The Consciousness Test

John Thornton

The School of Philosophy, Australian National University, Canberra and
The Institute for Integrated and Intelligent Systems, Griffith University, Brisbane
j.thornton@griffith.edu.au

In 1994, Todd Moody[1] argued that a zombie community would not develop in the same way as a human community, because their lack of consciousness would subtly alter their speech behaviour – particularly their philosophical talk of phenomenal consciousness. Subsequently, even those who accept the conceptual possibility of zombies have baulked at the idea that zombies could behave differently to their human counterparts. One reason must be that to accept such a difference comes at a high price: the denial of the causal closure of the physical. In particular, David Chalmers[2] has given a detailed account of how this 'paradox of phenomenal judgment' can be answered. For Chalmers, the paradox is how a human being, whose behaviour is entirely determined by interactions between microphysical entities that obey physical laws, can come to make correct judgments about phenomenal experiences. His answer is given in a detailed analysis of the formation of direct phenomenal beliefs. This analysis explains how, although we form direct phenomenal concepts that possess genuine phenomenal content, the role of such concepts in the formation of beliefs and the subsequent production of behaviour can be explained in entirely functional, and hence physical terms.

In this paper I shall argue that even if we accept Chalmers' account, there are still capacities of human concept formation and judgment that remain unexplained. For example, consider Moody's community of zombies, and the situation of their never having had any direct or indirect contact with any conscious entity. I suggest that one concept such a community would lack, would be the concept of a *zombie,* i.e. the concept of an entity physically identical to themselves that lacks phenomenal consciousness. I shall argue that the reason for this lack is that zombies, by definition, lack the *direct knowledge of being conscious* required to form such a concept.

1 The Consciousness Test

To make the issues clearer, consider the following thought experiment: In a nearby world, there exists a cloning machine that produces molecularly type-identical copies of any object placed in its cloning space. It works by scanning the originals and assembling the copies out of a supply of inert chemical compounds. Unknown to the machine's inventors, for an organism to be conscious it must be composed of *living* cells. And (in this world) cells are only alive if they are causally connected to, and materially constituted by, a process of cell division involving previously existing living cells. As the cloned cells are assembled out of non-living material, they are not alive and are therefore unable to support consciousness. So, when the machine is presented with a conscious living organism it will produce an unconscious physical replica, or, as we would say, a *zombie.*

Until the invention of the machine, this world was indiscernible from our own both in terms of the physical evolution of events and in terms of the consciousness of the organisms experiencing those events. Once the machine was assembled, and the first humans were cloned, the non-cloned humans devised a *consciousness test* to decide if the cloned humans were phenomenally conscious. The test requires a non-cloned human subject with the following characteristics:

1. Complete ignorance of the phenomenal concepts to be grasped in the test.
2. Complete amnesia in relation to any previous experience of phenomenal consciousness.

These stipulations ensure that a cloned subject does not answer questions on the basis of previously learnt concepts or on the basis of comparison with previous experience. At the same time, the ignorance and amnesia are not so severe as to affect the subject's ordinary discourse involving non-phenomenal concepts and memories. For the purposes of the test the precise distinction between what counts as a phenomenal concept and a non-phenomenal concept is not crucial. What is required is that the clone is materialised in the same epistemic situation as a zombie from the isolated zombie community mentioned in the introduction.

The first subject of the test, called Eric, is placed in a hermetically sealed and sound-proofed room with internal oxygen and power supplies. Eric is seated on a chair in front of a computer running a consciousness test program, and the entire room is located within the cloning space of the cloning machine. At the appointed moment, the cloning machine is activated. Simultaneously, a second device effects Eric's amnesia, and an instant later a cloned room is materialised containing Eric's cloned twin. At this point the two rooms are molecule-for-molecule type-identical and sealed from any further external influences.

The consciousness test program, called Alan, is a Turing Test certified dialogue program, such that most human experts are unable to reliably detect that Alan is not human, even after an hour of textual interaction. Alan's task is to answer the participant's questions about where they are, and what they are doing, and then lead them into the test. The following is a transcript of the first part of the test during which the dialogue in the two rooms remained indistinguishable:

Alan: 'Do you see the coloured patch on the screen?'
Eric: 'Yes'
Alan: 'What colour is it?'
Eric: 'Green'

Alan: 'Do you agree that any normal conscious person looking at that patch would also call it green?'

Eric: 'Yes'

Alan: 'Do you agree that there is something going on in you when you look at the green patch that we could call your conscious experience of the colour of green?'

Eric: 'I'm not sure, I haven't thought about it before'

Alan: 'Well, do you agree that there are photons being emitted from the screen that are hitting your retina?'

Eric: 'Yes'

Alan: 'Do you think those photons are coloured green?'

Eric: 'Well, no. I see what you mean. If the photons aren't coloured green then neither is the patch on the screen, at least if I consider it independently of my looking at it'

Alan: 'Yes'

Eric: 'So the patch on the screen only *looks* green to me'

Alan: 'Exactly'

Eric: 'And my conscious experience of the colour of green is somehow caused by the photons hitting my retina and then stimulating my brain'

Alan: 'Precisely'

Eric: 'Gosh'

Alan: 'Now, given your understanding of the conscious experience of the colour of green, can you conceive it possible that someone else could have a *different* conscious experience of the colour of green, say the kind of conscious experience you have when you see a red patch, even though they would still call the same things green as you do?'

Eric: 'Well yes, that does seem possible. That person could have something different going on in their brain.'

Alan: 'Yes, but think more carefully. What if *exactly the same things* were going on in that person's brain as are going on in yours. Suppose there was an exact clone of you, sitting in exactly the same situation as you are in now. Can you conceive it possible that your clone could have a different conscious experience of the colour of green to the one that you are having now? Or even that he could be having *no conscious experience whatsoever*?...'

The entire consciousness test thought experiment hinges on this moment, when each considers the situation of his identical twin. The question is: is it possible for conscious Eric (Eric$_c$) to respond differently to zombie Eric (Eric$_z$)? In the rest of the paper, I will argue that it is.

Clearly the consequences of accepting such an argument are significant. For it is built-in to the experiment that there are no relevant initial physical differences between the situations in the two rooms. The *only* difference that could make a difference is that Eric$_c$ is conscious and Eric$_z$ is not. So if Eric$_c$ responds differently it must be because he is conscious, and the fact of his responding differently entails that his possession of consciousness is, in some unspecified way, downwardly causally effective on the physical. If this is the case, then the causal closure of the microphysical is false.

The argument I shall be defending is that Eric$_c$ is able to conceive the possibility that his twin clone has a different phenomenal experience to himself because his being conscious enables him to conceive the possibility that his phenomenal experience is not determined by physical relations. Conversely, Eric$_z$ is unable to conceive the possibility that his twin clone has a different phenomenal experience to himself because his *not* being conscious entails that he is unable to conceive the possibility of his phenomenal experience being determined by anything *other than* physical relations.

This reduces to the claim is that consciousness is a difference-making cause of the ability to conceive the possibility that phenomenal experience is not determined by physical relations. I will argue that this claim can be justified on the basis of a priori reasoning in combination with phenomenological reflection. The role of this reflection is to confirm that being conscious, in and of itself, is constitutive of a direct knowledge of being conscious. I shall then argue that this direct knowledge enables us, as conscious individuals, to conceive the possibility that our phenomenal experience is not determined by physical relations. Conversely, I shall argue that an unconscious entity, such as Eric$_z$, although able to *mimic* the possession of such a concept, will not be able to *discover* the concept in an environment where no pre-existing example of the concept is accessible.

2 Phenomenal Judgment

Phenomenal judgments are judgments about states of consciousness. We express such judgments whenever we talk of conscious experience *as* experience, rather than talking of the objects and states of affairs that we encounter on the basis of such experience. If we accept that consciousness is something ‘over and above’ the physical, that cannot be reduced to or explained in terms of physical or functional concepts, then the question naturally arises as to whether consciousness has any *independent* causal efficacy. That is, whether consciousness makes any material difference to my behaviour, rather than simply supervening on physical processes that themselves are entirely determined by the operation of physical law. It is here that the paradox of phenomenal judgment arises. For, if we accept the causal closure of the physical, not just as a useful assumption for the purposes of the development of the physical sciences, but as a universal principle, then it becomes difficult to understand how physical brain processes could ever come to express phenomenal judgments about conscious states. For causal closure entails that those conscious states have no way of physically influencing the brain states that are forming correct judgments about those very states.

It is this paradox that Chalmers sets out to address with his account of direct phenomenal belief. Here he defends the irreducibility of consciousness *and* the causal closure of the physical. The basis of his strategy is to argue that the brain is physically structured in such a way that physical law alone ensures that our speech matches the situation of our phenomenal experience. For such an argument to succeed requires, firstly, that all the phenomenal experiences about which we judge have suitable corresponding physical conceptual instantiations; secondly, that the brain is able to form beliefs on the basis of these concepts; and thirdly that the brain is able to correctly reason about such beliefs.

Of these requirements, it is the first that presents the greatest problem in relation to maintaining the irreducibility of consciousness: for how can a physically instantiated concept capture a phenomenal quality that cannot be physically reduced? Chalmers' answer is worth going into, as the core of our argument rests on the claim that there is an implicit knowing of self-consciousness that cannot be represented conceptually (and so cannot feature in a physical explanation of behaviour) and yet that can still form the basis of a phenomenal judgment.

2.1 Pure Phenomenal Concepts

Chalmers introduces a number of important conceptual distinctions in relation to phenomenal colour perception that can be illustrated via the consciousness test. These phenomenal concepts are distinct from the everyday *non*-phenomenal concepts of colour, such as the concept that picks out green as a property of an external object (e.g. when Eric first identifies the patch on the screen as being green). The use of phenomenal concepts presupposes a reflective stance that distinguishes between experience *as* experience and experience as an experience of things and states of affairs in the world. From such a stance, green can be seen as both a property of an object and as a phenomenological property or quality of the experience itself. That is not to say that the *experience* is green, it is rather to say that I can have an experience *of* green by bringing the phenomenal quality of green to explicit consciousness as the content of a suitable phenomenal concept. In intentional terms, we could say that a phenomenal colour concept presents a colour under a different aspect, *as* an experience of a colour rather than *as* the colour of an object, or that the quality of the colour is brought to the foreground while the object's being coloured recedes.

Eric first grasps a phenomenal concept of green after considering whether photons themselves are coloured. This leads him to distinguish between the patch of colour on the screen conceived as something emitting colourless photons and his phenomenal experience of the colour of green. Chalmers terms this an *individual relational concept* of green or $green_I$, i.e. the "phenomenal quality typically caused in me by paradigmatically [green] things."[3] Chalmers further distinguishes a *community relational concept* of green or $green_C$, i.e. the "phenomenal quality typically caused in normal subjects within my community by paradigmatically [green] things."[4] This is the concept Eric uses when he considers whether someone else could have a different experience of green to the one he is having.

Both *green*$_I$ and *green*$_C$ are termed *relational* because, although their content (G) is phenomenal, G is picked out by means of referring to an external object, e.g. the green patch on the screen. Chalmers goes on to distinguish a third *phenomenal demonstrative concept* (D) which Eric can employ while looking at his experience of the patch on the screen and thinking 'Does Alan mean *this* colour experience?' As with *green*$_I$ and *green*$_C$, D is *relational,* but in this case the reference is fixed in relation to the act of ostending.

Chalmers' major step is to recognise a fourth *non*-relational *pure phenomenal concept green*$_P$ that picks out phenomenal greenness directly "in terms of its intrinsic phenomenal nature."[5] This concept has the distinguishing feature that its epistemic intension is fixed in all possible worlds, whereas the epistemic intensions of the relational phenomenal concepts can vary. So, for example, if we allow that a physically identical but phenomenally spectrum-inverted world is epistemically possible, then the inverted-community, including my inverted-twin, will experience what-I-call and what-we-call red in exactly those situations where we experience what-I-call and what-we-call green. This means our three relational phenomenal concepts will have different intensions according to which world we currently treat as actual. In contrast, the intension of the pure phenomenal concept, *green*$_P$, picks out the intrinsic phenomenal quality of green without reference to any act of ostension or to any paradigmatic objects. Therefore it picks out green itself, without reference to the world in which the distinction is made. In fact, without such a concept, it would not be possible to talk about inverted worlds, because such talk presupposes there is some constant quality by means of which we can differentiate between the worlds.

2.2 Direct Phenomenal Concepts and Direct Phenomenal Beliefs

On the basis of this notion of a pure phenomenal concept, Chalmers introduces the notion of a *direct phenomenal concept.* Direct phenomenal concepts are formed while I am actually experiencing the quality to which a pure phenomenal concept refers in such a way that the experienced quality is explicitly taken up as the *content* of a direct phenomenal concept. Such a direct concept is only present for as long as I am correctly demonstratively attending to the experience which is constitutive of its content. This correct attentive taking up of the phenomenal quality into the content of the concept can be negatively characterised as not imposing a pre-existing concept that fails to match the experience in some way. That is not to say that a direct phenomenal concept has to perfectly capture all the hues and variations of a particular colour experience, it is rather to say that it must agree with the experience at whatever level of detail the concept itself specifies.

A direct phenomenal concept can then become the basis of a *direct phenomenal belief.* Such beliefs are formed "when a subject predicates the concept of the very experience responsible for constituting its content."[6] So, during the event of forming a direct phenomenal concept, a subject can explicitly form the direct phenomenal belief that 'this quality is x' where x is the direct phenomenal concept that is currently being demonstrated. For example, $D = green_P$ represents Eric's direct phenomenal belief that the quality he is demonstrating by looking at the greenness of the green patch on the screen, is $green_P$.

Chalmers' uses this account of direct phenomenal concepts and direct phenomenal beliefs to answer the paradox of phenomenal judgment, i.e. the puzzle of how, given the causal closure of the physical, we can form causally effective phenomenal judgments about our phenomenal experiences. Direct phenomenal beliefs escape the paradox because they are formed on the basis of *constitution* and not on the basis of causal connection. The crucial point is that the content of a direct phenomenal concept is constituted by the phenomenal experience, i.e. the quality of greenness is actually present in the experience, it does not *cause* the content of the concept, it *is* the content of the concept. The direct phenomenal belief in turn is a belief about the actual experience that is demonstrating the concept, and this experience is also an immediately given constitutive element of the belief. The belief does not require a causal connection with anything outside of itself, because the elements that constitute the belief, the direct phenomenal concept and the demonstrative act, are immediately co-present in the forming and sustaining of the belief.

At the same time, direct phenomenal concepts and direct phenomenal beliefs can be understood as having physical instantiations in the brain, and functional roles in the production of appropriate speech acts expressing phenomenal judgments. In this way, the causal closure of the physical is made consistent with the existence of irreducible phenomenal properties and the paradox of phenomenal judgment is resolved.

3 Direct Knowledge of Consciousness

For the purposes of the rest of the paper, I will accept Chalmers' claim that phenomenal experience can be taken up into direct phenomenal concepts and that direct phenomenal beliefs are constituted rather than caused by experience. This entails accepting that both Eric$_c$ and Eric$_z$ will be indiscernible in terms of their utterances concerning direct phenomenal beliefs. However, the central argument of the paper, viz. that consciousness is a difference-making cause of the ability to conceive the possibility that phenomenal experience is not determined by physical relations, is not affected by Chalmers' account of phenomenal judgment. The issue is not whether Eric$_c$ and Eric$_z$ will form the same beliefs concerning the phenomenal content of their experiences, but whether they will form the same beliefs concerning consciousness itself, considered independently of particular acts that demonstrate phenomenal experiences. The central claim is that being conscious is constituted by a direct, non-conceptual knowledge of being conscious, and that this knowledge is the difference that makes a difference in the consciousness test. I will defend this claim and its consequences via the elucidation of the following *reductio*:

C1) Consciousness makes no difference to the physical operation of the brain.
C2) If there is direct knowledge of being conscious, this knowledge cannot be explained in physical terms.
C3) There is direct knowledge of being conscious that is constitutive of being conscious.
C4) C3 can be justified by the immediate demonstration of being conscious.
C5) From C3 and C4, I can infer that I am not not conscious [now].
C6) From C1 and C5, I can conceive the epistemic possibility of zombies.
C7) From C5 and C6, I can infer that I am not a zombie.
C8) From C1, C2 and C3, a zombie does not have direct knowledge of being conscious.
C9) From C8, a zombie cannot conceive the possibility of being conscious.
C10) From C8 and C9, a zombie cannot conceive the possibility of not being conscious.
C11) From C9 and C10, a zombie cannot distinguish between a conscious individual and a non-conscious duplicate.
C12) From C7 and C11, consciousness makes a difference to the physical operation of the brain.

C1) Consciousness makes no difference to the physical operation of the brain: I assume C1 for the purposes of the *reductio*.

C2) Direct knowledge of being conscious cannot be explained in physical terms: C2 introduces the notion of a *direct* knowledge of being conscious, where the directness of the knowledge has the same *non-relational* and *non-inferential* sense as it does for a direct phenomenal belief or a direct phenomenal concept. So, a direct knowledge of being conscious is a knowledge of being conscious that is unmediated and accessible purely on the basis of being conscious. This will be clarified in relation to C3.

C2 follows from C1, for, if I have direct knowledge of being conscious, this knowledge cannot be explained as the taking up of physically instantiated content into a physically instantiated direct phenomenal concept, because, from C1, the presence of consciousness can make no difference to the physical operation of the brain. Consequently, there is no physical basis upon which such a concept can form. This contrasts with the forming a direct phenomenal belief, where the presence of absence of consciousness makes no difference to the physical story. For instance, both Eric$_c$ and Eric$_z$ are able to form physical representations of direct phenomenal concepts of phenomenal qualities because the presence of a phenomenal quality has a physical correlate in the brain. We can illustrate this by conceiving of the phenomenal quality of a particular shade of green as corresponding to a pattern of activity in the visual cortex, which itself is connected to another pattern of activity in the neocortex representing a direct phenomenal concept of that quality. But, in the case of Eric$_c$'s being conscious in the sense that Eric$_z$ is *not* conscious, there can be no such distinguishing pattern of physical activity, because we have already assumed that the physical processes occurring in their two brains are identical (at least to begin with).

C3) There is direct knowledge of being conscious that is constitutive of being conscious: I will take it that C2 is fairly uncontroversial. The main issue is C3, the possibility of a direct knowledge of consciousness. The problem is this: we have already assumed that the presence or absence of consciousness will make no physical difference to the operation of the brain. Therefore there can be no physical representation of a *direct* concept of being conscious. The only kinds of concepts that are possible are indirect or relational, such as the concept of being conscious in relation to being awake, or thinking 'I am conscious,' or understanding and conversing appropriately in a human language. The archetypal sign of consciousness, viz. the having of phenomenal experiences, still picks out being conscious in relation to the phenomenal qualities that are experienced, i.e. one is transitively conscious-*of* the phenomenal quality, one is not conscious of consciousness itself. As Moore famously said:

> . . . the moment we try to fix our attention upon consciousness and to see what, distinctly, it is, it seems to vanish: it seems as if we had before us a mere emptiness. When we try to introspect the sensation of blue, all we can see is the blue: the other element is as if it were diaphanous.[7]

However, on our account, it is a *necessary feature* of consciousness that it is diaphanous in the precise sense that it cannot be directly captured in any physically instantiated concept structure. Therefore, if a direct knowledge of consciousness is possible, it must not only be non-relational and non-inferential, but also *non-conceptual.* This non-conceptual character is distinct from the notion of non-conceptual phenomenal content, because, on Chalmers' account, it must always be *possible* to form a direct phenomenal concept that can take up such phenomenal content. The content is therefore non-conceptual in the sense that it is not currently conceptualised, not in the sense that it *cannot* be conceptualised. In contrast, the claim here is that consciousness *cannot* be directly conceptualised, because a concept requires a physical instantiation in order to function as a concept.

Chalmers' account of direct phenomenal belief also has the consequence that consciousness, in being non-conceptual, cannot be represented and therefore cannot become an intentional object for an intentional consciousness. Here we consider intentional consciousness to be any consciousness that has a definite object, i.e. content that can be represented conceptually. As we have already accepted that all phenomenal content can become the content of a pure phenomenal concept, it follows that all *transitive* consciousness is conceptual or intentional in this sense, and therefore that all transitive *consciousness-of such-and-such* is intentional. So, on the basis of Chalmers' nonreductive representationalism,[8] we can conclude that there can be no transitive consciousness-of consciousness.

According to the account so far, if a direct knowledge of consciousness is possible, then such knowledge must be non-relational, non-inferential, non-conceptual, non-intentional, non-transitive and immediately accessible. Therefore, any attempt to demonstrate consciousness as if it were a kind of phenomenal content should fail. For example, consider the demonstrative concept D_C of the form '*This* consciousness.' Such a concept has a slot awaiting the demonstrated phenomenal content. But there is no such content that can fill the slot, so the demonstrative intention falls back on itself and demonstrates nothing but itself. In falling back, the demonstration is an *unfulfilled* act of transitive-intentional consciousness, an act that failed to reach an object, and so failed in its transitive intention, and so failed to be transitive. But the act is still *conscious* in its failure to grasp its object within a concept. It is this consciousness of failure, of having not reached to any object that is of interest.

More generally, we can observe that every act of transitive consciousness has an accompanying consciousness that *the act is the act that it is.* This can be expressed as follows: let E be a transitive consciousness of the form, *conscious-of such-and-such.* Now observe that in any act of transitive consciousness, I am also conscious that there is such an act. We can express this non-transitive consciousness as the *consciousness-that* $E = E$, i.e. the consciousness that there is *identity.* This consciousness introduces no new content. Neither is it a consciousness-*of* an identity. It is rather the consciousness-*that* the transitive act *is happening.*

The claim is that the accompanying *consciousness-that* $E = E$ *is* my direct, non-conceptual knowledge of consciousness, and that this is the immediate and only justification necessary for any explicit assertion to the effect that 'I am conscious now.' This claim currently rests on three pieces of evidence: Firstly, that the consciousness-that $E = E$ has the correct *form,* i.e. it is non-relational, non-inferential, non-conceptual, non-intentional, non-transitive and immediately accessible. Secondly, that it is possible, through phenomenological reflection, to distinguish, within a transitive, intentional consciousness-of x, a consciousness-that there is a consciousness-of x. And thirdly, that you can immediately demonstrate a *direct* knowledge that you are conscious [now] by observing that you know that you are conscious [now] without referring to anything outside your immediate experience and without being able to demonstrate or represent any experiential or phenomenal content that corresponds to your being conscious.

The basic form of a consciousness-that is as follows:

> conscious-that ((conscious-of (such-and-such)) = (conscious-of (such-and-such)))
> which can be simplified to:
> conscious-that [conscious-of (such-and-such)]

Here the square and round brackets distinguish that the operation of a *consciousness-that* is *not* the same as the operation of a *consciousness-of* such-and-such. Neither is the *that* of a consciousness-that a form of assertion, although it can be the *justification* for an assertion. Also, although a consciousness-that can been represented as the consciousness-that there is identity, such a conceptual representation is indirect and so cannot be fulfilled by any experience of a consciousness-that. To be precise, a consciousness-that there is identity does *not* hold two objects and compare them, that would be a consciousness-of identity. Rather, a consciousness-that is a *direct knowledge* of identity. This is the knowledge that any given consciousness-of such-and-such is a consciousness-of such-and-such.

C4) Direct knowledge of being conscious is justified by the immediate demonstration of being conscious: Consciousness-that is justified knowledge because the knowledge claim: *that this experience is this experience* is immediately given as the non-conceptual content of the experience. This knowledge is an *unmediated* or direct knowledge of being conscious in the strong sense that it does not refer to any content, concept or belief that lies either inside or outside an experience. It is rather *constitutive* of being conscious to have direct knowledge of being conscious, just as phenomenal content is constitutive of a pure phenomenal concept. In both cases, direct evidence of constitution can only be given via the conscious experience of such constitution, as neither the direct knowledge of being conscious nor the consciousness-that of a consciousness-of phenomenal content, has any physical instantiation.

The justificatory role of consciousness-that is also present in the notion of justification by acquaintance. In Chalmers' discussion of the justification of direct phenomenal belief he reintroduces the notion of acquaintance first developed by Russell.[9] On this account, acquaintance is a relation between a subject and a phenomenal property, such that the subject is directly aware of that property. Chalmers' uses this notion as an epistemic justification of pure phenomenal beliefs. He also notes that:

> Some philosophers ... have held that we are "acquainted with acquaintance", and have made the case of its existence that way. I think there is something to the idea that our special epistemic relation to experience is revealed in our experience, but I note that the proponent of acquaintance is not forced to rely on such a thesis.

So Chalmers assumes that there is such a thing as acquaintance, but only allows it to provide epistemic justification for direct phenomenal beliefs, because such beliefs involve a direct awareness of *non-relational* content. Here the question is what justifies acquaintance? That is, on what basis do I know that I am acquainted with phenomenal content? On the current account, acquaintance is a consciousness-that there is consciousness-of phenomenal content. As such, a direct phenomenal belief is *negatively* justified by the fact that the content of the belief does *not* refer outside the act of forming the belief. It therefore does not inherit the doubt that accrues to external reference. However, the *positive* justification of acquaintance is that a *subject* is acquainted with the phenomenal content of the belief, i.e. that there is some property in the relation of acquaintance that guarantees its veridicality. Chalmers signals that this justification could be that we are "acquainted with acquaintance" but does not explore further. This leaves us with an unanalysed subject and the threat of recursion, i.e. that I can be acquainted with being acquainted with acquaintance, and so on.

In the current account, I have rejected a subject-based ontology of experience for the Humean reason that no such content can identified within conscious experience. By staying with the phenomenological evidence, we can identify a consciousness-of such-and-such and a consciousness-that there is a consciousness-of such-and-such. On this basis, by taking a subject to be a conceptual *objectification* of a consciousness-that we can translate the acquaintance relation into terms that directly express the operations of consciousness, as follows:

subject acquainted-with content → conscious-that [conscious-of (content)]
where subject = conscious-that and acquainted-with = conscious-of

Here it is the consciousness of the subject, the consciousness-that, that provides the justification of the acquaintance relation, where the justification is the identity that the content is the content that it is. In this translation, consciousness-that is the ground of acquaintance, or that on the basis of which we not only have epistemic certainty of being conscious but also the ground on which we have epistemic certainty of the phenomenal content of a pure phenomenal belief.

The structure of conscious-that[conscious-of(such-and-such)] has obvious parallels with Sartre's account of pre-reflective self-consciousness. In Sartrean terms, a consciousness-that expresses the implicit self-givenness or 'for-itself' of consciousness, the 'only mode of existence that is possible for a consciousness of something.'[10] This is in contradistinction to higher-order accounts of consciousness,[11] that take self-consciousness to be an extrinsic property of mental states, i.e. an intransitive consciousness that occurs on the basis of a first-order mental state becoming an object for an external higher order mental process. Consciousness-that cannot be understood in these terms because it is both non-conceptual, non-objective and therefore cannot be known as an object of explicit conceptual reflection. That is not to say a consciousness-that cannot be understood in terms of a reflection. But it is an immediate reflection that reveals the identity of an experience with itself. It is this primal reflection which forms the basis for the further conceptual reflection of a consciousness-of. Such higher-order reflection takes a consciousness-of as its object and objectifies it as an experience, rather than taking it directly as a consciousness-of the world.

For example, in objectifying a consciousness-of a tomato as an *experience-of* a consciousness-of a tomato, I can conceptualise the phenomenal quality of the experience and form a direct phenomenal belief to the effect that '*this* quality is red_P. The structure of this act of reflection can be expressed as:

conscious-that[conscious-of(experience-of(phenomenal red))]

Here the experience-of has become the intentional object of a reflective consciousness-of. Such reflection represents experience as experience, rather than representing the objects of experience directly. The three-level structure expresses the reflection of a higher order intention on a lower level intention, and shows the distinction between such explicit reflection or self-consciousness and the implicit or immediate self-consciousness of a consciousness-that. The structure also shows that there is no problem with a recursion of consciousness, i.e. a normal state of consciousness is always a consciousness-that I am conscious-of such-and-such, where, if the content of such-and-such is intended as a conscious state, it can only be an objectified conscious state, something represented (i.e. as an experience), and not an original consciousness, as, if our reasoning is correct, an original consciousness cannot be conceptualised. So there cannot be a consciousness-of a consciousness-of such-and-such because a consciousness-of is a consciousness-of a such-and-such that can be conceptualised and, once a consciousness-of is conceptualised, it is no longer a consciousness-of but an *experience-of such-and-such.* The same argument applies to a consciousness-of a consciousness-that.

C5) From C3 and C4, I can infer that I am not not conscious [now]: As Nagel said 'an organism has conscious mental states if and only if there is something that it is like to *be* that organism – something it is like *for* the organism.'[12] Consciousness-that provides a more precise characterisation of the 'what it is like' of first-person consciousness. For consciousness-that is not 'like' anything – knowing what it is like is the knowledge that there is *identity,* not likeness. We can characterise this identity as a reflection of a consciousness-of on itself, with the proviso that it is the reflection – the consciousness-that – that brings the consciousness-of into being. The 'of' here expresses that there is a gap or opening of reflection within which such-and-such *becomes* conscious as a phenomenal-intentional representation.

Direct knowledge of being conscious is therefore not a formal or empty notion, it contains the knowledge of *what* it is to be conscious, not just the knowledge-*that* there is consciousness. It this *substantive* knowledge of consciousness that comes closest to Nagel's idea of 'what it is like.' For a consciousness-that is not abstract, it inhabits the reflection within which a consciousness-of is known, and constitutes what we conceptualise as the first-person perspective.

As it is a moment of the direct knowledge of being conscious, this substantive aspect of knowing what it is to be conscious cannot be directly conceptualised. However, it does have epistemic consequences: in substantively knowing what it is to be conscious, *I know that being conscious is something rather than nothing.* Put another way, in knowing what being conscious is, I can form the concept of something's *not* being conscious, in the precise sense that something's not being conscious would make a difference to it – it would lack my substantive knowledge of being conscious.

The point is not whether there is a metaphysical possibility that there could be non-conscious things. The point is that my being conscious allows me to conceive the epistemic possibility that there could be non-conscious things. This is a simple point of logic: *only on the basis of my knowing what x is and that x exists can I conceive of x not being a property of y.* For example, I cannot conceive of there not being bears in the forest if I have never heard of such things as bears.

Putting these points together: (i) as I am able to conceive the possibility of things that are not conscious, and (ii) as, from C3, I have direct knowledge of being conscious, and (iii) as, from C4, I can immediately demonstrate direct knowledge that I am conscious [now], it follows that (iv) I am not not conscious [now].

C6) From C1 and C5, I can conceive the epistemic possibility of zombies: If consciousness makes no difference to the physical operation of the brain (as assumed in C1) and as it is epistemically possible to conceive of things that are not conscious (as argued in C5), it follows that it is epistemically possible to conceive of a physical duplicate of my brain and body that lacks consciousness (as we do in the consciousness test). Therefore I can conceive the epistemic possibility of zombies.

C7) From C5 and C6, I can infer that I am not a zombie: From C5, I know that I am not not conscious [now], from C6, it follows that if I were a zombie then I would not be conscious [now], therefore I am not a zombie.

C8) From C1, C2 and C3, a zombie does not have direct knowledge of being conscious: From C1, consciousness makes no difference to the operation of the physical brain. Therefore there will be no physical difference in the operation of the brain of a zombie and of a physically identical conscious individual. From C3, a conscious individual has direct knowledge of being conscious that is constitutive of being conscious. From C2, such knowledge cannot be explained in physical terms. From C1, C2 and C3, as a zombie is not conscious and as direct knowledge of being conscious is, in being constitutive of consciousness, a self-knowledge of consciousness only available to consciousness, it follows that a zombie does not have direct knowledge of being conscious.

C9) From C8, a zombie cannot conceive the possibility of being conscious: Here the claim is *not* that a zombie cannot come to possess the physical trace of a concept of being conscious. It is that any such concept that a zombie may possess will lack the direct knowledge of consciousness to which the conception of the possibility of being conscious refers. This reference is not a reference to something that could be the content of a concept, it is a reference to a non-conceptual knowledge, that, from C8, a zombie cannot possess. So, any concept that a zombie may possess concerning the possibility of being conscious will differ from that of a conscious individual in that its reference will be empty. This does not necessarily imply a physical difference between a zombie and its conscious physical duplicate, as, for example, the zombie may have acquired the physical trace of a concept of the possibility of being conscious on the basis of observation and mimicry.

C10) From C8 and C9, a zombie cannot conceive the possibility of not being conscious: From C8, a zombie does not have a direct knowledge of being conscious, and so, from C9, any concept it may have formed about being conscious will be empty, in the sense of not referring to a direct knowledge of being conscious. Therefore, the negation of the possibility of being conscious, i.e. the concept of the possibility of not being conscious, will also be empty.

C11) From C9 and C10, a zombie cannot distinguish between a conscious individual and a non-conscious duplicate: As the zombie's concepts of the possibility of being conscious (C9) and the possibility of not being conscious (C10) both have empty references, the zombie will not be able to distinguish between a conscious individual and a non-conscious duplicate. That is, the basis of the distinction will rest on the reference, which refers to the direct knowledge of being conscious in one case, and the absence of the direct knowledge of being conscious in the other. For the zombie both references point to nothing and so are equivalent, meaning there is no distinction to be made.

C12) From C7 and C11, consciousness makes a difference to the physical operation of the brain: From C7, I, as a conscious individual can infer that I am not a zombie, because I have substantive direct knowledge of being conscious, on the basis of which I can make a distinction between myself and a non-conscious physical duplicate. From C11, a zombie, lacking a substantive direct knowledge of being conscious, will be unable to distinguish between itself and a conscious physical duplicate, except on the basis of an external reference or comparison. That is, the zombie cannot make the distinction on the basis of its immediate internal state. It must rather make the distinction on the basis of some external evidence of a direct knowledge of consciousness, which implies the existence of a physical cause-effect relation with an entity that does possess such knowledge.

The argument is that the inability of a zombie to distinguish between itself and a conscious individual, without reference to external evidence, represents a difference in the physical behaviour of the zombie's brain in comparison with the behaviour of the brain of a conscious individual – because a conscious individual will be able to make such a distinction purely on the basis of a direct knowledge of being conscious.

As this is a *physical* difference, it should be possible to envisage a scenario where the difference makes an observable difference to the behaviour of an individual, according to whether or not the individual is conscious. This is exactly the purpose of the consciousness test.

4 The Consciousness Test Revisited

The consciousness test scenario asserts that both Eric$_c$ and Eric$_z$ initially lack an explicit concept that refers to a direct knowledge of being conscious. For, although Eric$_c$'s being conscious *constitutes* a direct knowledge of being conscious, because of his amnesia, he has no access to any concept or memory involving an explicit representation or recognition of a direct knowledge of being conscious. So there is no physical trace in Eric$_c$ of any effect of consciousness that he could rely on as the basis for forming an explicit concept of a direct knowledge of being conscious. In order to form such a concept, he must *discover* a direct knowledge of being conscious on the basis of an immediate demonstration of being conscious.

The moment of truth for the consciousness test is when Alan asks if either Eric$_c$ or Eric$_z$ can conceive it possible that their physical duplicate, in a physically identical environment, could have a different experience of the colour of green to the one that they are having. The argument is that in order to conceive such a possibility, it is necessary to conceive the possibility that there could be a difference in conscious experience without a corresponding difference in the relevant physical situation.

4.1 The Case of Eric$_c$

The claim, argued in C3, C4 and C5, is that Eric$_c$'s being conscious provides a direct *substantive* knowledge of *what it is* to be conscious. It is this knowledge that reveals the 'what it is like' of being conscious, i.e. its phenomenal character. If we combine this knowledge with C7, that Eric$_c$ can infer that he is not a zombie, it follows that Eric$_c$ can conceive of a zombie that would lack such phenomenal experience.

It is a simple step from here to assume that Eric$_c$, in the consciousness test scenario, conceives the possibility of his twin clone being a zombie, and is therefore able to conceive the possibility that his twin clone has a different phenomenal experience to himself, i.e. no phenomenal experience whatsoever.

4.2 The Case of Eric$_z$

The case for Eric$_z$ has already been argued in C8, C9, C10 and C11. If we accept C11, viz. that a zombie cannot distinguish between a conscious individual and a non-conscious duplicate, it follows that Eric$_z$, if he reasons correctly, will not be able to conceive that his physical duplicate could have a different experience of green to himself, given all the other conditions of the test.

If the central claim of the paper is accepted, viz. that there is direct knowledge of consciousness, and that Eric$_z$ does not have access to such knowledge, it could still be maintained that Eric$_z$ could *simulate* the possession of such knowledge in such a way that his behaviour remains indistinguishable from Eric$_c$'s. In that case, it would have to be shown how Eric$_z$ could come to utter the belief that his experience could be different from that of his physical duplicate. I cannot see how this can be done, unless we allow for the possibility that Eric$_z$ could come to act in this way by reasoning incorrectly, i.e. by mistake or by accident. We can block this possibility by assuming that both Eric$_c$ and Eric$_z$ are ideal reasoners, at least in relation to the tasks set them in the consciousness test, and that the ability to reason ideally is a consequence of the physical structure of the brain that Eric$_z$ and Eric$_z$ share. In that case, Eric$_z$ must reason correctly, and, according to the premisses of the thought experiment, he should correctly reason that there can be no difference in the experience of physical duplicates situated in physically identical environments.

4.3 The Causal Non-Closure of the Physical

The final upshot of this argument is that the two scenarios of the consciousness test result in different responses to Alan's final question, and as the two scenarios began as physical duplicates, only differing in the respect that Eric$_c$ is conscious and Eric$_z$ is not, it follows that the presence of consciousness is a difference-making cause of Eric$_c$'s differing response. Therefore the causal closure of the physical is false.

The core claims of the argument can be reduced to the following (i) that there is direct knowledge of consciousness (ii) that this knowledge is constitutive of being conscious, (iii) that this knowledge has no corresponding physical manifestation and (iv) that this knowledge enables conscious entities to make distinctions that they otherwise could not make.

If we accept (i) and (ii), the controversial claim for a phenomenal realist who also accepts the causal closure of the physical is (iii). Here, causal closure would require that there is some physical basis or representation of a direct knowledge of consciousness, in the same way that Chalmers' argues there are corresponding physical representations of direct phenomenal concepts. If such a physical representation can be demonstrated then this could provide a corresponding physical account of our ability to distinguish between zombies and conscious entities.

However, someone who is willing to accept the epistemic possibility of zombies, and is also willing to accept that there is direct knowledge of being conscious, must accept as an analytic truth that a zombie will lack such knowledge. Therefore the knowledge cannot make a physical difference to the constitution of the zombie's brain and claim (iii) is true.

If this is correct, then causal closure requires that either (i) or (ii) or both are false. To attack (ii) would require that there is some other way of acquiring a direct knowledge of being conscious, i.e. other than by being conscious. More particularly, for the purposes of the argument, it requires that a *zombie* could have some other way of acquiring a direct knowledge of being conscious. And as a zombie, by definition, can only be affected by physical events, this argument fails for the same reasons as above, viz. someone who is willing to accept the epistemic possibility of zombies, and is also willing to accept that there is direct knowledge of being conscious, must accept as an analytic truth that a zombie will lack such knowledge.

So, finally, defending causal closure involves denying that there is direct knowledge of being conscious. Here we arrive at a kind of bedrock position. For the primary justification for the existence of such direct knowledge is the immediate confirmation of being conscious. And this can be denied, or at least it can be given inferior status as a form of evidence or justification, e.g. by calling it an *intuition* rather than a knowledge. On this basis, it can be asserted that the causal closure of the physical, as a principle, carries more weight than any deliverance of intuition.[13] Taking such a position leads us out of the realm of rational argument and into the realm of metaphysical belief, i.e. where an acceptance of the causal closure of the physical becomes a core belief, on the basis of which one argues, rather than being something that can be argued for. Conversely, the assertion that a direct knowledge of consciousness is a certain knowledge rather than a fallible intuition, appears as an alternative core belief, about which *we* cannot argue further.

According to our best current physics, it is epistemically *and* metaphysically possible that the causal closure of the physical is false. Therefore the causal closure of the physical is not knowledge. So the core issue becomes whether my direct knowledge of consciousness is knowledge, or whether it is a fallible intuition which can only act as questionable evidence for the causal non-closure of the physical. To decide *that* would involve a major detour into the epistemological foundations of knowledge. Such an investigation would need to show that inferences made on the basis of direct knowledge of consciousness have the same kind of certainty that we attach to other forms of a priori reasoning. While I believe that such a case can be made, it goes beyond the scope of the current paper to argue this further. Instead, I shall take a tangential approach, via Penrose's Gödel-Turing argument, to show that a direct knowledge of being conscious is also implied in Penrose's a priori reasoning.

5 Parallels with Penrose

The basic claim of Penrose's Gödel-Turing argument is that human mathematicians have access to understanding that cannot be explained in terms of Turing computability. He also claims that '[t]his non-computational process lies in whatever it is that allows us to become *directly* aware of something.'[14]

5.1 The Non-Computability of Conscious Thought

Penrose's argument is based on Gödel's incompleteness theorem and Turing's related work on the undecidability of the halting problem, and can be summarised as follows:[15]

Suppose we have a computational procedure A that encodes all human mathematical knowledge for deciding whether a computation C ever halts, where C takes any natural number n as an argument. Here C can be any member of the ordered set of all possible computations that can take n as an argument: $C_0, C_1, C_2, C_3, C_4, C_5, \ldots,$ Let us also suppose there is a computation $C_\bullet$ that gives the computation $C_q(n)$ when presented with q and n. A then takes two arguments, q and n, and halts only if $C_q(n)$ does not halt, i.e.:

If $A(q,n)$ halts, then $C_q(n)$ does not halt.

Now consider the substitution $q = n$:

If $A(n,n)$ halts, then $C_n(n)$ does not halt.

As $C_0, C_1, C_2, C_3, C_4, C_5, \ldots,$ is an ordering of *all* possible computations that can be performed on a single natural number n and as $A(n,n)$ is a computation involving only one natural number, it must be the case that there is a C_k such that $A(n,n) = C_k(n)$. As before, we can substitute $n = k$ to give $A(k,k) = C_k(k)$, which also gives:

If $A(k,k)$ halts, then $C_k(k)$ does not halt.

As $A(k,k) = C_k(k)$, it follows that:

If $C_k(k)$ halts, then $C_k(k)$ does not halt, therefore $C_k(k)$ does not halt.

The key point is that if $C_k(k)$ does not halt, this is the same as saying that $A(k,k)$ does not halt, which means that A is unable to report that $C_k(k)$ does not halt. However, *we* know that $C_k(k)$ does not halt because otherwise there will be a contradiction. So we can report something that A is unable to report, even though A encodes (by definition) all procedures known by human mathematicians for deciding whether a given computation $C_q(n)$ halts. Therefore, concludes Penrose, human mathematicians cannot be using a Turing computable procedure to ascertain mathematical truth.

Clearly, $A(k,k)$'s inability to report that it will halt is not due to the complexity of the problem, it is due to the *self-reference* of A's arguments. For, k is the *number* of A, i.e. when A's arguments are identical, then A is identical with the k^{th} computation in the $C_0, C_1, C_2, C_3, C_4, C_5, \ldots,$ ordering. The problem for A is that it is unable to report on its own behaviour, i.e. it cannot answer whether it itself will halt when given its own number as an argument. Again this is not due to any lack of computing power. It would be a simple matter for A to include a function $C_{\bullet\bullet}$ that returns the number of A, i.e. k. All A then needs to include is a simple test of the form:

if $a_1 = C_{\bullet\bullet}$ and $a_2 = C_{\bullet\bullet}$ then *contradiction found.*

where a_1 and a_2 are the two input arguments $A(a_1, a_2)$. So A can easily recognise that it is being asked to decide about whether it itself will halt. And it can reason that it is in an impossible situation: if it halts then that means it will *misreport* that it will not halt. Here Penrose assumes that A is *sound* in that it *cannot* misreport that a computation halts which does halt, but it can *fail* to report such a situation. So A must continue running, not because it cannot reason the situation out, but because it is *unable to report* what it has reasoned, without contradicting itself.

When put this way, it appears that the inability of A to report that it will not halt is an artefact of the way the situation has been engineered, i.e. if we change the rules and let A stop when $a_1 = a_2 = k$ and report the further fact that it has only stopped because it was given itself as an argument, then it appears the paradox is resolved.

However, the situation is not that simple. A represents *all* techniques known to human mathematicians for deciding whether a computation $C_q(n)$ will halt that it is *possible* to formally encode, i.e. that are Turing computable. So, if the way we are able to decide that $C_k(k)$ will not halt is *able* to be formally encoded, then it would already be in A. Therefore, the way we are able to decide that $C_k(k)$ will not halt *cannot* be formally encoded as a Turing computable procedure. To complain that A really 'knows' about the contradiction does not alter the result: when *we* are presented with the problem, 'Does $C_k(k)$ halt?' *we* are able to correctly answer 'No.' When A is presented with the same problem, it *cannot* answer. Therefore we *cannot* be using a Turing computable procedure to arrive at our answer.

5.2 I am not a Number

I will take it that it is possible Penrose is correct[16] and concentrate on explaining *why* he could be correct, i.e. *how* human mathematicians could arrive at conclusions unavailable to Turing computable procedures. It is here that the parallel with the consciousness test arises: if the consciousness test arguments are correct, then we already have an instance of a human reasoner arriving at a conclusion unavailable to Turing computable procedures, i.e. in Eric$_c$ concluding that his physical duplicate could lack consciousness.

Of course, this assumes that Eric$_z$'s physical brain processes can be exactly simulated by Turing computable procedures, or equivalently, that functionalism is true for a non-conscious brain. That is not to say that *our* brains can be simulated in this way, or to suggest one way or another that non-computable quantum effects can be used to explain the reasoning abilities of human mathematicians.

Given the computability assumption, we can say that Eric$_z$'s inability to distinguish between himself and Eric$_c$ is also the inability of the Turing computable procedures that represent his brain to correctly infer that there could be more to being conscious than being in a certain kinds of physical state. What Eric$_z$ as a formal system lacks,[17] is what A as a formal system lacks, that is, *consciousness.* And what consciousness provides in both cases, is the ability to make new inferences that are not available within the formal system itself.

Defining $\boldsymbol{A}$: To make this clear, consider the situation of the brain of a human mathematician. If we accept the computability assumption, then there will actually be a Turing computable procedure A that human mathematicians are using to decide whether a given computation will not halt, and this procedure will be realised in the physical brain of a *competent* human mathematician. Here, competency means that the mathematician reasons *soundly* in relation to whether a computation will not stop, i.e. the mathematician will never report that a computation will stop that in fact does not stop. So, if the mathematician has any doubt about the computation not stopping, she will simply not provide an answer.

Now, it may be the case that different mathematicians reason in different ways according to the precise configurations of their brains. But we are not concerned with these implementation details. All that matters for the argument is *functional equivalence* in terms of inputs and outputs. So, for us to discover the procedure A that represents a human mathematician's ability to answer questions of the kind: 'Will $C_q(n)$ fail to halt?' we only require a procedure that picks out the same set N of non-halting $C_q(n)$ computations as the human mathematician. As we have assumed that competent mathematicians reason soundly about non-halting problems, the only difference between individual mathematicians would be the size of the set N that they are able to calculate, i.e. $|N|$. To simplify the argument, I shall assume that $|N|$ will eventually be identical for all competent human mathematicians, on the proviso that each is given infinite time. In other words, I am assuming, with Penrose, there is *one* procedure A that represents all the valid methods that human mathematicians can use to prove that computations of the form $C_q(n)$ will halt.

Discovering the number of $\boldsymbol{A}$: The next assumption is that there exists a completed neuroscience that has discovered the principles on which the human brain functions to a degree that is sufficient to formally specify A. For example, it may turn out that the brain is structured to implement a certain kind of Bayesian network that forms beliefs on the basis of correct probabilistic reasoning. Given the right training and exposure to the right kind of environment, we assume this network can be shown to be equivalent to a certain idealised reasoning system R. We can further assume that the presence of R can be confirmed by a brain scan, such that a person can be certified as R *competent.*

The procedure A can now be considered as a particular procedure that an R competent mathematician can implement by setting out to solve a problem of the form: Given q and n and a function $C_\bullet$ that, when given q, generates a computation C_q

that accepts a single natural number as an argument, calculate whether $C_q(n)$ is a non-halting computation. You may only answer 'Yes' if the computation does not halt, otherwise you do not answer at all.

We also assume that the functions $C_\bullet$ and $C_{\bullet\bullet}$ are *included* in A, where $C_{\bullet\bullet}$ returns the *number of* A when A is given identical natural numbers as input. Here, $C_{\bullet\bullet}$ numbers A according to the same conventions that $C_\bullet$ uses to generate C_q from q. So, for example, if we present A with input q and n where $q = n$ then the procedure $A(n, n)$ will be equivalent to one of the set of C computations that take n as an argument. The function $C_{\bullet\bullet}$ returns the number k of this computation such that $A(n, n) = C_k(n)$.

In this case, $C_{\bullet\bullet}$ is not returning the number of an abstract procedure A that is similar to the procedure used by the human mathematician. Rather, we are assuming that $C_{\bullet\bullet}$ returns the number of the actual procedure that the human mathematician is using to decide the whether $C_q(n)$ does not halt.

Escaping the paradox: To recap, our human mathematician's task is, given q and n as input, to provide an answer 'Yes' if $C_q(n)$ does not halt, otherwise do nothing. We can now examine the situation of such an R competent human mathematician (let's call her *Persephone*) being presented with the identical arguments $n = q = k$. Here, from $C_\bullet$, Persephone will derive $C_k(k)$ and from $C_{\bullet\bullet}$ she will derive that the number of the procedure she is now using to work out whether $C_k(k)$ will not halt is also k. Therefore, Persephone can reason that if she reports $C_k(k)$ does not halt, that will be an example of $C_k(k)$ halting, producing a contradiction. Therefore, as a competent mathematician, mindful of not losing her R competent status, what is the answer? To keep silent?

If we accept Penrose's argument, then it must be possible for Persephone to answer 'Yes' without contradiction. The problem is therefore to explain *why* there is no contradiction. Persephone's answer is: '$C_k(k)$ does not stop because if it stopped there would be a contradiction, and I can report this, on the basis of a direct knowledge of being conscious that cannot be represented in A. This direct knowledge allows me to infer that *I am not* A and therefore that I am not bound by the condition that A cannot halt.

Persephone's reasoning again relies on there being direct knowledge of consciousness that cannot be represented in any physical, conceptual or procedural system. That means that A in and of itself cannot rely on any direct knowledge of consciousness as a step within its reasoning process. Otherwise that direct knowledge would be physically and procedurally represented in A. If a direct knowledge of being conscious *were* represented in A, it could only be represented *indirectly,* for example, as a rule that allows A to conclude that it will not stop. But such a rule will be of no use to A, because it will be *internal* to A, and therefore cannot be used to break out of the contradiction.

To avoid misunderstanding: the argument is not that Persephone can use some higher-order mental process to reason about A as an object 'from the outside.' For, if

Persephone were to do this, then presumably *that* higher-order mental process could also be encoded as a Turing computable procedure, and, as it is being used to reason about whether $C_k(k)$ will not stop, *that* procedure also belongs in A. We can finally argue that every capacity for thought that Persephone possesses could become part of A, i.e. all of R. But that makes no difference: A, insofar as it is a computational procedure, can only report 'Yes' or remain silent. However sophisticated the procedures it possesses, it is written into the Gödel-Turing argument that A cannot report that it stops when given $q = n = k$ as input.

The situation for Persephone in relation to A is analogous to Eric$_c$'s situation in relation to Eric$_z$. For A, as a computational procedure, must, by definition, exactly follow the rules and axioms it embodies. Therefore A will behave in the same way, whether or not it is realised as a conscious system (C1). And A cannot infer anything new on the basis of a direct knowledge of being conscious, insofar as such knowledge is non-conceptual and only accessible on the basis of being conscious, and cannot be expressed in terms of procedural instructions (C2). Therefore, Persephone can conclude that A is like a zombie, in that it cannot be affected by, or informed by a direct knowledge of being conscious.

In addition, by demonstrating a direct knowledge of being conscious (C3 and C4), Persephone can infer that she is not not conscious [now] (C5). Therefore, she can distinguish between herself and a computational procedure (C7), even if that procedure is the very procedure that is executing in her brain [now].

It is on the basis of this distinction, between herself as being conscious, and A as the procedure running in her brain, that Persephone can conclude that, although A cannot report that it will not stop, *she* can report this without contradiction, because she is not A, or rather, she is *more than A*.

The foregoing analysis places the Gödel-Turing argument in a new light. The point that the argument makes is not that there are Gödel sentences that will defeat a formal system on the basis of their sheer complexity – such as Chalmers' picture of the ' "outer limits" of Gödelization.'[18] It is rather that formal systems are self-limiting, unable to draw certain inferences about themselves as a whole, because to do so would lead them into contradiction. The argument here is that consciousness escapes these limits because, in reflecting a system to itself, consciousness provides a non-conceptual knowledge that exists outside or beyond the system that is reflected.

6 Conclusion

The main argument of the paper can be summarised as follows:

1. There is direct non-conceptual knowledge of being conscious only accessible on the basis of being conscious.

2. Direct knowledge of being conscious allows conscious entities to make distinctions between conscious and non-conscious entities that cannot be made by non-conscious entities.
3. The ability to distinguish between conscious and non-conscious entities has physical effects.
4. Therefore the causal closure of the physical is false.

Notes

[1] (Moody, 1994).

[2] (Chalmers, 2003).

[3] (Chalmers, 2003, p. 223).

[4] (Chalmers, 2003, p. 223).

[5] (Chalmers, 2003, p. 225).

[6] (Chalmers, 2003, p. 235).

[7] (Moore, 1903, p. 450).

[8] (Chalmers, 2006).

[9] (Russell, 2007).

[10] (Sartre, 2003).

[11] For example, see (Rosenthal, 2002).

[12] (Nagel, 1980, p. 160).

[13] For example, consider Jackson's final rejection of his Mary argument on the grounds that we must make 'a choice between going with science and going with intuition' (Jackson, 2003, p. 251).

[14] (Penrose, 2005, p. 53, my emphasis).

[15] A complete exposition of Penrose's Gödel-Turing argument is given in (Penrose, 2005).

[16] For further discussion on the correctness of Penrose's argument, see Chalmers' review of Shadows of the Mind (Chalmers, 1995) and Penrose's reply (Penrose, 1996).

[17] Here I am assuming (with Penrose) that a set of Turing computable procedures, such as those that comprise A, can be represented as an equivalent formal system comprising of a set of rules and axioms.

[18] (Chalmers, 1995)

References

Chalmers, D. J. (1995). Minds, machines, and mathematics: a review of Shadows of the Mind by Roger Penrose. *Psyche*, *2*(9).

Chalmers, D. J. (2003). The content and epistemology of phenomenal belief. In Q. Smith & A. Jokic (Eds.), *Consciousness: New philosophical perspectives* (p. 220-272). Oxford: Oxford University Press.

Chalmers, D. J. (2006). The representational character of experience. In B. Leiter (Ed.), *The future for philosophy* (p. 153-181). Oxford: Oxford University Press.

Jackson, F. (2003). Mind and illusion. In A. O'Hear (Ed.), *Minds and persons* (p. 251-272). Cambridge: Cambridge University Press.

Moody, T. (1994). Conversations with zombies. *Journal of consciousness studies*, *1*(2), 196-200.

Moore, G. E. (1903). The refutation of idealism. *Mind*, *12*(4), 433-453.

Nagel, T. (1980). What is it like to be a bat? In N. Block (Ed.), *Readings from the philosophy of psychology: volume one* (p. 159-168). Cambridge, Massachusetts: Harvard University Press.

Penrose, R. (1996). Beyond the doubting of a shadow: a reply to commentaries on Shadows of the Mind. *Psyche*, *2*(23).

Penrose, R. (2005). *Shadows of the mind: a search for the missing science of consciousness.* London: Vintage Books.

Rosenthal, D. M. (2002). Explaining consciousness. In D. J. Chalmers (Ed.), *Philosophy of mind: Classical and contemporary readings* (p. 406-421). Oxford: Oxford University Press.

Russell, B. (2007). On the nature of acquaintance. In R. C. Marsh (Ed.), *Logic and knowledge* (p. 127-174). Nottingham, England: Spokesman.

Sartre, J.-P. (2003). *Being and nothingness* (H. E. Barnes, Trans.). London: Routledge.