Against the necessity of functional roles for conscious experience: reviving and revising a neglected argument

Gary Bartlett

Central Washington University

1. Introduction

Functionalist theories of conscious experience identify experiences as states which have a certain type of functional role. This identification entails, firstly, the

Functional sufficiency thesis about experience (FST_E): having a state with a certain type of functional role is sufficient for having a conscious experience of a certain type,

and secondly, the

Functional necessity thesis about experience (FNT_E): having a state with a certain type of functional role is necessary for having a conscious experience of a certain type.

Each thesis can be formulated with logical or nomological force. My concern is with the FNT_E's nomological version (the falsity of which entails the falsity of any stronger version). Henceforth, by 'the FNT_E' and 'the FST_E' I refer to the *nomological* versions.

The FNT_E gets cashed out as a requirement that each type of experience entails certain behavioral, cognitive or affective dispositions. For example, being in pain might entail dispositions to grimace, to desire relief, and to be grumpy.

The FNT_E is at least as central to functionalist theories as is the FST_E. But it is also more widely held than the FST_E, for it is enforced not just by overtly functionalist theories (e.g., Van Gulick 1993). Many theories, such as Rey (1997), reject the FST_E by placing further conditions on experience, yet retain the FNT_E. Most representational theories of experience, whether first-order or higher-order, give dispositions a necessary role. Thus Michael Tye (1995, 2000) holds that experiences are necessarily disposed (as he says, 'poised') to affect beliefs and desires. Also committed to the FNT_E is Peter Carruthers' higher-order theory (2000; 2005), on which

conscious experiences are disposed to cause higher-order thoughts about themselves; see also Van Gulick (2004). In general, any representational theory that relies on a functional role (or 'consumer') semantics to individuate experiences will embrace the FNT_E.

Yet while the FST_E has received wide critical attention, the FNT_E remains largely under the radar. As illustration, consider Janet Levin's entry on functionalism in the *Stanford Encyclopedia of Philosophy*. She mentions anti-FNT_E arguments only very briefly, in a section titled 'Inverted and Absent Qualia', which – as that title itself suggests – focuses on anti-FST_E arguments like Block's (1978) 'Chinese nation'. Regarding anti-FNT_E arguments, Levin merely remarks that it has been argued that 'people may have mild, but distinctive, twinges that have no typical causes or characteristic effects' (2010, §5.5.1, ¶2).

I do not mean to pick on Levin, however. Her limited coverage of the FNT_E merely reflects the state of the literature.

I shall present an argument against the FNT_E. More precisely, I shall present the argument's second half. Its first half was convincingly presented by Michael Antony (1994), in a paper which, like anti-FNT_E arguments in general, has been sadly ignored. His presentation of the argument's second half, however, is *un*convincing. I intend to fill this gap.

Let me explain. Suppose a theory of conscious experience says that functional state *F* is necessary for an experience *E*. An anti-FNT_E argument will describe a system that lacks *F* but still has *E*. The argument thus has two stages: the first describes conditions under which the system lacks *F*, and the second argues that under those conditions the system still has *E*.

One can run this sort of argument in the way noted by Levin (2010), by arguing that *some* experiences simply lack a proprietary functional role. But such an approach is unconvincing (as is perhaps indicated by Levin's lack of elaboration). The argument I am pursuing aims for a larger goal: that *no* functional role is necessary for *any* experience.

Antony (1994; all citations of Antony will be to this paper) describes a manipulation of the brain of a subject who is (at least to begin with) in pain. The manipulation is designed to cause the brain state that is realizing the pain experience to lose its functional role. This is the first

stage of the anti-FNT_E argument. Antony spends most of his time defending this stage. His defense is compelling, and I hope readers will forgive me for not repeating his discussion in any depth. Antony's defense of the argument's second stage, however, is brief and (in my view) weak. I shall give a new defense of the second stage. This is, I think, a worthwhile endeavor. As I have indicated, a number of theories of conscious experience would be placed in serious jeopardy if the anti-FNT_E argument that he constructed could be made sound.

Here is the plan. §2 recaps the case on which Antony builds his argument. Again, I will not spend much time repeating his defense of the argument's first stage. I will, however, outline his very brief defense of the *second* stage, and explain its inadequacy. §3 introduces a principle that is central to my own defense of the second stage. In §4 I modify Antony's case so that my principle will apply to it. In §5 I formulate my argument, and §6 defends its key premise.

2. Antony's argument: the case of Sam

Antony imagines a person, Sam, who starts to experience pain at time t_1, when his brain's 'pain region', which for brevity I shall call p^S, goes into an active state that I shall call ϕ. (The shorthand terms are mine, not Antony's.)

The FNT_E entails that ϕ realizes pain in virtue of its causal relations to other brain states, which underpin Sam's pain-related dispositions. Thus ϕ is disposed to cause activity in other regions of his brain; for example, in the region responsible for action planning, so that Sam devises a plan for obtaining pain relief (say, by finding some aspirin). I shall call the action-planning region a^S. State ϕ of p^S, then, realizes pain partly in virtue of its disposition to activate a^S. Antony's case now goes as follows (I insert my terminology in brackets):

> Imagine that there exists some impressive technology by means of which one can instantly incapacitate (or destroy) any region of an individual's brain one chooses, right down to components of individual cells. That done, imagine… that activity [state ϕ] in Sam's pain region [p^S] has begun at t_1, and that the signal is traveling toward the action-planning region [a^S] but thus far has reached only the

> half-way point along the connecting neural pathways. Call the time at which the signal reaches the half-way mark 't_2'. Now assume that Sam's pain region [p^S] has been activated since t_1, and that it remains activated at t_2. Suppose, finally, that at t_2 the action-planning region of Sam's brain [a^S] is incapacitated with our imaginary technological wonder. (p. 107)

As of t_2, then, Sam cannot formulate a plan to get aspirin, or indeed any plan at all. So, says Antony, ϕ 'no longer bears the necessary causal relations for pain, according to functionalism, since it is no longer disposed to cause an action-planning process' (p. 108). This is the conclusion of the first stage of the anti-FNT_E argument: as of t_2, ϕ has lost the functional role of pain. If so, then according to the FNT_E *Sam is no longer in pain* – thanks simply to an incapacitation of an inactive part of his brain.

Let me pause to address the worry that Antony's argument is scientifically naïve. I foresee three specific concerns.

Concern (i): the brain has no 'pain region' or 'action-planning region'. We now know that pain has a highly distributed network of cerebral correlates, including in the thalamus, the insula, and the anterior cingulate cortex (e.g., Tracey & Mantyh 2007). And certainly action planning is not localized to a specific brain region; the main locus is the prefrontal cortex, which houses executive functions (e.g., Fuster 2008, Ch. 5), but other areas are also likely involved. However, Antony himself notes that his argument "does not assume that there is a single region of Sam's brain where all pains occur, or that there are brain regions that subserve only pain, but only that each pain occurs some place or other", and that the same goes for the 'action-planning region' (p. 107n). The argument will run on any two brain regions, or even two sets of regions, that are sufficiently anatomically distinct that signals may be sent between them and that one region may be incapacitated while the other is not. Nothing about this is impossible. And if you think it's impossible in the specific case of pain and action-planning, others could be substituted; pain and action-planning are simply examples.

Concern (ii): it is not possible to perform the required kind of anatomically specific and temporally precise neural incapacitation. Antony's 'impressive technology' is certainly fiction for now. But it is not nomologically impossible. We can already shut down one hemisphere of the brain by injecting sodium amobarbital via the carotid artery, in what is known as the Wada test for hemispheric localization of language and memory functions. What Antony envisions is a more precise version of that. We can imagine delivering a drug into the brain that immediately alters the chemistry of the neurons within a defined range of the region of release so that they cannot carry action potentials.

Concern (iii): it is an empirical matter whether the experience of pain would be cancelled (or even affected) by the incapacitation of knowledge of how to seek pain relief. It is crucial to see that this concern must be directed *at the functionalist*. A key point of my argument is that the FNT_E entails some very bold empirical predictions. For if it is true, then *the cancellation of a functional role entails the cancellation of any current experience for which that role is necessary*. For example, the cancellation of the pain role, perhaps by incapacitating the ability to plan the acquisition of relief, should cancel (or at least reduce) the subject's pain experience. I am questioning whether such predictions are plausible. Ideally, the answer would indeed be delivered empirically, by actually performing the sort of experiment Antony describes. But we currently lack the knowledge and technology to do this. As for existing research in cognitive neuroscience, it may be suggestive, but it will not supply a crucial test. For example, we might look at patients with prefrontal lesions that have impaired their executive functioning, and ask whether their ability to experience pain has been affected. I know of no studies indicating such an effect; but the functionalist might reasonably object that case studies, which are by their nature uncontrolled, are a poor test of her theory. She might also argue that cancelling just *one* of ϕ's causal relations will not suffice to cancel ϕ's realization of pain (see below); that is, that a wider set of lesions would be required. Of course, we are unlikely to find case studies in which just the right set of brain regions are damaged; and indeed we do not yet even know what the right set of brain regions would *be*. Therefore, at least for now, only in imagination can we carry

out the precise sort of neural manipulations that are called for. Empirical tests (if we decide they are worthwhile) must wait for the future.

Let us return to Antony's argument. As I have said, he focuses his defensive efforts on its first stage, which I described above. I can only sketch his points here. He considers and rebuts three specific ways of cashing out the most obvious objection, which is that canceling a single causal relation of ϕ could not cancel its entire realization of pain. The thrust of his rebuttals (see his §5) is that his case may be amended to involve canceling whichever causal relations, and however *many* relations, are deemed necessary to support the pain realization – by incapacitating not just a^S, but many regions of Sam's brain.[1] For ease of exposition I shall proceed as if only a^S is incapacitated, but it should be kept in mind that this restriction is not mandatory.

Another objection can be motivated by an analogy of a kind that is common in the literature on dispositions. Imagine encasing a fragile vase very thickly in bubble-wrap so that it is safe from damage. We nevertheless think the vase retains its fragile disposition. Similarly, can't the functionalist say that even though ϕ's disposition to activate a^S cannot be *manifested*, it still *exists* (since p^S is intact)? However, Antony points out that this objection gets functionalism wrong. The view it suggests results in an explosion, and thus a trivialization, of functional roles. The functionalist would end up saying that *anything* could realize *anything*. As Antony says, for example, on that sort of view "the temperature of my coffee cup could realize pain, since it is disposed to cause an action-plan-construction" (p. 108n) – for *if* the cup *were* hooked up to the right sort of system components, its temperature *would* cause construction of an action plan (and whatever other effects are part of the pain role). Here is another way to put the point, which will also be important below. It is precisely in order to avoid this explosion of functional roles that

[1] Teleofunctionalist theories may have a way out. If functional roles are defined purely biologically, ϕ might keep its functional role even if *all* of its causal relations are cancelled. However, I do not know if any teleofunctionalists would actually adopt this view, for it seems to entail that ϕ would still realize pain if p^S were removed from Sam's brain, put in a jar, and artificially stimulated. This sort of result is usually regarded as anathema by functionalists.

functionalists hold that a mental state's *core realizer* (in this case, activity ϕ in p^S) is *not sufficient* for realizing the mental state (in this case, pain). Also necessary is a set of causal relations between the core realizer and various other brain states. The core realizer plus those other states compose pain's *total realizer* (Shoemaker, 1981). The total realizer of Sam's pain thus includes, we assume, a^S's readiness to be activated. So if a^S is incapacitated, p^S no longer has the disposition to activate a^S.

This ends my discussion of the first stage of Antony's argument. Most readers will, I believe, be ready to accept the argument that far (and I commend Antony's discussion to anyone with further doubts). Since the functional roles of mental states are commonly taken to supervene on the brain, it will not be controversial that interference in the right places in Sam's brain would cancel the functional roles of his experiences, such as pain. What *will* be controversial, however, is the further claim that Sam's experiences themselves could *survive* such interference. This is the burden of the second stage of the argument, which is my concern in this paper.

In his argument's second stage Antony must argue that Sam would, in fact, still be in pain after the incapacitation of a^S. His defense of this claim is very brief. In the short third section of his paper, he says that it is supported by 'an *intuition* – one to the effect that manipulating a system's inactive or unused parts can have no effect on that system's experience at the time of the manipulation' (p. 113).

This intuition is compelling. I myself find it deeply appealing.[2] However, it will also attract some justified resistance. The resistance will be based on the countervailing intuition of the multiple realizability of the mental. It is widely held, especially (but not only) by functionalists,

[2] Maudlin (1989) gives an argument that relies on the same basic intuition (I discuss this argument in Bartlett 2012). However, he targets *computational* theories only, and says that functionalism is untouched. I am not sure Maudlin is right about this, but in any case the present paper tries to find another way of driving the argument home against functionalism.

that mental states can be realized by a variety of physical states. So why does the fact that a^S is physically inactive show that it is extraneous to Sam's pain experience?

Antony's response is to stand pat on his intuition. He says that describing the incapacitated brain region (in Sam, a^S) as 'inactive', 'unused', or 'inert' suffices to secure his argument in all cases concerning 'systems in which the realizations of conscious states and processes involve (neural) activity' (pp. 113-114). He thus appears to assume that it is simply evident that neural activity, and *only* neural activity, realizes conscious experiences in humans. However, this is *not* evident. Antony himself admits that how humans' experiences are realized is a contingent matter. But empirical evidence does not show that experience in humans is realized solely by neural activity. At best it shows that neural activity is *necessary*.

Moreover, functionalists have principled reasons to dismiss Antony's intuition. Functionalism views mental states as fundamentally relational. As I said earlier, causal relations between ϕ (the activity in p^S) and various other brain states are what enable ϕ to occupy the pain role. ϕ is only the core realizer; the total realizer includes a^S's intactness. For the functionalist, then, a^S is involved in realizing Sam's pain, so of course incapacitating it would result in the pain state's elimination. The fact that a^S is inactive does not prevent it from standing in a causal relation; but the incapacitation disrupts the causal relation that constitutes the pain state.

Now of course Antony need not accept the functionalist's relational view of experiences. But equally, the functionalist need not accept Antony's intuition that physically inert states cannot be part of the realization of experiences. At best we have a stalemate.

In short, the anti-FNT_E argument that Antony is attempting to provide still stands in need of its second stage. I propose to complete that stage of the argument in a way that works *within* the functionalist's own conceptual apparatus of core and total realizers. My defense will rest on the claim that, given that the incapacitation of a^S could not affect Sam's immediate thoughts or behavior, it could not eliminate (or even modify) his experience of pain.

3. Introducing the Constraint on Experiential Change

Central to my argument that Sam would remain in pain after the incapacitation of a^S will be the following principle:

> *The Constraint on Experiential Change*. Given a subject S who is attending to a current conscious experience E, changes in E will cause immediate change in S's dispositions to thought, affect, or behavior.

The Constraint, as I shall call it for brevity, captures the idea that changes to one's conscious experiences are able to immediately shape one's thought, affect, or behavior (or all of these). It is not about what is *constitutive* of experiences. It does not say that changes in cognitive, affective, or behavioral dispositions are a necessary condition on experiential change. It says only that such changes in dispositions are caused by changes in attended experience. So the Constraint is a pretty weak constraint on the attribution of experiential change. I submit that an adequate theory of conscious experience must satisfy it.

I have formulated the Constraint so as to skirt around doubts about the causal efficacy of experiences. It has been suggested that conscious experiences are merely epiphenomena of unconscious neural processes. Studies by Benjamin Libet (1985) and Daniel Wegner (2002) purport to show that we are conscious of some acts of will only after the decision to act has been made unconsciously. In a similar vein, studies of visual deficits such as blindsight and agnosia suggest that the visual system has two 'streams', and that we are conscious only of one of them even though it is the other that controls many of our behaviors (for a review, see Milner & Goodale, 2009). These claims are controversial. But even if they have merit, they do not show that *all* of our experiences are epiphenomenal. To the contrary, it seems clear that conscious experiences frequently shape our thought, affect, and behavior. Consider, for example, your ability to attend to your pain when your doctor asks you to rate its severity. In any case, certainly no functionalist will want to say that experiences are completely epiphenomenal, and so I shall

assume that they are not – at least in cases where the subject is actually attending to the experience. These are the cases to which the Constraint applies.

The Constraint embeds two claims. The first is that changes in attended conscious experience cause changes in our cognitive, affective, or behavioral dispositions. The second is that such dispositional changes occur more or less *immediately*: the experiential changes alter our dispositions right away, as opposed to only at a later time (perhaps mediated by independent causal factors). It is convenient to discuss these two claims in reverse order.

The second claim, that experiences can have immediate effects, is quite nicely captured by Michael Tye's (1995, 2000) claim that conscious experiences are 'poised' to affect one's beliefs and desires, and thus also one's behavior;[3] or Jesse Prinz's recent claim that the function of consciousness is to "provide a menu for action" (2012, p. 203) by making available a range of options extracted from the current situation. Thus a splitting headache causes you to grimace and try to remember where the aspirin is. The sound of your dog barking causes you to wonder what is bothering him, and maybe to go and check on him. And so on. Of course, an experience will not *always* affect your actual thoughts, emotions, or behavior. The claim is only that it *can* do so. Perhaps if you are distracted then an experience might have no effect even on your *dispositions* to thought, emotion, or behavior. But such cases fall outside the scope of the Constraint.

The Constraint's first claim is that changes to one's experiences bring about changes to one's cognitive, affective, or behavioral dispositions. I stress that this is not intended as a conceptual truth. The primary evidence for it lies in everyday observation. First-person acquaintance with our own experiences indicates that distinct experiences tend to have distinct effects on thought, emotion, and behavior. Such informal observations are also supported by empirical research into the causal powers of experiences (e.g., Morsella 2005).

[3] But I do not endorse Tye's claim that such poising is *constitutive* of experiences (see above).

The worry that some experiences may *not* be poised to affect thought, emotion, or behavior may be put aside if we focus on experiences that are central to the attention of a normal individual.[4] I doubt there could be a change to an attended experience that did not immediately alter the subject's cognitive, affective, or behavioral dispositions.

In reply, some might appeal to Putnam's (1963) 'super-spartans', who are trained to suppress all behavioral signs of pain: perhaps with even more training, they might be able to suppress the mental signs too, and not even think about the pain or have it change their mood. But (I reply) then they would fall outside the scope of the Constraint, for they would not be *attending* to the pain. And even if they were somehow still attending to it without *thinking* about it, they would not be without *dispositions* to think about (or act on) their pain. They would merely have trained themselves not to manifest those dispositions.

A more realistic concern might point to phenomena such as inattentional blindness and change blindness. Simons and Chabris (1999) found that if subjects were asked to track how many times a basketball was passed by a team of players in a video, many subjects would completely fail to notice a person in a gorilla suit walking across the middle of shot. Rensink, O'Regan and Clark (1997) found that it was extremely hard for subjects to notice some major changes in an image – such as a central object disappearing – if the original and the changed version were flashed successively and repeatedly with a mask in between. Such results might be taken to show that our experiences can undergo radical changes without our even noticing. However, it is far from clear that this is the correct interpretation. It may be that subjects simply don't *experience* the gorilla, or the object's disappearance, at all. Or more conservatively, even if they do experience the gorilla or the disappearance on some level, there is no reason to think that they are also *attending* to it. Indeed there is every reason to think that they are *not*. Therefore,

[4] Of course, many will want to say that Sam is far from normal. I address this concern in §6.

such cases fall outside the Constraint's scope. Certainly there is no positive reason to think that Sam's situation is anything like that of the subjects in these studies.

4. Preparation for the main argument: adapting the Sam case

I now introduce two alterations to the Sam case. The first is simply that the pain causes him to grimace, and that he will keep grimacing until the pain lessens. The second alteration will be more controversial: it is that a^S is not involved in causing the grimace. More specifically, I wish it to be the case that when a^S is incapacitated at t_2, *Sam's grimace is unaffected*.

You might be surprised that I call this alteration controversial. For isn't it *obvious* that Sam's grimace would not be affected by a^S's incapacitation? Grimacing is a natural and *unplanned* expression of pain. I agree, but I cannot use this as a reason for the alteration, because it is specific to this case. I have chosen the example of a grimace precisely *because* it seems unlikely to be causally connected to action planning. The functionalist is claiming that, insofar as a^S's intactness is a part of the total realizer of Sam's pain (see §2), its incapacitation will end his pain experience. In turn, that strongly suggests that the *effects* of that experience will also be affected – and that goes for *all* of the effects of the experience. The grimace is simply one example. I am looking to construct an argument that will apply to all of them. Indeed, I want my argument to apply not just to any effect of a pain experience, but to any effect of any *experience* at all. So if I were to simply say that grimacing is unplanned behavior and thus not likely to be affected by the incapacitation of a^S, that reason would not generalize, so it would not serve my purpose. I must give a reason for the second alteration that will apply in the general case.

An analogy will help illustrate what I am proposing. Consider a car's state of braking. The core realizer of that state will be a state of the brake system, which comprises the brake pedal, the pads, the discs and all the interconnections of these parts, including the master cylinder and brake lines. Call this system b^C. The core realizer of the state of braking, then, is a certain active state of b^C: minimally, the state of the pedal's being depressed and the pads' pressing against the

discs. Call that state β. Braking's *total* realizer will further include states of other parts of the car: its wheels, for example, since β's disposition to cause a slowing of the revolution of the wheels is constitutive of its being the core realizer of *braking*. Finally, a normal effect of braking is that the tail-lights illuminate – via a switch under the brake pedal which powers the tail-lights.

(A terminological note: I shall refer to the parts of a total realizer that are outside its core realizer – such as a^S in Sam, and the wheels in the car – as the 'non-core elements' of the realizer.)

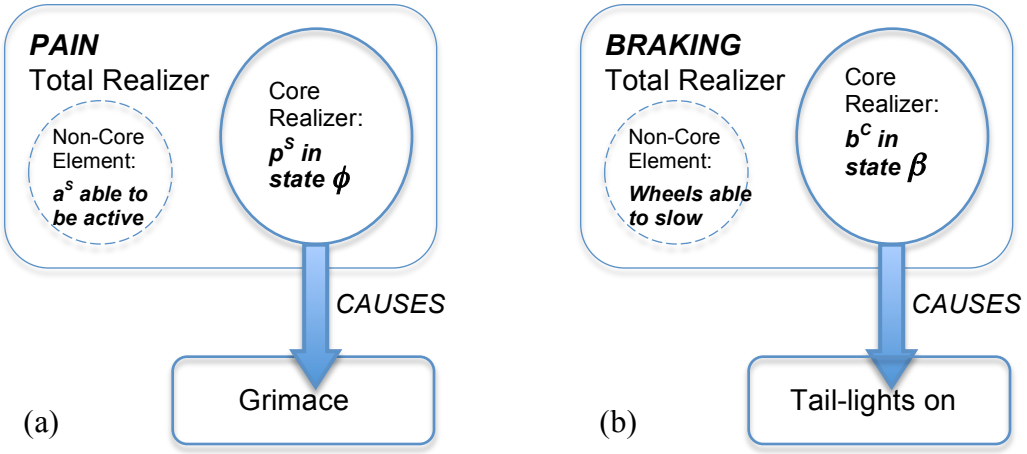

Figure 1. Comparison of Sam's realization of pain with a car's realization of braking.

Now although the disposition for the wheels to slow is constitutive of braking, and although braking normally causes the tail-lights to illuminate, the wheels play no part in causing that illumination. We could remove or incapacitate the wheels without changing the fact that state β causes the tail-lights to come on (Fig. 1b). I want to say that the situation in Sam is analogous (Fig. 1a). *His grimace will be unaffected by the removal or incapacitation of a^S, for a^S plays no part in causing the grimace*. This is by contrast with ϕ, the activation of p^S.

Of course, since the tendency to cause expressions of discomfort is part of pain's functional role (though this is not important to my argument), the grimace no doubt requires the manifestation of *some* disposition that is part of pain's total realizer. But this is no reason to think

that the incapacitation of a^S in particular would have to influence Sam's grimace. For there is no reason why a part of the brain that subserves action-planning should be directly causally linked to facial expression – any more than a car's wheels are directly causally linked to its tail-lights. If this reasoning is doubted in this particular case, we may switch to some other case (involving some other type of experience) where the equivalent premise is acceptable; but I shall proceed on the assumption that the present case is acceptable.

My general claim, from which my second stipulation concerning Sam's case is drawn, is that *there will be some non-core elements of the realizer of a token experience, and some immediate cognitive, affective, or behavioral effects of that experience, which are such that those non-core elements will lack a direct causal link to those cognitive, affective, or behavioral effects*. From this it follows that some case of the sort I am describing must be possible – that is, a case in which at least some of a token experience's actual effects remain unchanged by the canceling of one (or more) of the experience's non-core elements. Hence my stipulation that in the case at hand, some of the actual effects of Sam's pain, in particular his grimace, are unchanged when a^S is incapacitated.

Let me now consider two objections to the general claim that I have just made.

Firstly, someone might simply deny the claim. They might argue that an experience's total realizer must include only states that can influence *all* of the experience's typical effects, and then make an empirical bet that this criterion is met by states that are deemed (*a priori*) to be part of a given experience's total realizer – such as the intactness of a^S.

However, while there may *sometimes* be a direct causal link between a non-core element of the realizer of an experience and a thought, emotion, or behavior caused by that experience (perhaps because the non-core elements physiologically overlap their associated core realizer), so that interfering with the former will inevitably affect the latter, there is no reason to think this will *always* (or even often) be the case. Moreover, the functionalist cannot rely on an empirical bet (cf. Antony p. 110n). Her theory is meant to apply to all nomologically possible sentient creatures; yet it is certainly *possible* that some such creatures will have experiences whose

realizers have non-core elements which have no direct causal link to some of the effects typical of the experience.

Of course, if the disposition in question has actually manifested, *then* eliminating the relevant non-core element *would* have an effect. We would definitely alter Sam's behavior if we incapacitated a^S after he had begun to move toward the medicine cabinet! But not all of the dispositions associated with an experience will manifest every time the experience is tokened, so not all of its non-core elements will be causally active every time. Non-core elements must only support a *disposition* for the subject to think, feel, or act in a certain way; they need not *actually cause* the subject to think, feel, or act in that way every time the mental state is tokened. So unless a given non-core element is one whose supported disposition is actually manifesting, we may alter that element on that occasion without affecting the subject's responses. So even if someone rejects my claim that Sam's grimace is unaffected by the incapacitation of a^S, a similar claim must still be accepted in a wide range of cases – and our argument can be re-run using one of those cases. The case of Sam's pain causing his grimace is merely a representative example.

The second objection to my general claim also starts with the assertion that an experience's total realizer must include only states that can influence *all* of the experience's typical effects; but then it proceeds differently from the first objection. A psychofunctionalist, who is willing to let empirical findings determine the functional roles of mental states, might say that if there turns out to be even *one* effect of Sam's pain (like his grimace) that would be unaffected by a^S's incapacitation, then the intactness of a^S is simply *not* part of the pain's total realization. (By analogy, if the total realizer of braking includes only states that can influence *all* the effects of braking, then states of the wheels are ruled out, for they do not affect the illumination of the tail-lights.) Accordingly, the incapacitation of a^S would be irrelevant to the realization of Sam's pain, for its intactness would not be part of the pain role. Any physical insult that *did* cancel a non-core element of Sam's pain experience would have to influence *any and all* typical effects of that experience – just as would an insult to the core realizer itself. Generalizing the objection: a case

like the one I am describing, involving a physical insult to a supposed non-core element of a pain experience but *without* influence on any of the pain's typical effects, would be impossible.

This cure, however, turns out to be identical to the disease it is meant to target. The proposed restriction on non-core elements is meant to derail my argument against the FNT_E, but it actually just abandons the FNT_E – because it entails that experiences may have *no* functional role. For it might turn out that for some experience *E*, the only state that has influence over *all* of *E*'s typical effects is its core realizer. Then according to the proposed restriction, *E*'s core realizer and its total realizer would be identical. But since an experience's total realizer is sufficient for the experience itself, *E*'s core realizer would be sufficient for *E* itself. Hence *E* would have no *typical* causes and effects. No particular dispositions would be necessary for its realization. For example, pain might be realized simply by (e.g.) c-fiber activation, serving as core and total realizer all in one. Whatever causes and effects *E* had in one subject might not be duplicated in another. In short, *E* would have no functional role. Yet the FNT_E asserts the nomological necessity of functional roles for experience.

The problem I am identifying with this second objection, I emphasize, is not that the proposed restriction on the total realizers of experiences entails that there *actually will* be experiences whose core realizers exhaust their total realizers. (Though I suspect that the restriction does indeed have that entailment.) It is just that the restriction entails that such a case is *possible*. So the objection is self-defeating, because it denies the FNT_E.

This section has described the adapted version of the case of Sam that I need for my adapted argument, and defended the possibility of that version. I now present the argument itself.

5. The main argument

We are assuming that functionalists take the intactness of a^S to be part of the total realizer of Sam's pain experience. So they hold that Sam could not be *in pain* if a^S were incapacitated, much as a car could not be *braking* if its wheels were immobilized.

This functionalist view does not sit well with the Constraint on Experiential Change. Sam's pain is causing him to grimace. If his pain were to cease, or even just reduce in intensity, the Constraint says that this would cause an immediate change in his cognitive, affective, or behavioral dispositions. It is very likely, for instance, that his grimace would immediately cease or change in some way. Certainly, assuming that his new dispositions *could* be manifested, such a change would be (nomologically) *possible*. We would need an excellent reason to believe otherwise. Yet Sam's grimace *cannot* immediately cease or change as a result of the incapacitation of a^S, for a^S has no direct role in causing the grimace. Only via some indirect path could the incapacitation affect Sam's grimace (say, if it were replaced by a puzzled frown when he finds himself unable to remember how to get to his medicine cabinet, due to the incapacitation of a^S). Hence, I argue, incapacitating a^S cannot immediately affect his pain experience, for if it did then it would be possible for his grimace to change right away. So the intactness of a^S is not necessary for Sam to be in pain; it is not part of his pain's total realizer.

I do *not* deny that experiences possess associated dispositions that on many instances do not *manifest*. What I deny is the FNT_E, which implies that *the dispositions themselves* are necessary for the experience. I argue that since the presence or absence of those dispositions generally *cannot* causally influence an experience's cognitive, affective, or behavioral effects, they are not necessary for the experience itself. Here, then, is the argument:

(1) Incapacitating a^S cannot immediately cause Sam to stop grimacing.

(2) Eliminating Sam's pain experience very likely would immediately cause him to stop grimacing.

Therefore,

(3) Incapacitating a^S cannot eliminate Sam's pain experience.

Therefore,

(4) Sam remains in pain after a^S is incapacitated.

Premise (1) was defended in §4. So I now turn to (2), which is supported by the Constraint.

6. Defense of premise (2)

The functionalist may argue that eliminating Sam's pain experience would *not* necessarily cause him immediately to stop grimacing, and hence that, contrary to (4), he might no longer be in pain after t_2 – even though he continues to grimace as if he is.

There are two positions available regarding how the grimace might be maintained after t_2 in the absence of any pain experience. The first is that the grimace was actually not being caused by the pain experience even before t_2, but by some other state, which remains after t_2 even though the pain is gone. The second is that before t_2 the grimace was caused by pain, but after t_2 it is caused by some other, newly-obtaining state.

The first option threatens to lead to the result that pain is epiphenomenal. Since our argument is not specific to pain's causing of a *grimace*, one cannot escape it by just accepting that pain does not cause grimaces. Indeed, the argument is not even specific to the experience of pain, so nor can one escape it by accepting that pain lacks all causal powers. One would have to accept epiphenomenalism for all experiences generally. But functionalists are not amongst those who find epiphenomenalism palatable. For functionalists, after all, mental states are constituted by their causal powers. If experiences do not *have* causal powers then the objection mentioned by Levin (2010; see my §1) applies in spades: there cannot be a functional account of experience if experiences do not have functional roles!

So functionalists will surely prefer the second option. By analogy, there are ordinary situations in which a grimace of pain might continue even if the pain itself ceases. A person whose toothache is now actually gone might keep grimacing out of a desire for sympathy, for example. The functionalist may suggest that something schematically similar happens to Sam at t_2. Call the state brought about by the incapacitation of a^S 'schmain'. Before t_2, Sam was in pain. But with the incapacitation, the cause of his grimace switches from pain to schmain. This story avoids epiphenomenalism, for Sam's pain *was* causing his grimace before t_2.

On this proposal the functionalist is holding that schmain, while similar to pain in its realization, is qualitatively distinct; it is not *painful*. It may even be an entirely non-experiential state. After t_2, therefore, Sam is no longer in pain.[5]

If this is so, however, why is he still grimacing? The Constraint implies that if pain and schmain were qualitatively distinct, it would be *possible* (and remember, we are dealing with nomological possibility) for them to have distinct effects. One of those effects would be on his grimace. (Remember, though, that Sam may also continue to evince pain in other ways, such as by *saying* that he is in pain. The grimace is just one example.) So if his grimace cannot *possibly* subside, we ought to conclude that Sam is still in pain after t_2.

The functionalist may object that we cannot apply the Constraint in this unusual context. Perhaps eliminating an experience *in that way* would *not* affect the subject's thought, emotion, or behavior. After all, we lack empirical evidence about such cases, and the cognitive neuroscience literature is littered with novel brain lesions that had highly unexpected effects.

Sam's situation is indeed novel, so our judgment should be cautious. But if anyone is being *in*cautious, it is the functionalist. Her current proposal entails a remarkably bold empirical conjecture. Let me explain.

Remember first that Sam was not special. The functionalist must say that if we performed the same procedure on hundreds of subjects, *none* would cease grimacing when we incapacitated their brain's action-planning region, despite the supposed fact that none are in pain after that time.

---

[5] The functionalist may say that the incapacitation of a^S would not make Sam's experience *non*-painful, but rather a *different kind* of pain. But in the final analysis this will not change the outcome of my argument. Firstly, as I shall explain, the main problem is the sheer implausibility of the idea that there could be a change in Sam's experience with *no possibility* of his thoughts or behavior changing; and the implausibility of that idea is not much reduced if the experiential change is from one kind of pain to another, rather than from pain to non-pain. (Wouldn't Sam be able to say that his pain experience had altered?) Secondly, there must be some set of non-core elements whose incapacitation *would* eliminate his pain entirely. So we may simply re-run the argument on *that* incapacitation.

More generally still, the functionalist's current proposal applies to all types of experience (nociceptive, visual, tactile, etc). Her general claim runs as follows. Given a token experience which is causing a certain cognitive, affective, or behavioral response, that experience can be terminated by canceling one or more of its non-core elements (by incapacitating some part of the brain that supports those elements).

Now, with that termination, one of two things must happen. Either (i) the response will immediately change (due to some direct causal link to the canceled element(s)); or (ii) the response will be unchanged.

I noted in §4 that (i) may *sometimes* be the case, but that the functionalist cannot always or even often rely on it, for there is no reason to think that such direct causal links (e.g., between a^S and Sam's facial expression) are omnipresent or even common. So (ii) would have to be the case at least sometimes, and probably most of the time. Yet the idea that (ii) would be a common outcome of the imagined intervention is a remarkably bold empirical conjecture. It will not suffice just to say that the response (e.g., the grimace) will be unchanged because the canceled non-core element (e.g., a^S) was not causally connected to it. For if that were the case, we would also expect – per the Constraint – that the cancellation would not change the *experience* (e.g., we would expect Sam to still be in pain). But the functionalist is claiming that canceling the non-core element *does* change the experience. So we need an explanation of how the cancellation can do that *without* affecting the response. That is, the explanation for (ii) must explain both the unchanged response *and* the changed experience.

Consider an example of the sort of speculative hypothesis that would explain how (ii) could be true in Sam's sort of case. Perhaps the neural circuits that subserve action-planning play a role in *mediating* facial response, so that their incapacitation creates a dissociation which prevents a facial response to changes in nociceptive experience. This *could* be true. Yet the functionalist cannot stop with such a limited speculation. She must say that except for occasional cases in which (i) holds, a similar dissociation would be created by *every* incapacitation of a non-core element of *any* experience's realizer in *any* subject – so that in *no* such case could the subject's

thought, emotion, or behavior be altered, despite the experiential change. The sort of hypothesis that the functionalist would need is thus extremely broad: for example, that these sorts of cortical lesions *always* cause subjects to become unable to attend to their current experience (so that they do not immediately respond to the experiential change). I urge that such a hypothesis does not deserve to be taken seriously unless it has significant empirical motivation.

The functionalist's bet, then, is that the close causal tie between experience and the rest of our mental economy can be severed under a wide range of (hypothetical) circumstances involving precise and limited brain lesions, though she offers no explanation of how this severing would occur. By contrast, my bet is that in those circumstances, the causal tie would remain intact – simply because we know that it exists in normal cases, and because there is insufficient reason to suppose that it is severed in the cases under discussion. In the absence of a specific reason not to do so, I submit that we should infer from the general evidence supporting the Constraint to the conclusion that Sam remains in pain after t_2. And that entails that the FNT_E is false.

7. Summary and conclusion

I have endeavored to complete the anti-FNT_E argument that was first offered by Antony (1994) by arguing that the FNT_E violates a principle that we very justifiably believe about conscious experiences to which one is attending. That principle is that when one's attended experience changes, one's immediate cognitive, affective, or behavioral dispositions also change. The FNT_E, by contrast, entails that there can be changes in your attended conscious experience which *cannot* affect those dispositions. I have argued that this is reason to reject the FNT_E.

Let me remind the reader that my argument shows that functional states are not *nomologically* necessary for experience. So the conclusion is not just that functionalism fails to solve the 'hard problem' (Chalmers 1995) of explaining how experience could arise from mere physical stuff. Some philosophers might hope that even if functionalism cannot solve that problem, it can still identify which states are, as a matter of fact, those in virtue of which *we* have experiences.

Chalmers himself promotes just such a 'non-reductive' functionalism. But my conclusion is that our experiences do not occur in virtue of our being in certain functional states, for the former can remain in the absence of the latter. So functionalism – and indeed any theory that adheres to the FNT_E, which includes a number of theories that are not overtly functionalist, such as many representational theories – fails even as a non-reductive theory of experience.

Acknowledgements

For helpful discussion of the arguments in this paper, I wish to thank Brian McLaughlin, Tim Maudlin, Gene Witmer (who commented on an earlier version presented at the Southern Society for Philosophy & Psychology in 2009), two anonymous referees for this journal, and – last but not least – Michael Antony.

References

Antony, M. V. (1994) Against functionalist theories of consciousness, *Mind & Language*, **9** (2), pp. 105-123.

Bartlett, G. (2012) Computational theories of experience: between a rock and a hard place, *Erkenntnis*, **76** (2), pp. 195-209.

Block, N. (1978) Troubles with functionalism, in C. W. Savage (ed.) *Minnesota Studies in the Philosophy of Science, Vol. 9: Perception and Cognition*, Minneapolis: University of Minnesota Press.

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

Carruthers, P. (2005) *Consciousness: Essays From a Higher-order Perspective*, Oxford: Oxford University Press.

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

Fuster, J. M. (2008) *The Prefrontal Cortex* (4th ed.), London, UK/ Burlington, MA / San Diego, CA: Academic Press.

Levin, J. (2010) Functionalism, in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Summer 2010 Edition)*, [Online], http://plato.stanford.edu/archives/sum2010/entries/functionalism/ [10 Apr 2013].

Libet, B. (1985) Unconscious cerebral initiative and the role of conscious will in voluntary action, *Behavioral and Brain Sciences*, **8** (4), pp. 529-539.

Maudlin, T. (1989) Computation and consciousness, *The Journal of Philosophy*, **86** (8), pp. 407-432.

Milner, D. & Goodale, M. A. (2009) Visual streams: what vs how, in T. Bayne, A. Cleeremans, & P. Wilken (eds.) *The Oxford Companion to Consciousness*, Oxford: Oxford University Press.

Morsella, E. (2005) The function of phenomenal states: supramodular interaction theory, *Psychological Review*, **112** (4), pp. 1000-1021.

Prinz, J. J. (2012) *The Conscious Brain: How Attention Engenders Experience*, Oxford / New York: Oxford University Press.

Putnam, H. (1963) Brains and behaviour, in R. J. Butler (ed.), *Analytical Philosophy: Second Series*, Oxford: Basil Blackwell.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997) To see or not to see: the need for attention to perceive changes in scenes, *Psychological Science*, **8** (5), pp. 368-373.

Rey, G. (1997) *Contemporary Philosophy of Mind: A Contentiously Classical Approach*, Oxford: Blackwell.

Shoemaker, S. (1981) Some varieties of functionalism, *Philosophical Topics*, **12** (1), pp. 93-119.

Simons, D., & Chabris, C. (1999) Gorillas in our midst: sustained inattentional blindness for dynamic events, *Perception*, **28** (9), pp. 1059-1074.

Tracey, I., & Mantyh, P. W. (2007) The cerebral signature for pain perception and its modulation, *Neuron*, **55** (3), pp. 377-391.

Tye, M. (1995) *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*, Cambridge, MA: MIT Press.

Tye, M. (2000) *Consciousness, Color, and Content*, Cambridge, MA: MIT Press.

Van Gulick, R. (1993) Understanding the phenomenal mind: are we all just armadillos? in M. Davies & G. Humphreys (eds.), *Consciousness: Psychological and Philosophical Essays*, Cambridge, MA: Blackwell.

Van Gulick, R. (2004) Higher-order global states (HOGS): an alternative higher-order model of consciousness, in R. Gennaro (ed.), *Higher-order Theories of Consciousness: An Anthology*, Amsterdam: John Benjamins.

Wegner, D. M. (2002) *The Illusion of Conscious Will*, Cambridge, MA: MIT Press.