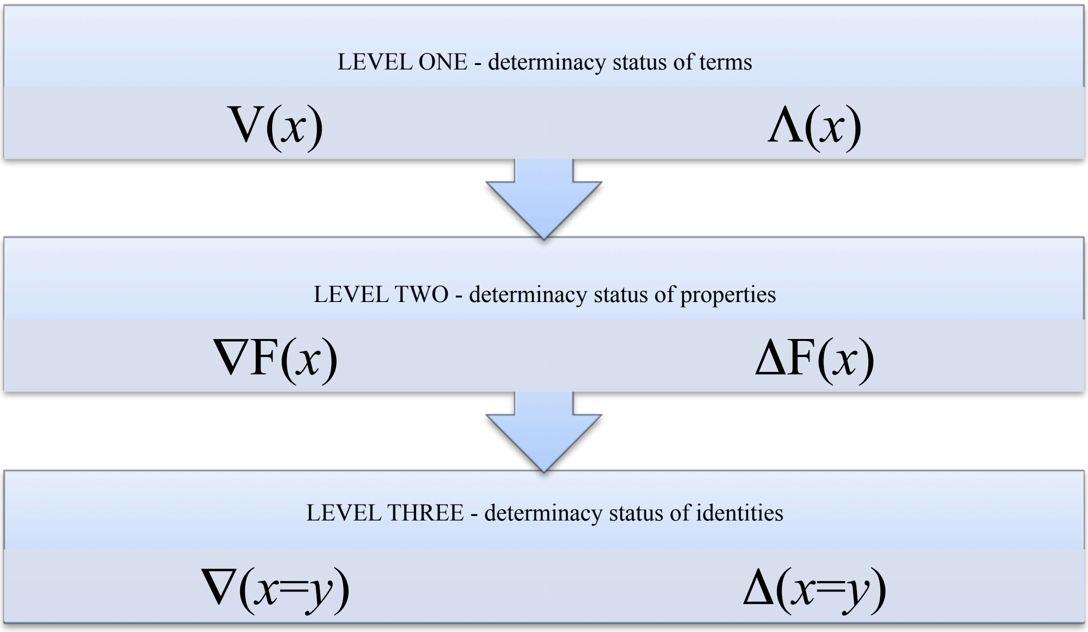
7. 2. 1. Three levels of determinacy status

- LEVEL ONE: determinacy status of terms. This is the level at which we state whether terms, such as 'P_h' and 'C-fibre firings' are vague or not. There are two statuses at this level: $V(x)$ and $\Lambda(x)$.
- LEVEL TWO: determinacy status of properties. There are two statuses at this level: $\nabla F(x)$ and $\Delta F(x)$.[268]
- LEVEL THREE: determinacy status of identities. This is the level at which we state whether identity statements, such as 'P_h = C-fibre firings', '$x=y$', are determinate or not. There are two statuses at this level : $\nabla(x=y)$ and $\Delta(x=y)$.[269]

[268] As stated by EP8, $\Delta F(x) \dashv\vdash \blacktriangle F(x) \vee \blacktriangle\neg F(x)$. Despite their different truth conditions, $\blacktriangle F(x)$ and $\blacktriangle\neg F(x)$ share the same determinacy status. The same applies to level three - $\blacktriangle(x=y)$ and $\blacktriangle(x\neq y)$ have identical determinacy status while having opposite truth conditions.

[269] The more natural and more general taxonomy will have determinacy status of statements as the third level, i.e. $\nabla(S)$ and $\nabla(S)$. Since the issue at hand only concerns identity statements, my categorizing is sufficient.

Figure 7-a


LEVEL ONE - determinacy status of terms
$\mathrm{V}(x)$
$\Lambda(x)$
LEVEL TWO - determinacy status of properties
$\nabla \mathrm{F}(x)$
$\Delta \mathrm{F}(x)$
LEVEL THREE - determinacy status of identities
$\nabla(x=y)$
$\Delta(x=y)$

Accordingly, the sixteen principles I have developed so far can be categorized into four groups:

- GROUP A: the *intra*-laws of LEVEL ONE. VP1 and VP2 belong to this group. They govern the *intra*-relations between $\mathrm{V}(x)$ and $\Lambda(x)$.

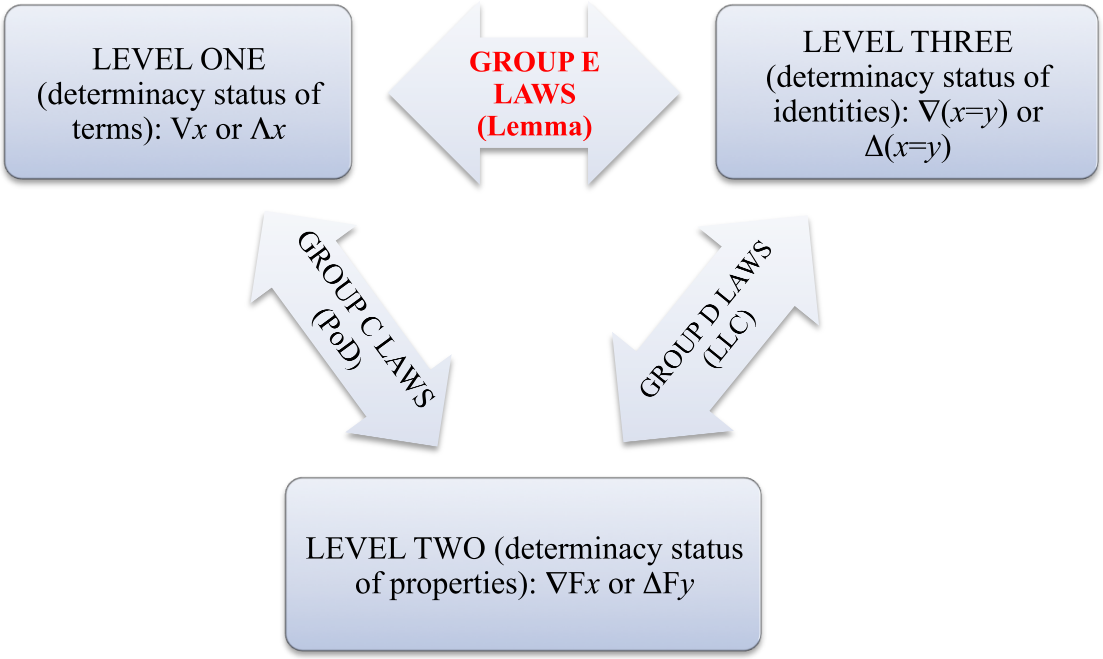
- GROUP B: the *intra*-laws of LEVEL THREE. EP1 to EP10 are principles at this level. EP1 to EP5 govern the *intra*-relations between $\nabla(x=y)$ and $\Delta(x=y)$; EP6 to EP10 are extensions to EP1~EP5 that take into account the ▲-variations of $\Delta(x=y)$.

- GROUP C: the *inter*-law of LEVEL ONE and LEVEL TWO. The *principle of difference* governs the *inter*-relation between the determinacy status of terms and the determinacy status of properties attributed to these terms.

- GROUP D: the *inter*-law of LEVEL TWO and LEVEL THREE. *Leibniz's Consequence* governs the *inter*-relation between the determinacy status of identity and the determinacy status of properties attributed to the flanking terms of the identity.

Using this taxonomy, we can quickly discover that Lemma, the thing we want to prove, is itself a prospective law that belongs to what might be called GROUP E. It is an *inter*-law of LEVEL ONE and LEVEL THREE. It governs the *inter*-relation between the determinacy status of terms and the determinacy status of identities that involve these terms.

Figure 7-b


LEVEL ONE (determinacy status of terms): $\mathrm{V}x$ or Λx
GROUP E LAWS (Lemma)
LEVEL THREE (determinacy status of identities): $\nabla(x=y)$ or $\Delta(x=y)$
GROUP C LAWS (PoD)
GROUP D LAWS (LLC)
LEVEL TWO (determinacy status of properties): $\nabla\mathrm{F}x$ or $\Delta\mathrm{F}y$

It follows that in order to prove a GROUP E principle, we can construct a *modus tollens* argument that only consists of two premises: a GROUP C principle and a GROUP D principle. In proving Lemma, the general form of the argument amounts to the following:

PREMISE ONE. An *inter*-law that says:

$\mathrm{V}x, \Lambda y \rightarrow$ such-and-such determinacy status(es) of Fx and Fy.

PREMISE TWO. An *inter*-law that says:

$\blacktriangle(x=y) \rightarrow$ NOT such-and-such determinacy status(es) of Fx and Fy.

CONCLUSION. $\mathrm{V}x, \Lambda y \vdash \neg\blacktriangle(x=y)$.

The role of PREMISE ONE is filled by the established *principle of difference*. We have also tried to fill PREMISE TWO with *Leibniz's Consequence* in the previous attempt but failed due to its inapplicability to predication of terms. In furtherance of producing a proof for Lemma, I think all there is left to do is to posit a tenable GROUP D principle that replaces *Leibniz's Consequence*. It will be preferably in this form:

$$\blacktriangle(x=y) \rightarrow \neg\exists\mathrm{F}(\nabla\mathrm{F}x \;\&\; \Delta\mathrm{F}y)$$

Hence, having the negation of PoD's consequent as the desired consequent of our new principle will be conducive to proving Lemma as they form a straightforward *modus tollens*.

7. 2. 2. The determinacy status of 'Π'

In pursuance of this important new principle, I am going to utilise the Surveyor Mango – Surveyor Nectarine example. Let us begin by adding a new term – 'NC', which stands for Surveyor Nectarine's crisp pegging of Mt. Cook. NC bears a relation, Π, to OE, namely that the entire NC is such and such kilometres away from OE. Due to its vagueness, OE lacks a clear boundary; it is thereby intuitive to say the exact distance between the NC and OE is Sorites-inducing. Thus, this particular relation is an indeterminate one:

(RM1) $\nabla\Pi(\mathrm{NC}, \mathrm{OE})$

In contrast, NC and NE are both crisp and therefore the distance between the two mountains is also clear and thereby determinate. In other words, the Π-relation that NC bears to NE is a determinate one:

(RM2) $\Delta\Pi(\mathrm{NC}, \mathrm{NE})$

Moreover, the same can be said for the different crisp terms of the exact same mountain such as ME and NE, in which case the distance between the pair is significantly shorter than the above pairing, but nevertheless determinate:

(RM3) $\Delta\Pi(\mathrm{ME}, \mathrm{NE})$

Furthermore, determinacy of the Π-relation also holds for identical crisp terms such as NC to itself, in which case the distance is zero, and determinately so:

(RM4) $\Delta\Pi(\mathrm{NC}, \mathrm{NC})$

The four cases above indicate a regularity concerning the different determinacy statuses of relations between terms of different determinacy statuses: RM1 shows that indeterminate relation holds for a vague term and a crisp term, whereas RM2, RM3 and RM4 exemplify the determinacy of the Π-relations held by crisp terms. Since the terms and the meaning of Π are arbitrarily chosen, we can reach a generalisation that works for all crisp terms and all binary relations between them:

Table 7-c

	Λy	Vy	Λx
Λx	$\Delta\Pi(x, y)$	$\nabla\Pi(x, y)$	$\Delta\Pi(x, x)$

Firstly, we can tell from the table that vagueness of a term results in the indeterminacy of the relation a crisp term bears to it. Call this the *Principle of vagueness to indeterminate relations* (PIR):

(PIR) $\forall\Pi((\Lambda x \,\&\, Vy) \rightarrow \nabla\Pi(x, y))$[270]

We can also conclude that when both terms are vague the relation that one term bears to another is determinate. This is the *Principle of crispness to determinacy relations* (PDR):

(PDR) $\forall\Pi(\Lambda x \,\&\, \Lambda y) \rightarrow \Delta\Pi(x, y))$

While being true, these two principles do not exhaust all possibilities of pairing of terms. So far we have considered the following: (Λz, Vx), (Λz, Λx), and (Λz, Λz). What about (Vx, Vx) – the pairing of identical vague terms? This brings us back to the familiar message voiced in Evans' argument – that self-identity is always determinate, even for vague terms. Thus, it is fair to say that all relations a vague term bears to itself must be determinate:

$$\forall\Pi(Vx \rightarrow \Delta\Pi(x, x))$$

Since determinacy of self-relations also applies to crisp terms, we can generalise the idea further and get the *Principle of determinate self-to-self relations* (PSR):

(PSR) $\forall\Pi(\Delta\Pi(x, x))$

So we have three principles in PIR, PDR, and PSR that tell us the determinacy status of the relations a term bears to itself in relation to the determinacy statuses of the corresponding terms. Here is a question: can we derive an axiom schema that captures all three principles? In other words, is there a common pattern that we can draw from the three types of pairing that yield $\Delta\Pi$, which is different from the one pairing that yields

[270] Hereinafter all variables are treated as free variables which are presumed to be universally quantified.

$\nabla\Pi$? I call attention to the determinacy status of identity statements that these pairings flank. Consider Table 7-d:

Table 7-d

Determinacy status of pairing/flanking terms	Λx, Vy	Λx, Λy	Λx, Λx	Vx, Vx
Determinacy status of $\Pi(x, y)$	$\nabla\Pi(x, y)$	$\Delta\Pi(x, y)$	$\Delta\Pi(x, x)$	$\Delta\Pi(x, x)$
Determinacy status of $x=y$	$\nabla(x=y)$	$\Delta(x=y)$	$\Delta(x=x)$	$\Delta(x=x)$

A pattern is visible. Namely, the determinacy statuses at the level of Π coincide with the determinacy statuses at the level of identity. To be precise, when the identity $x=y$ is determinate (this, of course includes both $\blacktriangle(x=y)$ and $\blacktriangle(x\neq y)$), the Π - relation that the flanking terms of the identity bear to each other is also determinate:

(PDD) $$\forall\Pi(\Delta(x=y) \rightarrow \Delta\Pi(x, y))$$

This *Principle of determinate identity to determinate relations* not only appears to be plausible from the above table; we have in fact proven its truth. With the introduction of the 'solid delta' - $\blacktriangle$, I have elucidated that $\Delta(x=y)$ includes two possibilities – $\blacktriangle(x=y)$ and $\blacktriangle(x\neq y)$. When the former is realised, that is, when x is determinately identical to y, then x and y are really one and the same thing. $\blacktriangle(x=y)$ is semantically equivalent to $\blacktriangle(x=x)$, which in turn, has the same truth condition as $\Delta(x=x)$. Consequently, $\Pi(x, y)$ must be determinate, just as $\Pi(x, x)$ is always determinate, as we have exemplified in RM4. On the other hand, when $x=y$ is determinately false, then, as in the RM2 and RM3 cases, $\Pi(x, y)$ is also determinate.

In PDD we have our first principle that governs the relation between the identity level and Π-level (i.e. a GROUP D law). Next, I am going to add a third item—a crisp term, z, to the level of terms, and investigate the determinacy status of $\Pi(z, x)$ and $\Pi(z, y)$. Given the crispness of z and the determinacy statuses of x, and y, we will have Table 7-e via PIR, PDR, and PSR:

Table 7-e

Determinacy status of pairing/flanking terms	$\Lambda z, \Lambda x, Vy$	$\Lambda z, \Lambda x, \Lambda y$	$\Lambda z, \Lambda x, \Lambda x$	$\Lambda z, Vx, Vx$
Determinacy status of $\Pi(z, x)$ and $\Pi(z, y)$	$\Delta\Pi(z, x)$ $\nabla\Pi(z, y)$	$\Delta\Pi(z, x)$ $\Delta\Pi(z, y)$	$\Delta\Pi(z, x)$ $\Delta\Pi(z, x)$	$\nabla\Pi(z, x)$ $\nabla\Pi(z, x)$
Determinacy status of $x=y$	$\nabla(x=y)$	$\Delta(x=y)$	$\Delta(x=x)$	$\Delta(x=x)$

We can therefore list three[271] more statements that contain the occurrences of z:

(ZPIR) $\forall\Pi(\Lambda x \,\&\, Vy \,\&\, \Lambda z) \rightarrow (\Delta\Pi(z, x) \leftrightarrow \nabla\Pi(z, y))$

(ZPDR) $\forall\Pi(\Lambda x \,\&\, \Lambda y \,\&\, \Lambda z) \rightarrow (\Delta\Pi(z, x) \leftrightarrow \Delta\Pi(z, y))$

(ZVV) $\forall\Pi(Vx \,\&\, Vy \,\&\, \Lambda z) \rightarrow (\nabla\Pi(z, x) \leftrightarrow \nabla\Pi(z, y))$

ZPIR and ZPDR are provable from, and thereby theorems of, PIR and PDR respectively. ZPIR indicates that the difference at the level of determinacy status of terms must manifest at the level of determinacy status of relations. In contrast, ZPDR says that crispness at the level of terms give you two determinate relations.

The remaining one—ZVV—is somewhat different, since it can be reduced to $\forall\Pi(Vx \,\&\, \Lambda z) \rightarrow (\nabla\Pi(z, x))$. Nevertheless, by putting forward ZVV, I wish to draw attention to a feature highlighted by ZPDR and ZVV. That is, when the terms have the same determinacy statuses (i.e. both crisp or both vague), the relations that these terms bear to a third and crisp term, z, must have the same determinacy statuses (i.e. both determinate or both indeterminate).

7. 2. 3. *Principle of harmonious relations*

In order to express this idea more clearly, we can merge ZPDR and ZVV to creat a single principle. Of course we can achieve this by simply conjoining the two and making a long and wordy universally quantified conjunctional statement: $\forall\Pi((\Lambda x \,\&\, \Lambda y \,\&\, \Lambda z) \rightarrow (\Delta\Pi(z, x) \leftrightarrow \Delta\Pi(z, y)) \,\&\, (Vx \,\&\, Vx \,\&\, \Lambda z) \rightarrow (\nabla\Pi(z, x) \leftrightarrow \nabla\Pi(z, x)))$. This looks long and ugly, indeed! To express the principle more efficiently and more elegantly, I suggest

[271] We could add to the list $\forall\Pi(\Lambda x \,\&\, \Lambda z) \rightarrow (\Delta\Pi(z, x))$ to capture the information in the fourth columns. But this formula is the same as the pre-established PIR, with only a change of variable.

that we can borrow the idea shown in PDD, that instead of making a GROUP C principle that governs the relation between the determinacy status of terms (LEVEL ONE) and the determinacy status of Π-relations (LEVEL TWO), we can have a GROUP E principle that bridges the level of identity (LEVEL THREE) and the level of Π-relations (LEVEL TWO). In more detail, we want to know that given a certain determinacy status of x=y, what is the determinacy status of the Π-relations that a crisp term, z, bears to x and y? Do the Π-relations in question share the same determinacy status?

With this rationale in mind, we now go back to the table and find that when x=y is determinate, Π(z, x) and Π(z, y) have the same determinacy status. Thus:

(PoR) $\forall\Pi(\Delta(x=y) \rightarrow (\Lambda z\rightarrow((\Delta\Pi(z, x)\leftrightarrow\Delta\Pi(z, y)) \vee (\nabla\Pi(z, x)\leftrightarrow\nabla\Pi(z, y))))$

This is the *principle of harmonious relations for determinate identity* (PoR for short). The name of this new principle might have a funny political or sociological connotation to it, but by *harmonious* it merely refers to the same determinacy status shared between the relations z bears separately to x and y, when x and y are determinately identical. The deltas always point in the same direction; they are always in *harmony*, metaphorically speaking.

The principle is still quite verbose in this initial form. Thanks to the equivalence principle EP1S, we know that $\Delta(x=y) \dashv\vdash \neg\nabla(x=y)$, so we only need one half of the disjunction. The principle thus can be reduced to:

(PoR) $\forall\Pi(\Delta(x=y) \rightarrow (\Lambda z\rightarrow((\Delta\Pi(z, x)\leftrightarrow\Delta\Pi(z, y))))$[272]

We can shorten it further by eliminating the Λz-condition. In other words, when x=y is a determinate identity, the relations that z bears to x and y will not be differed in determinacy status, *regardless of the determinacy status of z*. Table 7-f explains this point:

Table 7-f

Determinacy status of pairing/flanking terms	Vz, Λx, Λy	Vz, Λx, Λx	Vz, Vx, Vx
Determinacy status of	∇Π(z, x)	∇Π(z, x)	∇Π(z, x)

[272] By EP1S, it can be also written as $\forall\Pi(\Delta(x=y) \rightarrow (\Lambda z\rightarrow((\nabla\Pi(z, x)\leftrightarrow\nabla\Pi(z, y))))$.

$\Pi(z, x)$ and $\Pi(z, y)$	$\nabla\Pi(z, y)$	$\nabla\Pi(z, x)$	$\nabla\Pi(z, x)$
Determinacy status of x=y	$\Delta(x{=}y)$	$\Delta(x{=}x)$	$\Delta(x{=}x)$

This time we add a vague z instead of a crisp z to the equation and compile the determinacy status of Π that z bears to the flanking terms of $\Delta(x{=}y)$. One result remains unaltered: there is no difference in the determinacy status of $\Pi(z, x)$ and $\Pi(z, y)$. We shall therefore conclude with a final, updated version of PoR as follows:

(PoR) $\Delta(x{=}y) \rightarrow \forall\Pi(\Delta\Pi(z, x)\leftrightarrow\Delta\Pi(z, y))$

which in English reads:

(PoR) If an identity statement is determinate (determinately true or determinately false), then all relations a third term bears to the flanking terms of the identity must share the same determinacy status.

7. 2. 4. Proof. Lemma – the second attempt

With the addition of PoR, we can remove the troublesome *Leibniz's Consequence* and insert PoR in the proof of Lemma. Hence, our second attempted proof looks like:

Proof. Lemma – (2)

$Vx, \Lambda y \vdash \neg\blacktriangle(x{=}y)$

1.	Vx	Assumption
2.	Λy	Assumption
3.	$\blacktriangle(x{=}y)$	Assumed for *reductio*
4.	$\exists F(\nabla Fx \,\&\, (\blacktriangle Fy \vee \blacktriangle\neg Fy))$	1, 2, PoD
5.	$\exists F(\nabla Fx \,\&\, \Delta Fy)$	4, EP8
6.	$\exists F(\neg\Delta Fx \,\&\, \Delta Fy)$	5, EP2S
7.	$\neg\forall F(\Delta Fx \leftrightarrow \Delta Fy)$	6, quantification calculus
8.	$\blacktriangle(x{=}y) \rightarrow \forall\Pi(\Delta\Pi(z, x) \leftrightarrow \Delta\Pi(z, y))$	PoR, EP8
9.	$\forall\Pi(\Delta\Pi(z, x) \leftrightarrow \Delta\Pi(z, y))$	3, 8, →E
10.	$\forall\Pi(\lambda w(\Delta\Pi(z, w))x \leftrightarrow \lambda w(\Delta\Pi(z, w))y)$	9, property abstraction
11.	$\neg\blacktriangle(x{=}y)$	3, 7, 10, RAA

Let x be a vague term and y be a crisp term. Since the proof is via *reductio*, it is assumed that x=y is determinately true and a contradiction is derived. From the vagueness of x and crispness of y, it follows by *principle of difference* that there is a predicate such that x has it indeterminately and y has it or lacks it determinately. This in turn by quantification calculus and equivalence principles EP8 and EP2S can be reduced to saying it is not the case that for all predicates F, x has or lacks F determinately if and only if y has or lacks F determinately. Call this the *consequence of difference*. From the *reductio* assumption, it follows by *principle of harmonious relations* that for all relations, z bears a relation to x determinately iff z bears that relation to y determinately. The proof then uses lambda abstraction to turn this biconditional statement into: for all properties such that z bears Π to x, x has the property determinately iff y has the property determinately. Taking this with *consequence of difference*, a contradiction is yielded and thus the *reductio* assumption is negated, so Lemma is established.

This proof evades the problem of inapplicability of *Leibniz's Consequence* to terms, by not using any Leibniz's reasoning as its inference rules. It follows from the aforementioned grand rationale of proving a GROUP E principle from yielding a *modus tollens* with a GROUP C principle and a GROUP D one. More precisely, it tries to expose a contradiction between the *reductio* assumption that $\mathrm{V}x, \Lambda y \vdash \blacktriangle(x=y)$, and the joint truth of PoD and PoR. However, this also appears to be the biggest pitfall of *Proof. Lemma –(2).* Is the proof really uncovering a contradiction?

To reveal the answer to the above question, we need to scrutinise the proof's final steps. The conclusion is derived via *reductio ad absurdum* from line 3, line 7, and line 10. Line 7 and line 10 are purportedly generating a contradiction between them. Line 7, namely, $\neg\forall F(\Delta Fx \leftrightarrow \Delta Fy)$, the so-called *consequence of difference*, is an entailment of PoD. Its derivation is sound, barring criticism to PoD and the equivalence principles in use. To contradict the *consequence of difference* is therefore to show an example of NCD:

(NCD) $\forall F(\Delta Fx \leftrightarrow \Delta Fy)$

which means that all properties determinately possessed or lacked by *x* must also be possessed or lacked by *y*, and vice versa.

Line 10 supposedly conveys such a proposition. Line 10 is derived from line 8: $\blacktriangle(x=y) \rightarrow \forall\Pi(\Delta\Pi(z, x) \leftrightarrow \Delta\Pi(z, y))$, which in turn is an entailment of PoR. PoR states

$\Delta(x=y) \rightarrow \forall\Pi(\Delta\Pi(z, x) \leftrightarrow \Delta\Pi(z, y))$, and since $\Delta(S) \dashv\vdash \blacktriangle(S) \vee \blacktriangle\neg(S)$ (EP8), $\blacktriangle(x=y) \rightarrow \forall\Pi(\Delta\Pi(z, x) \leftrightarrow \Delta\Pi(z, y))$ also holds. Running *modus ponens* with line 8 and the *reductio* assumption $\blacktriangle(x=y)$ yields line 9 - $\forall\Pi(\Delta\Pi(z, x) \leftrightarrow \Delta\Pi(z, y))$. The next move, namely, the transition from 9 to 10, is rather tricky. 9 says for all Π, z bears or does not bear Π to x determinately if and only if z bears or does not bear Π to z. It is correct that we can run lambda abstraction on things inside a universally quantified scope, without doing universal instantiation first. In our case, doing so yields:

$$\forall\Pi(\lambda w(\Delta\Pi(z, w))x \leftrightarrow \lambda w(\Delta\Pi(z, w))y)$$

which says *for all Π, x has the property such that z bears or does not bear Π to x determinately iff z bears or does not bear Π to y determinately.*

So, this seems to be a way to say x has or lacks a certain property determinately iff *y* has or lacks a certain property, and hence $\Delta Fx \leftrightarrow \Delta Fy$.

Two worries emerge instantly. Firstly, does the lambda abstraction $\lambda w(\Delta\Pi(z, w))x$ have the same meaning as ΔFx? It is easy to argue no. In the former, Δ is inside the scope of λ, and it modifies Π. In the latter, however, Δ is placed immediately in front of F and thereby modifies F. A more accurate substitution for $\lambda w(\Delta\Pi(z, w))x$ should therefore be in the form of Fx, where F stands for *the property such that z bears or does not bear Π to x determinately*. $\lambda w(\Delta\Pi(z, w))x$ remains agnostic about whether this predication of *x* is determinate or not, contrariwise, ΔFx announces that Fx is determinate. In this regard, $\lambda w(\Delta\Pi(z, w))x \leftrightarrow \lambda w(\Delta\Pi(z, w))y)$ cannot be perceived as $\Delta Fx \leftrightarrow \Delta Fy$.

Even if we can read $\lambda w(\Delta\Pi(z, w))x \leftrightarrow \lambda w(\Delta\Pi(z, w))y)$ as saying $\Delta Fx \leftrightarrow \Delta Fy$, there is a second and more harmful problem. $\forall\Pi(\lambda w(\Delta\Pi(z, w))x \leftrightarrow \lambda w(\Delta\Pi(z, w))y$ does not mean $\forall F(\Delta Fx \leftrightarrow \Delta Fy)$. In the former, Π is being universally quantified where Π is just part of the meaning of the λ-abstracted property, whereas in the latter the entire property F is being quantified. In this regard, instead of $\forall F(\Delta Fx \leftrightarrow \Delta Fy)$, a more accurate substitution for $\forall\Pi(\lambda w(\Delta\Pi(z, w))x \leftrightarrow \lambda w(\Delta\Pi(z, w))y$ would be just $\Delta Fx \leftrightarrow \Delta Fy$, which does not share the form of NCD.

In sum, there are two unwelcome mismatches between line 10 and the contradiction of line 7 (NCD). NCD says that for *all* properties, *x* has it determinately iff *y* has it determinately, but Line 10 only shows that for *a particular kind of property and all* *instances of this property*, *x* has it determinately iff *y* has it determinately. It follows that the proof has not yet produced a genuine contradiction under $\blacktriangle(x=y)$. The *reductio* assumption is still not turned. As a result, line 11 is not a valid inference, and the entire proof is not a valid one.

7. 3. The final proof

Proof. Lemma – (2) is unfortunate indeed. But there is a silver lining out of this. To be specific, we now understand the general structure of how to prove Lemma. In my next attempt, I shall keep utilising PoD (GROUP C principle) and find another GROUP D principle in place of PoR. In other words, my third proof can retain the form and the top half (up to line 7) of my second proof:

Proof. Lemma – (3)

$Vx, \Lambda y \vdash \neg\blacktriangle(x=y)$

1.	Vx	Assumption
2.	Λy	Assumption
3.	$\blacktriangle(x=y)$	Assumed for *reductio*
4.	$\exists F(\nabla Fx \,\&\, (\blacktriangle Fy \vee \blacktriangle\neg Fy))$	1, 2, PoD
5.	$\exists F(\nabla Fx \,\&\, \Delta Fy)$	4, EP8
6.	$\exists F(\neg\Delta Fx \,\&\, \Delta Fy)$	5, EP2S
7.	$\neg\forall F(\Delta Fx \leftrightarrow \Delta Fy)$	6, quantification calculus
8.	*GROUP D principle*	
9.	$\forall F(\Delta Fx \leftrightarrow \Delta Fy)$	
10.	$\neg\blacktriangle(x=y)$	

Since I am going to use the same framework I developed for the second proof, the job at hand is, again, to find a legitimate GROUP D law and insert it into subsequent lines of the third proof. Hopefully, together with the *reductio* assumption, the prospective GROUP D law can entail a contradiction to line 7, which resembles the form of NCD. Looking at the list of vagueness and indeterminacy principles, we have two established GROUP D principles in LLC and PoR:

GROUP D laws:	
Leibniz's Consequence -	
(LLC)	$\forall F(\blacktriangle(x=y) \rightarrow (Fx \leftrightarrow Fy))$
Principle of harmonious relations -	
(PoR)	$\Delta(x=y) \rightarrow \forall\Pi(\Delta\Pi(z, x)\leftrightarrow\Delta\Pi(z, y))$

The latter, as I explained at the end of the last section, does not generate NCD with the assumption of $\blacktriangle(x=y)$. The former, namely LLC, on the other hand, seems a perfect fit for our requirement. Given $\blacktriangle(x=y)$ and LLC, NCD is yielded via *modus ponens*. Neat and tidy, nothing fancy is needed. An obstacle to this approach is, of course, the lesson we learned from the first proof – that LLC is inapplicable to the predication of terms. But if we can somehow remove this obstacle by limiting the inapplicability to some but not all types of predication of terms, LLC can be used to prove Lemma. Supposing this could be done, this would give us the following proof:

Proof. Lemma – (3)

$Vx, \Lambda y \vdash \neg\blacktriangle(x=y)$

1.	Vx	Assumption
2.	Λy	Assumption
3.	$\blacktriangle(x=y)$	Assumed for *reductio*
4.	$\exists F(\nabla Fx \ \& \ (\blacktriangle Fy \vee \blacktriangle\neg Fy))$	1, 2, PoD
5.	$\exists F(\nabla Fx \ \& \ \Delta Fy)$	4, EP8
6.	$\exists F(\neg\Delta Fx \ \& \ \Delta Fy)$	5, EP2S
7.	$\neg\forall F(\Delta Fx \leftrightarrow \Delta Fy)$	6, quantification calculus
8.	$\forall F(\Delta Fx \leftrightarrow \Delta Fy)$	3, LLC, *modus ponens*
9.	$\neg\blacktriangle(x=y)$	3, 7, 8, RAA

As discussed, the transition from 7 to 8 is only questionable on the basis that we have obtained a strong resistance to applying Leibniz's style inference rule. To get rid of such resistance, we need to revisit the source of it. Recall *Proof. Lemma – (1)*:

Proof. Lemma – (1)

$Vx, \Lambda y \vdash \neg \blacktriangle(x=y)$

1.	Vx	Assumption
2.	Λy	Assumption
3.	$\blacktriangle(x=y)$	Assumed for *reductio*
4.	Vy	1, 3, LLC, ↔E
5.	$\neg \Lambda y$	4, VP1
6.	$\neg \blacktriangle(x=y)$	2, 5, 3, RAA

7. 3. 1. Terms *used* & terms *mentioned*

The resistance to using LLC as a inference rule for line 4 of *Proof – (1)* (henceforth, line 1-4[273]) is motivated by the 'Albert' example, detailed in 7.1.3:

AL1. ($\blacktriangle$(Al the person = Bert the person) → ∀F(F(Al the person) ↔ F(Bert the person))

AL2. ($\blacktriangle$('Al' the name = 'Bert' the name) → ∀F (F('Al' the name) ↔ F('Bert' the name))

I concluded that counterexamples to AL2 could be produced without hard effort, and since the application of LLC in line 1.4 is more akin to AL2 than AL1, LLC is not permitted there. I think this conclusion is still correct. The hidden message that I haven't mentioned is that if the application of LLC resembles cases like AL1, then the application is legitimate.

What makes AL1 fundamentally different from AL2 then? I think one of the essential differentiating points is the *use-mention* distinction that this pair of statements showcases.

Definition. A term *t* is *used* in a sentence iff *t*'s occurrence in that sentence has the purpose of referring to *t*'s referent.

Definition. A term *t* is *mentioned* in a sentence iff *t*'s occurrence in that sentence does not have the purpose of referring to *t*'s referent.

In AL1, 'Al' and 'Bert' are *used* to refer to the person that these names refer to. As a result, 'F' in the second part of AL1 - ∀F(F(Al the person) ↔ F(Bert the person)) refers

[273] The first number denotes which proof; the second number denotes the line number. For example, line 6 of proof 2 will be referred to as line 2-6.

to properties of the person, not properties of the terms. Contra wise, in AL2, 'Al' and 'Bert' are merely *mentioned*. The terms do not refer to the person but only the terms themselves. As a result, 'F' in the second part of AL2- $\forall F(F($'Al' the name$) \leftrightarrow F($'Bert' the name$))$ refers to properties of the terms, not properties of the person those terms refer to.

It is common practice to signify this distinction by adding quotation mark to terms *mentioned*.[274] Thus, to make the distinction more apparent in our current example, we can rewrite the pair of sentences as:

AL1. $(\blacktriangle(\mathrm{Al} = \mathrm{Bert}) \rightarrow \forall F(F(\mathrm{Al}) \leftrightarrow F(\mathrm{Bert}))$

AL2. $(\blacktriangle(\text{'Al'} = \text{'Bert'}) \rightarrow \forall F\ (F(\text{'Al'}) \leftrightarrow F(\text{'Bert'}))$

It is still undisputed that AL2 is wrong for there are easy counterexamples. Hence, we can confirm that LLC fails for terms *mentioned*. In contrast, AL1 is a legitimate instantiation of LLC. It has the form of the standard Leibniz's Law except for the addition of '▲'. In a sense, if Leibniz's Law is applicable to describe the properties of Al the person and Bert the person given their identity, then *Leibniz's Consequence* must also be applicable to describe the properties of Al the person and Bert the person given their *determinate* identity. *Leibniz's Consequence*, in its essence, is just a weaker variant of Leibniz's Law. For this particular reason, we can also confirm that LLC works for terms *used*.

Moreover, to accommodate my agnosticism about ontic vagueness, which leads to my stipulation that 'V' and 'Λ' are properties of terms *mentioned*, not terms *used*, my intended uses of '*x*' and '*y*' in line 1.4 and its derivation are terms *mentioned*, not terms *used*. Therefore, LLC cannot be used to infer 1.4, and this ultimately costs *Proof. Lemma – (1)*'s validity.

But this doesn't seem to be the case for line 3.8. In deriving $\forall F(\Delta Fx \leftrightarrow \Delta Fy)$, LLC does not apply to any properties of terms *mentioned*. The derivation from LLC and $\blacktriangle(x=y)$ via *modus ponens*, where '*x*' and '*y*' in $\blacktriangle(x=y)$ are terms *used*. Furthermore, this is compatible with the nature of '*x*' and '*y*' in lines 3.4, 3.5, 3.6, 3.7. Thus, there is no mismatch of variables in lines 3.3, 3.7, and 3.8, since the '*x*' and '*y*' in those lines are all

[274] Cappelen, Herman and Lepore, Ernest, "Quotation", *The Stanford Encyclopedia of Philosophy* (Spring 2012 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2012/entries/quotation/>.

terms used. Hence, the final step of *Proof. Lemma – (3)* is also explained and immune to the 'Albert' counter example.

Having the *use-mention* distinction allows us to ascertain the boundary of LLC's usage. In particular, it explains away the worry of applying LLC to terms *used*, despite the fact that the 'Albert' counterexample certainly exposes a group of terms to which LLC is inapplicable, namely, the group of terms *mentioned*. The correct message out of this should not be that we need a proof of Lemma that does not appeal to any Leibniz's reasoning at all. Rather, the way forward is to come up with a proof for Lemma that does not involve the application of LLC to terms *mentioned*. As I have already suggested, *Proof. Lemma – (3)* is such a proof.

Nevertheless, there is one final difficulty that this way of explaining away the 'Albert' worry entails. That is, the *principle of difference* would have to consist of both '*x*' and '*y*' *used* and '*x*' and '*y*' *mentioned*. The principle states that $\mathrm{V}x \;\&\; \Lambda y \rightarrow \exists F(\nabla Fx \;\&\; (\blacktriangle Fy \vee \blacktriangle \neg Fy))$. Due to my stipulation of vagueness and crispness as properties of terms *mentioned*, the part on the left hand side of '→', the antecedent, contains terms *mentioned*. Due to my desire to legitimately apply LLC, the part on the right hand side of '→', the consequent, contains terms *used*. *Prima facie*, one might consider this discrepancy a pitfall – how could a single statement contain both '*x*'-*used* and '*x*'-*mentioned*? On second thought, this is hardly a problem at all. "The name of Al is 'Al'" is an example of such a statement, and there are lots of examples of this kind that we commonly use. On a more positive note though, I think the more important point that this worry might be revealing is a fascinating fact about the role of PoD. The principle, in the way I have described so far, is a law that transfer the properties of terms *mentioned* to the properties of *terms used*. It can be said that this conditional principle is, in fact, designed to include the *use-mention* discrepancy in its consequent and antecedent, respectively. It regulates that when two *mentioned* terms *x* and *y* have differentiating determinacy statuses, there must be a property such that *x-used* possesses or lacks indeterminately and *y-used* possesses or lacks determinately.

If my analysis so far is correct, we finally have a proof of Lemma that works. It relies on the principles of vagueness and indeterminacy I have developed:

List 7-g

Equivalence principles of V and Λ (GROUP A principle)*:*	
(VP1)	$\mathrm{V}(t) \dashv\vdash \neg\Lambda(t)$
(VP2)	$\mathrm{V}(t) \vee \Lambda(t)$
Equivalence principles of ∇ and Δ (GROUP B principle)*:*	
(EP1)	$\Delta(S) \dashv\vdash \neg\nabla(S) \vee \neg\nabla(\neg S)$
(EP2)	$\nabla(S) \dashv\vdash \neg\Delta(S) \vee \neg\Delta(\neg S)$
(EP1S)	$\Delta(S) \dashv\vdash \neg\nabla(S)$
(EP2S)	$\nabla(S) \dashv\vdash \neg\Delta(S)$
(EP3)	$\Delta(S) \dashv\vdash \Delta(\neg S)$
(EP4)	$\nabla(S) \dashv\vdash \nabla(\neg S)$
(EP5)	$\Delta(S) \vee \nabla(S)$
Equivalence principles of ∇ and ▲(GROUP B principle)*:*	
(EP6)	$\blacktriangle(S) \dashv\vdash S \,\&\, \Delta(S)$
(EP7)	$\blacktriangle\neg(S) \dashv\vdash \neg S \,\&\, \Delta\neg(S)$
(EP8)	$\Delta(S) \dashv\vdash \blacktriangle(S) \vee \blacktriangle\neg(S)$
(EP9)	$\blacktriangle(S) \vee \blacktriangle\neg(S) \dashv\vdash \neg\nabla(S)$
(EP10)	$\nabla(S) \dashv\vdash \neg\blacktriangle(S) \,\&\, \neg\blacktriangle\neg(S)$
Principle of difference (GROUP C principle)*:*	
(PoD)	$\mathrm{V}x \,\&\, \Lambda y \rightarrow \exists F(\nabla Fx \,\&\, (\blacktriangle Fy \vee \blacktriangle\neg Fy))$
Leibniz's Consequence (GROUP D)*:*	
(LLC)	$\forall F(\blacktriangle(x=y) \rightarrow (Fx \leftrightarrow Fy))$
Principle of harmonious relations (GROUP D principle)*:*	
(PoR)	$\Delta(x=y) \rightarrow \forall\Pi(\Delta\Pi(z, x)\leftrightarrow\Delta\Pi(z, y))$

After surveying alternative options in *Proof Lemma – (1)* and *Proof Lemma – (2)* and scrutinising every putative and tentative logical flaw thereof, I now present what I believe to be the most persuasive proof for VTIWW as below:

Proof. Lemma

$\mathrm{V}x, \Lambda y \vdash \neg\blacktriangle(x=y)$

1.	$\mathrm{V}x$	Assumption
2.	Λy	Assumption
3.	$\blacktriangle(x=y)$	Assumed for *reductio*
4.	$\exists F(\nabla Fx \ \& \ (\blacktriangle Fy \vee \blacktriangle\neg Fy))$	1, 2, PoD
5.	$\exists F(\nabla Fx \ \& \ \Delta Fy)$	4, EP8
6.	$\exists F(\neg\Delta Fx \ \& \ \Delta Fy)$	5, EP2S
7.	$\neg\forall F(\Delta Fx \leftrightarrow \Delta Fy)$	6, quantification calculus
8.	$\forall F(\blacktriangle(x=y) \rightarrow (Fx \leftrightarrow Fy))$	LLC
9.	$\forall F(\Delta Fx \leftrightarrow \Delta Fy)$	3, 8, MPP
10.	$\neg\blacktriangle(x=y)$	3, 7, 9, RAA

Proof. Lemma to VTIWW

1.	$\mathrm{V}x, \Lambda y \vdash \neg\blacktriangle(x=y)$	Lemma
2.	$\mathrm{V}x, \Lambda y \vdash \nabla(x=y), \blacktriangle(x\neq y)$	EP10
3.	$\mathrm{V}x, \Lambda y, \neg\blacktriangle(x\neq y) \vdash \nabla(x=y)$	Negation left (¬L)

As explained in 7.3.3, the complete proof of the vagueness to indeterminacy theorem is two-fold. The first part is a proof of Lemma in the form of natural deduction. The reasoning goes: let *x-mentioned* be a vague term and *y-mentioned* be a crisp term and let us assume for *reductio* that *x-used* and *y-used* are determinately identical. From the assumption of *x-mentioned* being a vague term and *y-mentioned* being a crisp term, it follows from the *principle of difference* that there is a property *F* such that *x-used* possesses *F* indeterminately and *y-used* either possesses *F* determinately or lacks *F* determinately. This in turn yields the *consequence of difference* - it is not the case that for all *F*, *x-used* has or lacks *F* determinately iff *y-used* has or lacks *F* determinately. Next, via *modus ponens*, *Leibniz's Consequence* and the *reductio* assumption yields that for all *F*, *x-used* has or lacks *F* determinately iff *y-used* has or lacks *F* determinately – the negation of the *consequence of difference*. A contradiction is therefore derived and the *reductio* assumption is turned.

The second part then uses sequent calculus to transform Lemma into VTIWW. The steps are comparatively simple and straightforward: first, we turn $\neg\blacktriangle(x=y)$ into $\nabla(x=y) \vee \blacktriangle(x\neq y)$ by the equivalence principle EP10. The formula consequently becomes: $Vx, \Lambda y \vdash \nabla(x=y), \blacktriangle(x\neq y)$. We then apply the 'negation left' rule to swing $\blacktriangle(x\neq y)$ from the right to the left of the turnstile and add to it a negation mark: $Vx, \Lambda y, \neg\blacktriangle(x\neq y) \vdash \nabla(x=y)$. VTIWW is established.

7. 4. Conclusion

We finally have an error-free *positive* argument that explains why and how a vague term and a crisp term lead to indeterminate identity that holds between these terms. Having this *positive* argument and the *negative* argument that rejects Evans' proof for the absurdity of indeterminate identity, I believe I can confidently conclude the truth of the V2 in my argument against Putnam. A small modification on the exact wording of this premise is required, however. The original V2 states that *if one flanking term of an identity statement is vague, and the other flanking term is crisp, then the identity statement is indeterminate*. This amounts to VTIW, which, as I detailed earlier, is wrong. Instead, the correct way of putting the vagueness to indeterminacy theorem is VTIWW: *If one flanking term of an identity statement is vague, and the other flanking term is crisp, and they are not determinately non-identical, then the identity statement is indeterminate*. This modification could invite a criticism that renders my argument against Putnam circular. But I have provided in 7.1.2 an explanation for why such criticism fails to undermine the indeterminacy objection.

To summarise the results of Part II, two important things have been achieved. First, the foundation of a new logic of identity has been laid. Second, this new logic has been applied in order to undermine Putnam's multiple realisability objection against identity theory. My overall reply to Putnam is a logic-tinted one. It is through this logic-lens that I am envisaging not only an indeterminate mind-brain identity but also a theory of indeterminate identity in general. Specifically, List 7-g is a compilation of every principle concerning vagueness and indeterminacy that I have proposed during the course of the current dissertation. Some of them are purely definitional, such as VP1. Some of them are postulated and I have explained their plausibility, such as PoD and PoR. The list does not exhaust the scope of all the rules there are to know about V and ∇. For example, there are many helpful theorems concerning the relationship between ∇ and ▲ that we can derive from the current list, such as $▲\neg(S) \rightarrow \neg\nabla(S)$[275]. In short, the formalisation of vagueness and indeterminacy is in its infancy, and it's yet to be applied to the mind-body problem in a thorough going way.

Finally, mind-brain identity theory has gone through its rise and fall since its inception. Among those arguments that have contributed to its fall, Kripke's modal argument, Chalmers' zombie argument, and Putnam's multiple realisability argument are undeniably three major components. The mirroring reply to Kripke and Chalmers pinpoints a problem in their reasoning, namely, that for each of their argument has a corresponding mirror argument that is deductively valid and has a conclusion contradicting the conclusion of the mirrored argument. In order to break the mirror, the most tenable option is to construct an argument that blocks the mirroring conceivability premise of the mirror argument, and doing so would in itself prove the falsity of identity theory, independently of the original modal argument and zombie argument. However, since this master argument hasn't been produced yet, the burden of proof is not on the identity theorists' side. Thus, the mirroring reply has enabled them to put the ball in the opposition's court. As things stand, two famous objections to identity theory are suspended, until this master argument is constructed – if it ever is.

The indeterminacy reply to Putnam's multiple realisability argument, on the other hand, seems decisive. In the last three Chapters we have established that Putnam's multiple realisability objection fails to damage indeterminate identity theory since indeterminate identity is neither transitive nor super-transitive. However, one might voice a reservation about the value of the indeterminacy reply, namely, how attractive is the indeterminate form of mind-brain identity theory? To answer this question, let us revisit the dialectic so far. Putnam makes it clear that his objection is aimed at Smart's identity theory. Neither Smart, nor any other proponents of identity theory distinguished between determinate and indeterminate identity. However, Smart should have done so, since P_h = C must be an indeterminate identity, given the foregoing argument turning on the

[275] This theorem is derivable from EP1, EP3, and EP7.

vagueness of P_h. So, identity theory, charitably reconstructed, holds that $\nabla(P_h = C)$. Therefore, Putnam's objection undermines only the determinate form of the theory, which is equally undermined by the argument here. Thus, identity theorists have one less objection to worry about.

Given these results, what of my own position on the mind-brain problem? I stated at the beginning of this thesis that I am going to offer rejections to Kripke's, Chalmers', and Putnam's arguments against identity theory. I take it that these rebuttals have rehabilitated identity theory, in the sense that key objections to it had been cleared away, and are no longer impediments to endorsing the theory. However, my aim in this thesis has not been to attempt to establish the truth of mind-brain identity, but rather to explain why three famous objections to the theory are flawed. But I remain agnostic regarding the solution to the mind-brain problem.

References

Anscombe, G. E. M. & Gerach, P. T. (1961), *Three Philosophers*, Blackwell.

Bayne, S. R. (1988), ‘Kripke’s Cartesian Argument’, *Philosophia* 18(July), 265--270.

Bechtel, W. P. & Mundale, J. (1999), ‘Multiple Realizability Revisited: Linking Cognitive and Neural States’, *Philosophy of Science* 66(2), 175--207.

Bechtel, W. P., Mandik, P., Mundale, J. & Stufflebeam, R. S. (2001), ed., *Philosophy and the Neurosciences: A Reader*, Blackwell.

Block, N. (1980a), ‘Functionalism’, in Ned Block, ed., *Readings in the Philosophy of Psychology*, Cambridge: Harvard University Press.

Block, N. (1980b), *Readings in Philosophy of Psychology*, Cambridge: Harvard University Press.

Block, N. (1980c), ‘Troubles with Functionalism’, in Ned Block, ed., *Readings in the Philosophy of Psychology*, Cambridge: Harvard University Press.

Blum, A. (1989), ‘Bayne on Kripke’, *Philosophia* 19(4), 455--456.

Block, N. & Fodor, J. A. (1972), ‘What Psychological States Are Not’, *Philosophical Review* 81(April), 159--81.

Boyd, R. (1980), ‘Materialism Without Reductionism: What Physicalism Does Not Entail’, *in* Ned Block, ed., *Readings in the Philosophy of Psychology*, Vol. 1, pp. 1--67.

Braddon-Mitchell, D. & Jackson, F. (2007), *The Philosophy of Mind and Cognition*, Blackwell.

Broome, J. (1984), ‘Indefiniteness in Identity’, *Analysis* 44(1), 6--12.

Brown, R. (2010) ‘Deprioritizing the A Priori Arguments Against Physicalism’, *Journal of Consciousness Studies* 17/3–4: 47–69.

Burgess, J. A. (1989), ‘Vague Identity: Evans Misrepresented’, *Analysis* 49(3), 112--119.

Byrne, A. (2007), ‘Possibility and Imagination’, *Philosophical Perspectives* 21(1), 125--144.

Campbell, D. I., Copeland, J. & Deng, Z. (2017), ‘The Inconceivable Popularity of Conceivability Arguments’, *Philosophical Quarterly* 67 (267), 223--240.

Capitan, W. H. & Merrill, D. D. (1967), *Art, Mind, and Religion*, University of Pittsburgh Press.

Cappelen, H. & Lepore, E., ‘Quotation’, *The Stanford Encyclopedia of Philosophy* (Spring 2012 Edition), Edward N. Zalta (ed.), URL=<http://plato.stanford.edu/archives/spr2012/entries/quotation/>.

Carnap, R. (1953), ‘The Interpretation of Physics’, *in* H. Feigl & M. Brodbeck, ed., *Readings in the Philosophy of Science*, New York: Appleton-Century-Crofts, 309—18.

Castaneda, H.-N. (1967), *Intentionality, Minds and Perception*, Wayne State University Press.

Chalmers, D. J. (2002a), ‘The Components of Content’, *in* David J. Chalmers, ed., *Philosophy of Mind: Classical and Contemporary Readings*, Oxford University Press, 608—33.

Chalmers, D. J. (2002b), ‘Does Conceivability Entail Possibility?’ in Tamar S. Gendler & John Hawthorne, ed., *Conceivability and Possibility*, Oxford University Press, pp. 145--200.

Chalmers, D. J. (2002c), ‘On Sense and Intension’, *Philosophical Perspectives* 16(s16), 135--82.

Chalmers, D. (2002d). (ed.) *Philosophy of Mind: Classical and Contemporary Readings.* New York: Oxford University Press.

Chalmers, D. J. (2004), ‘Epistemic Two-Dimensional Semantics’, *Philosophical Studies* 118(1-2), 153--226.

Chalmers, D. J. (2006a), The Foundations of Two-Dimensional Semantics, *in* Manuel Garcia-Carpintero & Josep Macia, ed., *Two-Dimensional Semantics: Foundations and Applications*, Oxford University Press, pp. 55--140.

Chalmers, D. (2006b). ‘Scott Soames’ Two-Dimensionalism. For a session at the meeting of the American Philosophical Association, Central Division, in April 2006. URL http://consc.net/papers/soames2d.pdf

Chalmers, D. J. (2010), *The Character of Consciousness*, Oxford University Press.

Chalmers, D. J. (2011), ‘Propositions and Attitude Ascriptions: A Fregean Account’, *Noûs* 45(4), 595--639.

Chalmers, D. (2012), *Constructing the World*, Oxford University Press.

Chalmers, D. and Jackson, F. (2001). 'Conceptual Analysis and Reductive Explanation' *Philosophical Review* 110: 315-61.

Cheyne, C. & Pigden, C. (2006), 'Negative Truths From Positive Facts', *Australasian Journal of Philosophy* 84(2), 249--265.

Chisholm, R. M. (1957), *Perceiving: A Philosophical Study*, Cornell University Press.

Copeland, B. J. (1994), 'Vagueness and Bivalence', *Proceedings of the Aristotelian Society* 68(n/a), 193--200.

Copeland, B. J. (1997), 'Vague Identity and Fuzzy Logic', *Journal of Philosophy* 94(10), 514--534.

Copeland, B. J. (2000), 'Indeterminate Identity, Contingent Identity, and Property Identity, Aristotelian-Style', *Philosophical Topics* 28(1), 11--25.

Copeland, B. J. (2002), 'The Genesis of Possible Worlds Semantics', *Journal of Philosophical Logic* 31(2), 99--137.

Couch, M. B. (2004), 'Discussion: A Defense of Bechtel and Mundale', *Philosophy of Science* 71(2), 198--204.

Cresswell, M. J. (1987), 'Aristotle's Phaedo', *Australasian Journal of Philosophy* 65(2), 131--155.

Cottingham, J., Stoothoff, R. & Murdoch, D. (1984), *The Philosophical Writings of Descartes*, Cambridge University Press.

Dummett, M. (1975), 'Wang's Paradox', *Synthese* 30(3-4), 201--32.

Evans, G. (1978), 'Can There Be Vague Objects?', *Analysis* 38(4), 208.

Feigl, H. (1958), 'The 'Mental' and the 'Physical'', *Minnesota Studies in the Philosophy of Science* 2, 370--497.

Feigl, H. & Brodbeck, M. (1953), ed., *Readings in the Philosophy of Science*, New York: Appleton-Century-Crofts

Feldman, F. (1973), 'Kripke's Argument Against Materialism', *Philosophical Studies* 24(November), 416--19.

Feldman, F. (1974), 'Kripke on the Identity Theory', *Journal of Philosophy* 71(October), 665--76.

Fine, K. (1975), 'Vagueness, Truth and Logic', *Synthese* 30(3-4), 265--300.

Fodor, J. A. (1968), *Psychological Explanation: An Introduction To The Philosophy Of Psychology*, New York: Random House.

Fodor, J. A. (1974), 'Special Sciences', *Synthese* 28(2), 97--115.

Frankish, K. (2007) 'The Anti-Zombie Argument', *Philosophical Quarterly* 57/229: 650–666.

Garcia-Carpintero, M. & Macia, J. (2006) ed., *Two-Dimensional Semantics: Foundations and Applications*, Oxford University Press.

Garrett, B. J. (1988), 'Vagueness and Identity', *Analysis* 48(3), 130--134.

Garrett, B. (1991), 'Vague Identity and Vague Objects', *Noûs* 25(3), 341--351.

Gendler, T. S. & Hawthorne, J. (2002), *Conceivability and Possibility*, Oxford University Press.

Guttenplan, S. D. (1994), ed., *A Companion to the Philosophy of Mind*, Cambridge: Blackwell.

Hardcastle, V. (2001), The Nature of Pain, *in* William P. Bechtel; Pete Mandik; Jennifer Mundale & Robert S. Stufflebeam, ed., *Philosophy and the Neurosciences: A Reader*, Blackwell, , pp. 295--311.

Hartshorne, C. (1965) *Anselm's Discovery: A Re-Examination of the Ontological Proof for God's Existence*. La Salle, IL: Open Court.

Hume, D. (2000), *A Treatise of Human Nature*, Oxford University Press.

Jackson, F. (1982), 'Epiphenomenal Qualia', *Philosophical Quarterly* 32, 127--36.

Jackson, F. (2003), Mind and Illusion, *in* Anthony O'Hear, ed., *Royal Institute of Philosophy Supplement*, Cambridge University Press, pp. 421--442.

Jackson, F., Pargetter, R. & Prior, E. W. (1982), 'Functionalism and Type-Type Identity Theories', *Philosophical Studies* 42(September), 209--25.

Johnsen, B. (1989), 'Is Vague Identity Incoherent?', *Analysis* 49(3), 103--112.

Kendel, E., Schwartz, J. & Jessell, T. (2000), ed. *Principles of Neural Science*, McGraw-Hill.

Kim, J. (1972), 'Phenomenal Properties, Psychophysical Laws and the Identity Theory', *The Monist* 56(April), 178--92.

Kim, J. (1992), ‘Multiple Realization and the Metaphysics of Reduction’, *Philosophy and Phenomenological Research* 52(1), 1--26.

Kim, J. (2011), *Philosophy of Mind*, Westview Press.

Kirwan, C. (1993), *Metaphysics: Books Gamma, Delta, and Epsilon*, Clarendon Press.

Kripke, S. A. (1959), ‘A Completeness Theorem in Modal Logic’, *Journal of Symbolic Logic* 24(1), 1--14.

Kripke, S. A. (1963a), ‘Semantical Analysis of Modal Logic I. Normal Propositional Calculi’, *Zeitschrift fur mathematische Logik und Grundlagen der Mathematik* 9(56), 67--96.

Kripke, S. A. (1963b), ‘Semantical Considerations on Modal Logic’, *Acta Philosophica Fennica* 16(1963), 83--94.

Kripke, S. A. (1977), ‘Identity and Necessity’, *in* Schwartz, S. P. (1977), (ed.), *Naming, Necessity, and Natural Kinds*, Cornell University Press, 66—101.

Kripke, S. A. (1980), *Naming and Necessity*, Harvard University Press.

Lewis, D. K. (1966), ‘An Argument for the Identity Theory’, *Journal of Philosophy* 63(1), 17.

Lewis, D. (1972), ‘Psychophysical and Theoretical Identifications’, *Australasian Journal of Philosophy* 50(December), 249--58.

Lewis, D. (1980a), ‘Mad Pain and Martian Pain’, *in* Ned Block, ed., ‘Readings in the Philosophy of Psychology’, Harvard University Press, , pp. 216--222.

Lewis, D. (1980b), ‘Review of Putnam’, *in* Ned Block, ed., ‘Readings in Philosophy of Psychology’, Cambridge: Harvard University Press, pp. 1--232.

Lewis, D. (1983), *Philosophical Papers Vol. I*, Oxford University Press.

Lewis, D. K. (1986), *On the Plurality of Worlds*, Blackwell Publishers.

Lewis, D. (1988), ‘Vague Identity: Evans Misunderstood’, *Analysis* 48(3), 128--130.

Lewis, D. (1990), ‘What Experience Teaches’, *in* William Lycan, ed., ‘Mind and Cognition: An Anthology’, Blackwell, pp. 29--57.

Lewis, D. (1994), ‘Reduction of Mind’, *in* Samuel Guttenplan, ed., ‘Companion to the Philosophy of Mind’, Blackwell, pp. 412--431.

Lewis, D. (1995), ‘Should a Materialist Believe in Qualia?’, *Australasian Journal of Philosophy* 73(1), 140--44.

Lewis, D. K. (1999), *Papers in Metaphysics and Epistemology*, Cambridge University Press.

Lycan, W. G. (1974), ‘Mental States and Putnam’s Functionalist Hypothesis’, *Australasian Journal of Philosophy* 52(May), 48--62.

Lycan, W. G. (1990), *Mind and Cognition: An Antholog*y, Wiley-Blackwell.

Malcolm, N. (1960) ‘Anselm’s Ontological Arguments’, *Philosophical Review* 69: 41–62.

Marton, P. (1998) ‘Zombies versus Materialists: The Battle for Conceivability’, *Southwest Philosophy Review* 14/1: 131–8.

McGarth, P.J. (1990) ‘The Refutation of the Ontological Argument’, *Philosophical Quarterly* 40/159: 195–212.

McGinn, C. (1977), ‘Anomalous Monism and Kripke’s Cartesian Intuitions’, *Analysis* 2(January), 78--80.

McLaughlin, B. P. & Walter, S. (2006), *Oxford Handbook to the Philosophy of Mind*, Oxford University Press.

Merricks, T. (2001), ‘Varieties of Vagueness’, *Philosophy and Phenomenological Research* 62(1), 145--157.

Mucciolo, L. F. (1975), ‘On Kripke’s Argument Against the Identity Thesis’, *Philosophia* 5(October), 499--506.

Murdoch, D. (1993), ‘Exclusion and Abstraction in Descartes’ Metaphysics’, *Philosophical Quarterly* 44(170), 38--57.

Murdoch, D. (1999), ‘The Cartesian Circle’, *Philosophical Review* 108(2), 221--244.

Nagel, T. (1986), *The View From Nowhere*, Oxford University Press.

Nicholls, J. et al. (2001), ed., *From Neuron to Brain*, MA: Sinauer Associates.

Noonan, H. W. (1982), ‘Vague Objects’, *Analysis* 42(1), 3--6.

Noonan, H. W. (1984), ‘Indefinite Identity: A Reply to Broome’, *Analysis* 44(3), 117--121.

Noonan, H. W. (1990), ‘Vague Identity Yet Again’, *Analysis* 50(3), 157--162.

Noonan, H. W. (1991), ‘Indeterminate Identity, Contingent Identity and Abelardian Predicates’, *Philosophical Quarterly* 41(163), 183--193.

Papineau, D. (2001), *Thinking About Consciousness*, Oxford University Press.

Papineau, D. (2007), ‘Kripke’s Proof is Ad Hominem Not Two-Dimensional’, *Philosophical Perspectives* 21(1), 475--494.

Parsons, T. & Woodruff, P. (1995), ‘Worldly Indeterminacy of Identity’, *Proceedings of the Aristotelian Society* 95(n/a), 171--191.

Parsons, T. (2000), *Indeterminate Identity: Metaphysics and Semantics*, Clarendon Press.

Place, U. T. (1956), ‘Is Consciousness a Brain Process?’, *British Journal of Psychology* 47(1), 44--50.

Plantinga, A. (1967), ‘Comments’, *in* Hector-Neri Castaneda, ed., *Intentionality, Minds and Perception*, Wayne State University Press, 201--5.
Plantinga, A. (1974), *The Nature of Necessity*. Oxford: Oxford University Press.

Polger, T. W. (2002), ‘Putnam’s Intuition’, *Philosophical Studies* 109(2), 143--70.

Puccetti, R. (1977), ‘The Great C-Fiber Myth: A Critical Note’, *Philosophy of Science* 44(June), 303--305.

Putnam, H. (1967a), ‘The Mental Life of Some Machines’, *in* Hector-Neri Castaneda, ed., *Intentionality, Minds and Perception*, Wayne State University Press, 177--200.

Putnam, H. (1967b), ‘Psychological Predicates’, *in* W. H. Capitan & D. D. Merrill, ed., *Art, Mind, and Religion*, University of Pittsburgh Press, 37--48.

Putnam, H. (1967c), ‘Rejoinder’, *in* Hector-Neri Castaneda, ed., *Intentionality, Minds and Perception*, Wayne State University Press, 206--213.

Putnam, H. (1973), ‘Meaning and Reference’, *Journal of Philosophy* 70(19), 699--711.

Putnam, H. (1975a), *Mind, Language, and Reality*, Cambridge University Press.

Putnam, H. (1975b), ‘Minds and Machines’, *in Mind, Language, and Reality*, Cambridge University Press, 362--85.

Putnam, H. (1975c), ‘Philosophy and Our Mental Life’ *in* ‘Mind, Language, and Reality’, Cambridge University Press, 291--303.

Putnam, H. (1988), *Representation and Reality*, MIT Press.

Quine, W. V. O. (1951), 'Two Dogmas of Empiricism', *Philosophical Review* 60(1), 20--43.

Rey, G. (1997), *Contemporary Philosophy of Mind: A Contentiously Classical Approach*, Blackwell.

Robinson, H. (2012) 'Dualism', in Zalta E.N. (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2012.). URL http://plato.stanford.edu/archives/win2012/entries/dualism/

Rocca, M. D. (1993), 'Kripke's Essentialist Argument Against the Identity Theory', *Philosophical Studies* 69(1), 101--112.

Rorty, R. (1965), 'Mind-Body Identity, Privacy, and Categories', *Review of Metaphysics* 19(September), 24--54.

Russell, B. (1985), *The Philosophy of Logical Atomism*, Open Court.

Sainsbury, R. M. (1995), *Paradoxes*, Cambridge University Press.

Sainsbury, R. M. (1995), 'Why the World Cannot Be Vague', *Southern Journal of Philosophy* 33(S1), 63--81.

Salmon, N. (1981), *Reference and Essence*, New York: Prometheus Books.

Schwartz, S. P. (1977), (ed.), *Naming, Necessity, and Natural Kinds*, Cornell University Press.

Searle, J. R. (1992), *The Rediscovery of the Mind*, MIT Press.

Shaffer, J. A. (1961), 'Could Mental States Be Brain Processes?', *Journal of Philosophy* 58(December), 813--22.

Shapiro, L. A. (2004), *The Mind Incarnate*, MIT Press.

Sher, G. A. (1977), 'Kripke, Cartesian Intuitions, and Materialism', *Canadian Journal of Philosophy* 7(2), 227--38.

Shoemaker, S. (1975), 'Functionalism and Qualia', *Philosophical Studies* 27(May), 291--315.

Shoemaker, S. (1981), 'Some Varieties of Functionalism', *Philosophical Topics* 12(1), 93--119.

Sturgeon, S. (2000) *Matters of Mind: Consciousness, Reason and Nature*. London: Routledge.

Soames, S. (2003), *Philosophical Analysis in the Twentieth Century Vol. 1*, Princeton University Press.

Soames, S. (2005), *Reference and Description: The Case Against Two-Dimensionalism*, Princeton: Princeton University Press.

Soames, S. (2006), Kripke, the Necessary a Posteriori, and the Two-Dimensionalist Heresy, *in* Garc, ed., 'Two-Dimensional Semantics', Oxford: Clarendon Press, pp. 272--292.

Thomason, R. H. (1982), 'Identity and Vagueness', *Philosophical Studies* 42(3), 329--332.

Turing, A. (1936), 'On Computable Numbers, with an Application to the Entscheidungsproblem', *Proceedings of the London Mathematical Society* 42(1), 230--265.

Tye, M. (1990), 'Vague Objects', *Mind* 99(396), 535--557.

Tye, M. (1994), 'Sorites Paradoxes and the Semantics of Vagueness', *Philosophical Perspectives* 8, 189--206.

Tye, M. (2000), 'Vagueness and Reality', *Philosophical Topics* 28(1), 195--210.

Wiggins, D. (1986), On Singling Out an Object Determinately, *in* P. Pettit & J. McDowell, ed., 'Subject, Thought and Context', Clarendon Press.

Williamson, T. (1994), *Vagueness*, Routledge.

Yablo, S. (1993), 'Is Conceivability a Guide to Possibility?', *Philosophy and Phenomenological Research* 53(1), 1--42.

Yablo, S. (1999) 'Concepts and Consciousness', *Philosophy and Phenomenological Research* 59/2: 455–63.

Yablo, S. (2000), 'Textbook Kripkeanism and the Open Texture of Concepts', *Pacific Philosophical Quarterly* 81(1), 98--122.

Yablo, S. (2002), Coulda, Woulda, Shoulda, *in* Tamar S. Gendler & John Hawthorne, ed., *Conceivability and Possibility*, Oxford University Press, pp. 441--492.

Zadeh, L. A. (1975), 'Fuzzy Logic and Approximate Reasoning', *Synthese* 30(3-4), 407--428.

Zemach, E. M. (1991), 'Vague Objects', *Noûs* 25(3), 323--340.

Zemach, E. M. (1994), ‘Identity and Epistemic Counterparts’, *Philosophia* 23(1-4), 265--270.

Zupanc, G. (2010), *Behavioral Neurobiology*, Oxford University Press.