Erkenntnis (2020) 85:693–714
https://doi.org/10.1007/s10670-018-0044-2

ORIGINAL RESEARCH

No Pain, No Gain (in Darwinian Fitness): A Representational Account of Affective Experience

Benjamin Kozuch[1]

Received: 11 November 2016 / Accepted: 10 July 2018 / Published online: 24 July 2018

Abstract
Reductive representationalist theories of consciousness are yet to produce a satisfying account of pain's affective component, the part that makes it pain*ful*. The paramount problem here is that that there seems to be no suitable candidate for *what* affective experience represents. This article suggests that affective experience represents the Darwinian fitness effects of events (roughly, the effects that an event has on a creature's chances of propagating its genes). I argue that, because of affective experience's close association with motivation, natural selection will work to bring affect into covariance with the average fitness effects of types of event, and that this covariance makes fitness effects a promising candidate for what affect represents. I also argue that this account is to be preferred to Cutter and Tye's recent proposal that affect represents harmfulness, and answer an objection that Aydede and Fulkerson recently offered against representational accounts of affect.

1 Introduction

By mid-last century, materialism had become the default view among philosophers and psychologists. Fifty-odd years later, and there still is no widely accepted materialist account of consciousness. Nonetheless, there is one approach to naturalizing consciousness showing promise, this being *reductive representationalism* (Harman 1990; Dretske 1995; Tye 1995; Clark 2000). The strategy is a two-step reduction: First, all experiential properties are reduced to representational properties, then these representational properties are reduced to physical properties. It has often been argued that the first step of the reduction must fail, since

✉ Benjamin Kozuch
bkozuch@ua.edu

[1] Philosophy Department, College of Arts and Sciences, The University of Alabama, 336 Ten Hoor Hall, Tuscaloosa, AL 35487-0218, USA

experiential properties, in general, cannot be reduced to representational properties (see, e.g., Block 1990, 1996). But some have claimed that even if these wide-ranging arguments were defeated, there are extra barriers to be overcome in the case of certain kinds of experience, one of the more prominent ones being *pain*.[1]

Consider that, for any type of experience to be comprehensively reduced, it must be the case that the phenomenal character of that experience is *exhausted* by whatever representational content it has. This is not obviously the case when it comes to pain. True, certain aspects of it look representational. Pains, for instance, seem to be represented as having a location and shape. However, pain experience also has an *affective* component, this being the part that makes it *hurt*, that makes it pain*ful*. Some have doubted that this affective component is representational (Aydede 2005; Aydede and Fulkerson 2014; cf. Klein 2015), the biggest problem here being that it is not clear *what* it represents: According to reductive representationalism, representation consists of some kind of covariance (or "tracking")[2] relation. But what suitably naturalistic and objective property is it that affective experience covaries with?

In this article, I offer a candidate. Here is the idea: Affective experience (I will often just say "affect"), of both positive and negative type, is closely connected to motivation, in that when an event causes positive or negative affect, it motivates one to repeat or avoid (respectively) events of the same type in the future. Because of this connection, an evolutionary advantage accrues to any individual whose affective experience is attuned to the Darwinian fitness effects of types of event. This is just to say that an individual having a *negative* affective response to injury is better at passing on its genes (has more "fitness") than one having a *positive* affective response, since only the latter is motivated to repeat events of that type. Given this, natural selection will work to shape affect so that it covaries with the average fitness effects of whatever event-type causes it. I argue that there being this selective pressure makes fitness effects a good candidate for what affect represents, and that the reduction of affect therefore poses no special problem for representationalism.

Here is the plan. In Sect. 2, I describe why affect is thought to be problematic for reductive representationalism. I also consider Cutter and Tye's (2011) proposal that affect represents aptness to harm, finding it to lack the right kind of covariance with affect. In Sect. 3, I argue that fitness effects *do* have the right kind of covariance, and therefore are a promising candidate for what affect represents. In Sect. 4, I answer an objection due to Aydede and Fulkerson (2014), one that probably applies to any account of affect given in terms of a tracking relation. I also revisit Cutter and Tye's theory, explaining why the account offered here should be preferred to theirs.

[1] Henceforth, I use "representationalism" and "reductive representationalism" equivalently.

[2] The terms "tracking" and "covariance" are used equivalently in this article.

2 Reductive Representationalism and Affective Experience

In the first step of the reductive representationalist strategy, experiential properties are reduced to representational properties. If reductive representationalism is to retain its attraction as a materialist theory of the mind, said reduction should be comprehensive: It should be that *all* experiential properties turn out to be representational properties. Many types of perceptual experience accommodate the representationalist strategy relatively well: Experiences such as seeing or hearing seem to have accuracy conditions, a hallmark of representation (Siewart 1998, Chap. 7). It is because of these accuracy conditions that an experience as of greenness can be called illusory if nothing green is before the subject. There is, moreover, a case to be made for perceptual experiences such as seeing or hearing being *exhausted* by their representational content. The claim typically made here is that experience is *transparent*, in that any attempt to introspect properties of an experience returns only those properties that the experience represents objects to have (Harman 1990; Lycan 2001; Tye 2002). If a subject, for example, introspects her experience as of a red apple, the only redness to be found belongs to the apple, with no redness belonging to the experience itself. So the idea that paradigmatic perceptual experiences like seeing or hearing are entirely representational has at least initial plausibility (but see Block 1996).

Not so with "bodily sensations," things like tickles, itches, and orgasms. It is questionable whether such experiences are assessable for accuracy, or if they are transparent the way that visual or auditory experiences seem to be. (Can a tickle be falsidical? What does an orgasm represent?) The representationalist has been especially pressed to come up with an account of pain experience (McGinn 1982; Block 1996, 2005; Aydede 2005, 2009). To be sure, certain aspects of pain seem representational. Pains appear to be represented as having a volume, shape, and location (e.g., a small, prickly pain in one's leg), and temporal characteristics (e.g., persistent, pulsating). However, these aspects of pain experience all belong to what pain researchers refer to as its *sensory* component. But pain also has an *affective* component, this being its awful, unwanted quality, the part that makes it pain*ful*.[3] The problem here is that it is not clear what this affective component might represent.[4]

Consider how for many types of experience, it is not difficult to produce a candidate for what it represents: A visual experience as of something being cube-shaped, for instance, plausibly represents some (equilateral) geometric property; a tactile experience as of something being bumpy plausibly represents some (uneven) textural property; and so on (cf. Aydede and Fulkerson 1962 forthcoming). But in

[3] That pain has both sensory and affective components is something vividly demonstrated in the effects of a *cingulotomy* (removal of the anterior cingulate cortex), an operation performed on patients with chronic, excruciating pain. After the operation, patients say that they still feel the pain (i.e., they have the sensory component), but that they do not mind it (i.e., they lack the affective component) (Damasio 1994, 1999). In like fashion, subjects under the influence of morphine rate the affective component of their pain as diminished, but not the sensory component (Kupers et al. 1991). Other experiments demonstrate similar dissociations (Rainville et al. 1997, 1999; Hofbauer et al. 2001).

[4] The representationalist can deny that affect has any phenomenal character, relieving him of the burden of explaining it. Tye took this approach in the past (1995, Tye 1997; cf. Armstrong 1962; Pitcher 1970), but appears to have now abandoned it.

the case of pain's affective component—the part of pain experience that *hurts*—finding a candidate is not so easy. What property is it that bodily events as diverse as sore throats, pulled muscles, and lacerations all have in common, that it could be what affect represents? Indeed, some philosophers argue that there is no good candidate, meaning affect is probably not representational (Aydede 2005; Block 2005; Aydede and Fulkerson 2014, forthcoming; cf. Klein 2015, Chap. 3).

This is not to say that no candidates have been offered—we look at one shortly. But let us first discuss what is desired in such a candidate, if it is to fit comfortably in a reductive representationalist theory. First, the candidate property must be *objective* (i.e., mind-independent), lest the reduction becomes viciously circular (Tye 1995, Chap. 5; Lycan 2001). This is why representationalists bend over backwards trying to explain how colors are objective (Tye 2000, Chap. 7), a long tradition of color irrealism among philosophers and scientists notwithstanding. Second, the property should be *naturalistic*. Unfortunately, what makes a property naturalistic is no settled matter. So as to not bog down, I adopt a criterion that harmonizes with one popular way of understanding naturalism (Quine 1981, Chap. 1), saying that a property is naturalistic iff it plays a part in our final science. Finally, the candidate property should *covary* with affective experience, in a certain way. Representationalists typically favor some kind of *tracking* theory of representation, according to which (stating it generically) S represents P iff S covaries with P under optimal conditions (Tye 1995, Chap. 4; 2002, Chap. 6; Dretske 1988, Chap. 3; 1995 Chap. 1; cf. Stampe 1977; Stalnaker 1984).[5,6] Optimal conditions are ones conducive to the representational system fulfilling its function, e.g., ones in which there are no environmental aberrations, and the representational system works as it should. It is particularly important that we mind this "under optimal conditions" clause: It means that some property P might covary poorly with affect and yet still be a good candidate for what affect represents, since abductive reasoning about the available evidence might nonetheless point to P being what affect *would* covary with, under optimal conditions. Summing up, the challenge reductive representationalism faces when it comes to affect is to produce an objective, naturalistic property that covaries with affect under optimal conditions. I refer to this as the *Covariation Problem*. The primary goal of this article is to solve the Covariation Problem.

Now, it is true that providing an account of affect in terms of a tracking relation is not the only option. Receiving much attention lately are theories where affect is taken to be some kind of *imperative* content, one which says, roughly, "Make this stop!" (Klein 2007, 2015; Martínez 2011; cf. Hall 2008). However, it has been

[5] Tracking theories of representation typically also include a *causal* clause, so that S represents P iff S is what *causes* P under optimal conditions (see Tye's formulation in the next footnote; cf. Fodor 1990).

[6] According to Tye's theory of representation, "S represents that P = df If optimal conditions were to obtain, S would be tokened in c iff P were the case; moreover, in these circumstances, S would be tokened in c because P is the case" (2000: 136). According to Dretske's theory of representation, some state S represents property P iff (a) S and only S covaries with P, and (b) it is the function of S to act as an indicator of P. I take Dretske's appeal to the "function" of S to mean that we would expect S to covary with P under optimal conditions (since this is when it would successfully fulfill its function), and so we can consider Dretske's theory also to be captured by the generic formulation of a tracking theory I just gave.

argued that imperative theories cannot account for *positive* affect (i.e., pleasure), or the fact that affect comes in degrees (Cutter and Tye 2011). Whatever these arguments' merit, I take it that an account of affect appealing only to a tracking theory of representation (i.e., one not resorting to imperative content)[7] would possess significant theoretical virtue. Thus it is worthwhile to see whether such an account can work.

So far, there have been few attempts to solve the Covariation Problem. The most well-developed is due to Cutter and Tye (2011; see also Tye 2005; Bain 2013), who argue that affect represents something *about* a "bodily disturbance" (e.g., a cut); namely, it represents the bodily disturbance as being *apt to harm*, where this harm is indexed to the subject experiencing the affect (i.e., apt to harm *me*).[8] One might worry whether a suitably naturalistic account could be given of "aptness to harm," but Cutter and Tye (hereafter C&T) believe that an appeal to teleology would do the trick,[9] and I will not press the point. Nonetheless, C&T's theory seems to have loose ends.

To see how, let us start by asking how we should understand "harm." Perhaps it is *physiological* harm. This makes sense, given pain's association with tissue damage. However, something seems not quite right about the covariational relationship between affect and physiological harm. The act of being socially excluded poses no physiological threat to a person (it involves no "bodily disturbance"), but nonetheless can cause strongly negative affect.[10] Consider too that C&T's theory is meant to also account for *pleasure*, i.e., *positive* affective experience, where this represents bodily events as "apt to benefit." But an orgasm causes strongly positive affect, and yet is of no known extraordinary physiological benefit.[11] Also problematic is the idea that the aptness to harm is indexed to the subject: Seeing one's offspring severely injured causes strongly negative affect, and yet is of no direct physiological threat to the subject. Overall, the covariational relationship between affect and physiological harm/benefit is not what we would expect, were the former to represent the latter.[12]

[7] Or to something like Millikan's pushmi-pullyu states (Millikan 1995), a primitive type of intentional content described as having both indicative and imperative content.

[8] Bain (2013) agrees with Cutter and Tye insofar as he believes that affect represents something about a bodily disturbance, namely that it is bad for the person undergoing it; he also accepts *aptness to harm* as a good candidate for what is meant by "bad" here. To this extent, Bain's account is susceptible to the same criticisms I will make about Cutter and Tye's view.

[9] Write Cutter and Tye: "We can understand the notion of harm in relation to the notion of a teleological system. Very roughly, something harms a teleological system to the extent that it hinders that system (or one of its subsystems) from performing its function(s)" (2011: 99–100).

[10] Empirical support for social exclusion causing negative affect can be found in Eisenberger et al. 2003. In this study, subjects were put in a computer-simulated game of catch in which none of the other virtual participants would throw the ball to them.

[11] Though orgasms might have modest physiological benefits such as temporarily boosting one's immune system (Haake et al. 2004) or ability to tolerate pain (Whipple and Komisaruk 1988), no literature supports the idea that they dramatically affect one's health.

[12] One might object that C&T's account is only meant to apply to *bodily* pain, and not the types of *emotional* pain being described here. There are, however, reasons to think that the affect accompanying bodily and emotional pain are fundamentally the same (see fn. 14), and therefore deserve a common explanation. At very least, an account providing a common explanation (like the one I offer below) should be preferred to one not doing so, *ceteris paribus*.

Responses could be made here on C&T's behalf. One could argue, for instance, that there *is* physiological harm or benefit in the above cases, since such events might raise or lower one's stress levels. Or it could be claimed that "harm" should be construed so as to include *psychological* harm, since events like seeing one's offspring grievously injured plausibly cause this. But what is probably most important to consider here is something stressed above, which is that these failures of covariation alone cannot convict C&T's theory, since they are consistent with affect covarying with aptness to harm or benefit *under optimal conditions*. However, a more advantageous time to evaluate these considerations comes after some conceptual resources have been developed, and so I postpone discussion of them until 4.2. My present goal is only to show that it is unclear whether C&T have solved the Covariation Problem, so as to motivate consideration of alternatives. Offering one alternative is what I do next.

3 Solving the Covariation Problem

Some features of affect should be pointed out, as they play a central role in my argument.

The discussion above focused mostly on *negative* affect, the kind associated with pain. But also mentioned was *positive* affect, the kind associated with pleasure. The pleasurable feelings experienced during a back rub, for example, not only have sensory dimensions (things like location and shape), but also an affective dimension, this being what makes it *feel good* (in that location, in that shape). So affect can be positive or negative. Let us refer to this attribute of affect as its *polarity*.

Something else to note about affective experience is that each instance of it has a certain *intensity*. An affective experience can be anywhere from barely noticeable to overwhelming: Being jabbed with a finger causes negative affect of low intensity; being stabbed with a knife causes negative affect of high intensity. Parallel remarks could be made about positive affect.

Here, then, is some terminology adopted in this article: First, there is the *polarity* of an affective experience, which refers to whether it feels *good* (positive) or *bad* (negative); then there is the *intensity* of an affective experience, which refers to *how* good or bad it feels.[13]

Something else to note about about affective experience is that not all instances of it involve overt bodily occurrences, such as cuts or caresses. Events that also cause positive or negative affect include receiving a smile or frown, witnessing an act of generosity or cruelty, or hearing a joke or insult.[14] In addition, affective

[13] I stress that the terms "polarity" and "intensity" are technical terms created for use specifically in this article; thus these words' meanings should not be inferred from some of the ways in which they have been used in emotion research (e.g., the way "polarity" has been used in the debate over whether affect is "independent" or "bipolar"; see fn. 15).

[14] It might be doubted whether bodily and non-bodily (or "emotional") affective experience should be categorized together, but see Helm (2002), Eisenberger and Lieberman (2004), Prinz (2010) and Corns (2014). In any event, my argument does not ultimately depend on bodily and non-bodily affective experience being the same (whatever that amounts to), only their being closely connected to motivation, something to be shown next.

experience might be complex, simultaneously involving affect of both polarities;[15] an example of this might be a prolonged run while downhill skiing, where one feels both a burning sensation in her legs, and the thrill of the descent.[16]

So, each affective experience has a polarity and an intensity, and not all affective experiences involve overt bodily occurrences. Having pointed these things out, I can now present my proposal.

3.1 Affective Experience and Motivation

A view prevalent in both philosophy and psychology is that the affective component of pain experience is closely associated with motivation.[17] The observation usually made is that negative affect is accompanied by an urge to terminate the event causing it (e.g., Klein 2007; Bain 2013). For instance, if one stands close enough to a fire such that it becomes uncomfortable, the negative affect motivates the person to move away. But notice that negative affect frequently does more than this, since it also often creates motivation to avoid future encounters with the type of event in question: One will be less likely to stand that close to comparable fires henceforth. This is not to say that some event causing negative affect *guarantees* that one terminates the event in question, or that one avoids events of the same type in the future; one might have contrary motivations greater than the motivation created by the negative affect (Bain 2013). For example, one might voluntarily stand jacketless in the cold if there is someone with whom they greatly wish to speak; similarly, a sprinter might frequently undergo the pain of rigorous training because of a strong desire to reach the Olympics. However, because of negative affect's close link with motivation, one would probably never carry out these actions in absence of contrary motivations like those just mentioned.[18] Parallel observations could be made about *positive* affect and motivation.[19]

Note now that motivation possesses properties analogous to the polarity and intensity possessed by an affective experience: Each instance of motivation has a

[15] The view that positive and negative affect can occur simultaneously in the same subject—that is, that they are (as they say in emotion research) "independent" of one another (see, e.g., Watson and Tellegen 1985)—is not entirely uncontroversial. According to the opposing view, positive and negative affect are bipolar opposites, with an occurrence of one entailing an absence of the other (see, e.g., Russell and Carroll 1999). (For review of the debate, see Colombetti 2005). I lack space for entering the debate here, but it is worth pointing out that the bipolar model fails to capture cases like the one just discussed (the downhill skier), suggesting that the bipolar model is—at best—only able to accurately describe affective states like *moods*.

[16] Indeed, sometimes the *same event* might cause both positive and negative affect, such as when muscle soreness brought on by strenuous exercise feels good or satisfying in some way.

[17] Philosophers holding this view include (Helm 2002; O'Sullivan and Robert 2012; Bain 2013; Cohen and Fulkerson 2014; Corns 2014; Aydede and Fulkerson 1962). Psychologists holding this view include (Melzack and Casey 1968; Leventhal 1993; Berridge 2004; Leknes and Tracey 2008).

[18] Note that I am not claiming that affective experience is the *only* source of motivation; the sprinter, for instance, might have a standing urge to follow in his Olympian father's footsteps.

[19] Corns distinguishes between "motivation" and "motivational oomph", where they are both constituted by a drive toward or away from some object, but only the former is a "cognized, intentional state" that requires "goal-directedness and flexible, means-end reasoning" (2014: 245). My use of the term "motivation" encompasses both motivation and motivational oomph.

polarity, in that it either causes one to seek an event-type (*positive* motivation), or to avoid an event-type (*negative* motivation). And each instance of motivation has an *intensity*, since one can be motivated to seek or avoid an event-type in varying degrees. A strong urge to get a milkshake, for instance, consists of strongly positive motivation, and a mild reluctance to get off the couch consists of weakly negative motivation.

The thesis that I now advance is that, in cases where affect causes motivation (and when the affective/motivational systems work as they should), the affective experience and the ensuing motivation tend to covary in their intensity and polarity. More precisely:

(Affect-Motivation Symmetry)

If subject S has affective experience E, and event-type T is recognized as a cause of E, then:

(a) Whether S gains motivation to avoid or pursue T is determined by the polarity of E
(b) The intensity of the motivation that S gains to avoid or pursue T is determined by the intensity of E

The thesis of Affect-Motivation Symmetry is nothing too novel. It shares much with the behaviorist's Law of Effect (Thorndike 1913), which says that if a response made by a subject has a "satisfying" or "discomfiting" effect, it will raise or lower (respectively) the likelihood of that response being repeated in the future. But we need no law of psychology to tell us that this kind of relationship obtains between affect and motivation; we need only consider everyday examples: If a person suffers a painful cut when using a knife, we expect that person to be more careful with knives in the future. And if a child enjoys her first taste of candy, we expect that child to seek more. Such observations provide reason to think that the polarity of an affective experience determines the polarity of the ensuing motivation. Similar observations show the same kind of relationship to obtain between the *intensity* of the affective experience and the ensuing motivation: We would expect a person whose action results in a hard knock to the head to be more strongly disinclined to repeat this action than one whose action resulted in a stepped-upon toe. This, of course, is because the first creates stronger negative affect than the second. We would, furthermore, expect the difference in motivation that the two events create to be proportional to the difference in the intensity of affect that the two events caused.[20,21]

[20] One interesting issue arising here—not just for the present theory, but for any theory hypothesizing a link between affect and motivation—concerns *which* of the multiple event-types in the vicinity of an affective experience the subject will become motivated to avoid or pursue. Space constraints dictate that this issue be left for future research.

[21] Note that the motivation that affect creates seems to be general, in that it *potentially* motivates one to do *any* action that might prevent or promote the type of event that caused the affect. For instance, the negative affect caused by a beating brings about a general motivation to stop the beating, leading to one being potentially motivated to fight back, run away, or do whatever else might end the beating. Interestingly, affect can motivate one to prevent events no longer possible, such as when one ruminates over how one could have prevented the death of her child.

Some remarks and clarifications: First, affect is not only able to create motivation to avoid or pursue the event that caused affect, but in some cases might motivate one to act to add or remove the affect itself; a situation like this might be when one takes aspirin for a headache. Second, the scenario that I am describing here is not one in which affect is *constituted* by motivation (the view held by Bain 2013; Cohen and Fulkerson 2014; Aydede and Fulkerson 1962), but rather is what frequently *causes* motivation (Berridge 2004; Corns 2014). While the constitutive view is thus far more popular, affect and motivation dissociate often enough to justify thinking that they are distinct (Corns 2014). Addicts, for instance, might be highly motivated to ingest something that they nonetheless experience as unpleasant (ibid.), and some experiments have been performed in which a reward stimulus produced increased motivation, but not affect, in genetically altered mice (Peciña et al. 2003; see Berridge 2004 for review). Now, it might be thought that such dissociations can also be used to undermine Affect-Motivation Symmetry, since they show affect to not ineluctably lead to motivation. But Affect-Motivation Symmetry is only a thesis about how the two relate when affective-motivational systems operate as they should, not when they malfunction due to aberrant circumstances (drug addiction, genetic manipulation). In addition, the overall argument I am making only requires Affect-Motivation Symmetry to obtain with *some* reliability, this being enough to make affective experience subject to the kind of evolutionary selective pressure described next.

Affect-Motivation Symmetry provides the first piece needed to solve the Covariation Problem. Now we look at the second.

3.2 Affective Experience and Fitness Effects

Affect shares a close association not only with motivation, but also emotion (see, e.g., Colombetti 2005; esp. Charland 2005). Indeed, it is questionable whether a psychological state feeling neither good nor bad could rightly be considered an emotion. Perhaps there is something to learn by considering affect in context of emotion.

Evolutionary accounts of emotions are increasingly popular. Here, emotions are thought of as modes of psychological and physiological function evolved by natural selection, designed to produce adaptive behavior in recurring situations of evolutionary importance (Nesse 1990; Cosmides and Tooby 2000; Nesse and Ellsworth 2009; see also Trivers 1971). The emotion of fear, for instance, happens in response to danger (e.g., the presence of a predator), manifests as a state of readiness for action, both physiologically (increased heart rate, circulation, and muscle tone) and psychologically (rapidity of thought, a focus on the threat and means of escape), all of which helps the agent evade or defeat the threat.

One way emotions increase the prospects for survival and reproduction is by teaching one to pursue positive and avoid negative outcomes (Cosmides and Tooby 2000). For example, the fear caused when confronted by a predator in an enclosed space not only aids one's escape, but also teaches one to avoid similar areas in the future. Given affect's close association with motivation, it is plausible that emotion facilitates this kind of learning largely in virtue of its affective component (e.g., it is

because the fear that one experienced when cornered by the predator was laden with negative affect that one avoids visiting places like that again). The question to ask now is: How precisely would affect most effectively promote this kind of learning?

It is in the answer to this question that a solution to the Covariation Problem lies, since it seems that affect best promotes this kind of adaptive learning by being *tuned* to the evolutionary importance of events, the polarity of affect being positive or negative according to whether the event causing the affect helped or hurt one's prospects for survival and reproduction, and its intensity matching *the degree to which* that event helped or hurt. Put another way (and using jargon to be explained momentarily), we should expect natural selection to shape affect so that it *covaries* with the average fitness effects of the type of event causing it. Now I unpack these ideas, starting with an explanation of fitness effects.

An individual's *fitness*—as it is understood here—is an index of the expected effectiveness with which that individual will propagate the genes that it possesses, whether these genes are within its own or others' bodies.[22] *Fitness effects* can be understood as changes in an individual's fitness that are caused when an individual undergoes events. The most effective way to propagate one's genes is through procreation, so any event directly impacting an individual's chances of procreating, such as finding a mate, affects the fitness of that individual. But events more distantly related to reproduction also have fitness effects: Since an animal must find sustenance if it is to have a chance of mating in the future, the event of finding sustenance also has fitness effects, if smaller than in the case of finding a mate. In addition, an individual's fitness can be affected by events in which the fitness of a *relative* is affected, since a relative would share a significant number of genes with that individual. And so finding sustenance for one's progeny would also be an event that raises one's fitness.

Just as token events have fitness effects, *types* of event have *average* fitness effects, this being determined by the fitness effects that events of that type tend to have. The average fitness effects of event-type *procreation* are positive and

[22] The term "fitness" is not univocal, and so I should explain in more detail how it is to be understood in this article: First, fitness is a *propensity*: It is an index of the *prospects* an individual has for propagating its genes, rather than how successful it actually turns out to be (Mills and Beatty 1979). Second, fitness is *inclusive*, so that an individual's fitness is a sum, not just of that individual's propensity to propagate whatever genes are in its own body (its "personal fitness"), but also of its propensity to help propagate duplicates of its genes existing in others' bodies (e.g., a son's or a sister's) (Hamilton 1964a, b; cf. Grafen 2006; but see Nowak et al. 2010; Rousset and Lion 2011 for a reply). Finally, fitness should be considered *dynamic*. The concept of fitness is often used as a *static* measure of whatever propensity an individual has to propagate its genes at the beginning of its life. However, for present purposes, an individual's fitness should be understood as something that changes when the individual undergoes events affecting its chances of propagating its genes in the future.

The notion of fitness used here focuses on selection occurring at the level of individuals, what is sometimes called "Darwinian fitness" (Darwin 1859). But the term "fitness" can be applied to individual genes, in which case "fitness" refers to a gene's propensity to spread copies of itself. These gene-centered views of fitness have been shown to be mathematically equivalent to Hamilton-style inclusive fitness (see, e.g., Grafen 2006). Not everyone agrees that the only relevant types of fitness are individual- or gene-based. Some evolutionary biologists and philosophers hold that selection also happens at the level of the group (Sober and Wilson 1999), or can involve an organism's standing attributes (this is "developmental systems" theory; see Gray 1992). The theory of affect offered in this article could be enriched to accommodate these possibilities.

relatively high, since each instance of procreation has a positive and sizable impact on one's fitness. And the average fitness effects of the event-type *finding sustenance* are positive but comparatively small, since each instance of finding sustenance has smaller positive effects on one's fitness. Importantly, an event-type can have high average fitness effects even if individual instances of it tend to have negligible fitness effects. Many instances of copulation have insignificant fitness effects, since many times copulation is not procreative. Still, *some* instances are procreative, and whenever one is, it monumentally increases one's fitness. Thus the event-type *copulation* is of high average fitness effects, even though most instances of it fail to boost one's fitness.

Note now that, just as motivation did, fitness effects have properties analogous to those possessed by affect: Fitness effects have *polarity*, since an instance of fitness effects can be positive or negative, according to whether they increase or decrease one's fitness. And fitness effects have something akin to *intensity*, since an instance of fitness effects can be of greater or smaller magnitude; I will refer to this as the *amount* of fitness effects.

Machinery in place, I can present the following thesis:

> *(Affect-Selection Symmetry)*
> For any event-type T and affective experience E, there will be selective pressure for the polarity and intensity of E to come to covary with the polarity and amount of the average fitness effects of T

(I hasten to stress that Affect-Selection Symmetry does not hypothesize that there *actually is* tight covariance between affect and fitness effects, rather just that there is selective pressure for such covariation; this is all that my overall argument requires.)

A first reason to think that this thesis is true comes from the fact that Affect-Motivation Symmetry seems to *entail* Affect-Selection Symmetry. That it does can be seen by considering how there being covariance between affect and motivation leads to there being selective pressure for affect and average fitness effects to covary in their polarity. Say some event-type T has negative average fitness effects. Now consider two distinct phenotypes,[23] P-1 and P-2, where P-1 has *negative* affect in response to T, and P-2 has *positive* affect in response to T. Because the polarity of an affective experience determines the polarity of the ensuing motivation, P-1 would have a selective advantage over P-2: Whereas P-1 would be motivated to avoid an event-type having negative average fitness effects, P-2 would be motivated seek it out. It is clear, then, that creatures whose affect covaries in its polarity with the average fitness effects of event-types will be more fit than those whose affect does not.

Similar observations can be made in the case of the *intensity* of affect. Say some event-type T has negative average fitness effects that are modest. Now consider two possible phenotypes, P-1 and P-2, the first of which has negative affect of *low* intensity in response to T, the second of which has negative affect of *high* intensity in response to T. In such a scenario, P-1 would likely have an evolutionary advantage over P-2, since P-2 would be motivated to avoid T to a maladaptive

[23] A phenotype is an expression of a genotype, the behavior or trait that a gene gives rise to.

degree. For example, if event-type T was the receiving of a scratch to the arm, P-2 might abstain from crawling through a thicket even in cases where there was a badly needed meal on the other side. On the other hand, P-1 would overcome any such reluctance because the expected amount of negative affect would not overwhelm the creature's motivation to obtain the needed sustenance. Because of situations like this, P-1 would have a selective advantage relative to P-2. It looks, then, as if creatures whose affect covaries in its intensity with the amount of average fitness effects of event-types will be more fit than those whose affect does not.

A second reason to think that Affect-Selection Symmetry is true comes in the form of notable instances in which affective experience and the expected average fitness effects of event-types appear to covary. Event-types that we would guess are of large positive average fitness effects, like orgasm, are accompanied by particularly intense positive affect. Likewise, event-types that we would guess are of very small positive average fitness effects, such as eating a morsel, produce positive affect of much weaker intensity. On the other hand, event-types that we would guess are of high negative average fitness effects, such as being struck in the testicles, bring about high intensity negative affect, while event-types that we would guess are of more modest negative average fitness effects, such as being scratched by a thorn, produce low intensity negative affect. That affect enjoys these kinds of covariance is probably best explained by there actually being selective pressure for covariance between affect and the average fitness effects of event-types.

One might object, however, that there also are cases speaking *against* Affect-Selection Symmetry. Receiving a vaccination and giving birth to a child both have or tend toward positive fitness effects, and yet are associated with mostly (if not only) negative affect. So is a migraine headache, though it does not seem associated with any event-type bearing fitness consequences. However, vaccinations are a recent invention, and natural selection works slowly, meaning it is too soon for natural selection to have fostered covariance between vaccinations and positive affect (if it ever does; see next paragraph). In the case of childbirth, we should expect any potential there was for covariation between positive affect and childbirth to be overwhelmed by the strongly negative affect brought on by the physical trauma concomitant to childbirth, a result of the birth canal being just large enough to fit an infant's head. Finally, a migraine probably involves *misrepresentation*, this being a case where sensory systems mistake a physiologically abnormal occurrence for an event-type that typically has large fitness effects (e.g., head trauma), in turn causing affective systems to represent that event's average fitness effects.

In any event, there being cases where covariation does not obtain between affect and fitness effects probably cannot speak strongly against that idea that there is selective pressure for such covariation, as there are general evolutionary reasons for thinking that the hypothesized selective pressure would not guarantee tight covariation between the two. As a rule, designs that natural selection settles upon are not optimal, just marginally better than whatever other flawed designs it has to choose from (Jacob 1977). Part of the problem here is that natural selection is unable to select for "stepping stone" designs, ones that are on their way to a more adaptive design, but which offer no selective advantage themselves, or are of selective disadvantage. As a result, most potentially beneficial adaptations turn out

inaccessible to natural selection, just because they are more than one step away in design space (Gould 1980). Another problem comes from the fact that a species' environment frequently changes. This often causes the average fitness effects of event-types to change, making fitness effects all the harder for natural selection to track. Expect this problem to be particularly pronounced in humans, who live in surroundings far removed from the Pleistocene environment to which they are adapted (Cosmides and Tooby 1997).[24] All in all, it would be quite surprising if affective experience faithfully covaried with the average fitness effects of each event-type, whether or not Affect-Selection Symmetry is true.

It looks, then, as if the fact that there are instances in which affect and average fitness effects fail to covary does not threaten Affect-Selection Symmetry. On the other hand, there exist many key cases in which the two do covary (e.g., an orgasm produces strongly positive affect, eating a morsel produces mildly positive affect), lending plausibility to the idea that there is indeed selective pressure for covariance between affect and fitness effects. But even if these notable cases of covariation were unconvincing, the reasoning presented several paragraphs above seems enough to establish Affect-Selection Symmetry: The very fact that, given affect's association with motivation, an advantage accrues to any individual whose affect is better attuned to the average fitness effects of event-types, seems in and of itself to guarantee that such selective pressure exists.

In conclusion, there are good reasons for thinking that Affect-Selection Symmetry is true. As I intend to show next, this makes fitness effects a good candidate for what affect represents.

3.3 Solution to the Covariation Problem

Remember what is needed to solve the Covariation Problem: some objective, naturalistic property that covaries with affect under optimal conditions. Let us look at how well fitness effects (i.e., the average fitness effects of event-types) satisfy these criteria.

First, fitness effects are objective. It is tempting to think that they are merely subjective, since there are no fitness effects without some individual affected by them. But remember that what we are concerned with here is avoiding the vicious circularity that arises if the candidate property is explicated in terms of the kind of experience to which it gives rise. This need not be done for fitness effects, and so fitness effects are objective in the relevant sense.

Second, fitness effects are naturalistic. The criterion adopted above says that a property is naturalistic iff it plays a part in our final science. Given that fitness effects are frequently and widely appealed to in evolutionary biology (e.g., in explanations as to how one trait comes to be promoted relative to another), it is quite likely that fitness effects will have a place in our final science.[25]

[24] This probably explains things such as why modern humans are prone to overeating, are susceptible to drug addiction, and sometimes favor video games to social interaction: While these activities plausibly tend to decrease fitness, opportunities to engage in them have not been around long enough for natural selection to make them less pleasurable.

[25] Or at least some other scientific property to which fitness effects are eventually reduced.

Finally, fitness effects covary with affect in the requisite way. According to the kinds of tracking theory to which reductive representationalists typically subscribe, some state S represents property P iff S covaries with P *under optimal conditions*.[26] Optimal conditions are typically understood as conditions under which a cognitive system is able to do what it is designed to do, conditions in which there are "no distorting factors, no anomalies or abnormalities" to prevent it from fulfilling its function (Tye 1995: 101). That affect covaries with fitness effects under optimal conditions can be derived from Affect-Selection Symmetry, which says that there is selective pressure for affect to covary with fitness effects. There being this selective pressure means that even when affect does not in fact covary with the average fitness effects of some event-type, the system governing affect is designed such that, if it *were* the case that conditions were optimal—i.e., if the environment were relatively stable, if the affective system did not malfunction—then affect *would* covary with that event-type's average fitness effects.

We have a solution to the Covariation Problem: Fitness effects are an objective, naturalistic property that covaries with affect under optimal conditions.[27,28] This seems like a significant step toward a having a reductive representationalist account of affect.

4 Objections and Competitors

Earlier in the article (Sect. 2), we discussed Cutter and Tye's account of affect, according to which affect represents aptness to harm or benefit. This proposal has recently been criticized by Aydede and Fulkerson (2014). Their criticisms probably also apply to the theory that affect represents fitness effects. The first thing that I do in this section is answer what I take to be Aydede and Fulkerson's central

[26] As indicated earlier (Sect. 2), the theory of representation just appealed to is a *generic* covariational theory, one broad enough to capture any of the theories that reductive representationalists typically employ (see fn. 6). Given this, the theory of affect offered in this article should turn out to be the correct way to analyze affect, regardless of which of these theories of representation prevail; indeed, this is probably the case even if some kind of consumer semantics (e.g., Millikan 1989) prevails, since it seems that the consumers of affective representations (i.e., the motivational systems) could properly perform their evolved function only when affect covaried with fitness effects.

[27] One might worry that the reasoning employed here leads to the idea that *all* adaptations represent fitness effects, since adaptations covary with fitness effects; namely, the positive fitness effects that justify their having been selected. There is, however, a conspicuous difference: In the case of whatever system produces affective states, there is another system "consuming" its states (the motivational system), and this other system depends on these states covarying with the fitness effects of event-types if it is to be able to properly perform its function (that of recalibrating a creature's behavior so that it is more fit). This is something far different from the way in which, e.g., giraffes' long necks covary with positive fitness effects.

[28] At this point, one might wonder whether all I have done is establish is that affect and fitness effects covary (under optimal conditions), without showing that the one actually *represents* the other. However, I *have* shown that fitness effects meet the criteria for representation as laid out by a generic teleologically based theory of content (of the type to which reductive representationalists typically appeal); and so to the extent that such a theory is correct, I *have* shown that affect represents fitness effects. True, whether such theories are correct continues to be controversial, but of course resolving that issue goes well beyond available space.

objection.[29] Then I explain why the account of affect given in this article should be preferred to Cutter and Tye's.

4.1 The Conceptual Priority Argument

In the last section, I offered a solution to the Covariation Problem, something claimed to be a *special problem* faced by attempts to reduce affective experience, i.e., something beyond the usual barriers thought to stand in the way of reducing experiential properties to representational properties. This, however, is not the only problem alleged to be unique to affect. Aydede and Fulkerson (2014) have argued, using Cutter and Tye's theory as an example, that attempts to reduce affect to the representation of *any* objective property ends up in conflict with a core commitment of reductive representationalism, making the reduction of affect in principle impossible.

The core commitment in question is what we might call the Transparency Thesis:

> Any quality that we (can) epistemically encounter in the introspection of an experience is a quality only (widely) represented by this experience, thus not a quality of the experience (p. 180)

As one would guess, the Transparency Thesis follows from representationalism's identification of experiential properties with representational properties. It is what allows the representationalist to explain away purported intrinsic mental properties (sometimes known as "qualia") that one might appear to find in introspection, casting them instead as worldly properties represented by the experience. So the Transparency Thesis is central to representationalism indeed.

Now, it is sometimes thought that if the Transparency Thesis is true, then a sort of *conceptual priority* should obtain: One should be unable to have a concept of experience E (what we could call an "experiential concept," or "e-concept") without also having a concept of whatever property P it is that E represents (what A&F call a "perceptual concept," or "p-concept").[30] The idea here is that, since the experiential property E is just the property of representing P, if a subject lacks a concept of P, she lacks the conceptual wherewithal to latch onto E, and ends up introspectively "blind" to the experience (Dretske 1995, Chap. 5; Aydede 2003; Aydede and Fulkerson 2014).[31]

This suggests to Aydede and Fulkerson (hereafter A&F) a way to test whether some type of experience is consistent with the Transparency Thesis: We simply ask whether it is possible to have an e-concept of that type of experience without having

[29] Thus the objection looked at below is one of a few interweaving (and complex) lines of argument appearing in Aydede and Fulkerson's article.

[30] P-concepts are the concepts used when information in an experience is used to non-inferentially form beliefs about environmental properties. So, if someone sees a red object and, on the basis of her experiencing its redness, forms the belief that the object she is looking at is red, the concept used to form this belief would be a p-concept. (For discussion, see Aydede and Fulkerson 2014: 182–183.)

[31] Here we are adopting the "displaced perception" view of introspection, something endorsed (in one form or another) by many reductive representationalists, but most closely identified with Dretske. Whether reductive representationalism has a satisfactory account of introspection is no settled matter (see Aydede 2003). Just as A&F do, I am taking it as an assumption that some kind of displaced perception view can account for introspection.

the corresponding p-concept. If it is possible, this constitutes a counterexample to the Transparency Thesis, and therefore to representationalism as well. Call this the *transparency test*.

A&F believe that most types of experience pass the transparency test. Letting the word "reddish" stand for the experiential property that constitutes an experience as of redness, A&F formulate the transparency test for reddish experiences as follows:

> …one has the concept REDDISH only if one has the concept RED (p. 193)

This looks plausible: It is hard to picture someone being able to conceptualize their experiences as of redness without having a concept of red as a property of objects, at least in some sense. Now, this raises questions concerning under what conditions one should be considered to possess the p-concept RED, but let us postpone them. Instead, we simply join A&F in the sensible assumption that—whatever these conditions turn out to be—experiences as of redness (along with many other types of experience) pass the transparency test.

A&F think that the transparency test poses a special problem in the case of affect. Cutter and Tye argue that affect represents aptness to harm, and A&F take this to mean that the transparency test for affect should look as follows:

> One has the concept PAINFUL only if one has the concept…HARMFUL (p. 194)

This thesis is false: One can say that one is in pain without knowing anything about pain's connection with harmfulness, perhaps without having a concept of harmfulness at all. So, if affect is supposed to represent harmfulness, affect looks like a counterexample to the Transparency Thesis, and therefore also to representationalism. Such, at least, is A&F's Conceptual Priority Argument.

It is easy to see how the Conceptual Priority Argument might be used against the idea that affect represents fitness effects, since it is possible to have a concept of one without the other; that is to say, this thesis is false:

> One has the concept PAINFUL only if one has the concept FITNESS EFFECTS

And so it appears that the Conceptual Priority Argument might be a threat to the theory that affect represents fitness effects. Indeed, A&F believe that their argument generalizes to any objective property that affect is hypothesized to represent.

I am not sure, however, whether the transparency test is consistently or sensibly applied in the Conceptual Priority Argument. The problem is that the thesis just above (let us refer to it as the *Fitness Effects Formulation*) assumes an impossibly strong standard for what it takes to possess a p-concept of what some type of experience represents. Consider the following: The idea that affect represents fitness effects—as presented in this article—is a hypothesis concerning the "deep essence" or "underlying nature" (Kripke 1972) of whatever it is that affect represents; it is meant to be akin to the hypothesis that water is H_2O, or that heat is mean molecular motion. It seems, then, that if we want to formulate the transparency test for affect as it is in the Fitness Effects Formulation, we must adopt a very demanding standard for p-concept possession, the following:

One has a p-concept of P only if one has a concept of the underlying nature of P

This standard, however, is too stringent. This quickly becomes evident if we try to apply it in the case of color experience. According to contemporary color science, something's being a certain color consists of a disposition to preferentially reflect light of certain wavelengths, what is known as a *spectral reflectance profile*. So, assuming that color experience is representational, the underlying nature of what color experience represents is probably spectral reflectance profiles (Tye 2000, Chap. 7). Given this, if we hold color experience to the same standard that affect is held to in the Fitness Effects Formulation, the transparency test for experiences as of redness should be formulated, not as A&F formulate it in their article, but rather like this:

One has the concept REDDISH only if one has the concept SPECTRAL REFLECTANCE PROFILE

Of course, reddish experiences fail *this* transparency test. As do many other types of experience: One can have a concept of an experience as of something being hot without having a concept of high mean molecular motion; likewise, one can have a concept of an experience as of something being high-pitched without having a concept of high-frequency sound waves. Since these are all types of experience that we would have thought to pass the transparency test (at least, all those party to the debate assume so in the case of color experience), it seems that the standard under consideration is too strong: Having the p-concept of what an experience represents cannot require having a concept specifically of whatever is the underlying nature of what that experience represents.[32] If so, the Fitness Effects Formulation is not an apt way to express the transparency test for affect. Furthermore, if Cutter and Tye's theory that affect represents aptness to harm similarly concerns the underlying nature of what affect represents, the "Harmfulness Formulation" is not apt either.[33]

The Conceptual Priority Argument, as it stands, appears ineffective against the theory that affect represents fitness effects. It at least cannot deliver the problem that A&F advertised as being *unique* to affect. Maybe the difficulty here is that we as yet

[32] Indeed, Tye has indicated that we should not expect the fact that one experiences some property P to guarantee any deep insight into what precisely P is. A&F point out this passage by Tye: "On my account, what it is exactly that a given experience or feeling represents need not be accessible to the subject's cognitive centers, including his or her powers of introspection, except in the most general and uninformative way (for example, as an experience of *this* sort)" (1996: 52). A&F are skeptical as to whether a bare-bones p-concept like this is "consistent with the demands" of the Transparency Thesis, but do not elaborate (2014: 195).

[33] There is an interesting question that has emerged in this discussion, which is why it is the case that introspection on a conscious state often returns something not easily recognizable as whatever property that state represents. For instance, when introspecting color experience, one does not find something resembling the dispositional property of a spectral reflectance profile, but rather what seems to be a stable categorical property of an object's surface; similarly, when introspecting an experience as of a high-pitched sound, one does not find something resembling discrete sound waves arriving at the ear at a rate of thousands per second, but rather a continuous and uniform pitch; and, finally, when introspecting affective experience, one does not find something resembling the average fitness affects of an event, but rather something that—putting it roughly—feels good or bad. Why there is this disconnect between the underlying nature of what a conscious state represents, and what we find in introspection, is an interesting question indeed, one that must eventually be answered as part of any comprehensive reductive representationalist theory of consciousness.

lack a good account of what a p-concept is, and that this makes it hard to faithfully execute the transparency test. Perhaps an account of a p-concept could be produced in such a way that affective experience, but not other kinds of experience, turns out to fail the transparency test. But this remains to be seen.

4.2 Competing Accounts of Affective Experience

Now we move on to the question of how the theory of affect offered in this article compares to Cutter and Tye's theory that affect represents aptness to harm/benefit. Let us call the first theory, "Fitness Effects," the second, "Harm." I will argue that Fitness Effects should be preferred to Harm since it explains a wider range of phenomena, and overall offers a more natural candidate for what affect represents.

Both Fitness Effects and Harm do well in cases involving physiological harm or benefit: Each theory predicts wounds to cause negative affect, since wounds are both physiologically harmful and of negative fitness effects; and each theory predicts resting when exhausted to cause positive affect, since doing so is both physiologically beneficial and of positive fitness effects. So far, so good. But now let us reexamine the scenarios discussed in Sect. 2, ones appearing problematic for Harm: Social exclusion causes negative affect though it produces no physiological harm; witnessing one's offspring injured causes negative affect, though it appears of no threat to the subject, physiological or otherwise; and an orgasm causes strongly positive affect, though it is of no comparable physiological benefit. Fitness Effects, however, can explain these scenarios by appealing to the average fitness effects of each event-type: Social exclusion produces negative affect because alienation from the group negatively affects one's ability to survive and reproduce in any number of ways (since it means, e.g., that one's alliances are weakening) (Kerr and Levine 2008); injury to offspring produces negative affect because it lowers the fitness of an individual sharing a significant number of genes with the subject; and orgasms produce strongly positive affect because they sometimes result in procreation, an event with highly positive average fitness effects. It seems that Fitness Effects gets all of the scenarios right that Harm does, and then some.

One might respond, however, that there is physiological harm/benefit involved in the cases just examined, since such events might cause or reduce stress, and stress creates health problems. Another response would be to claim that aptness to harm/benefit should be understood as including *psychological* harm/benefit, where this in turn is understood as something degrading or promoting a psychological system's ability to function properly.[34] Seeing one's offspring injured, for instance, evokes potent emotions, deleteriously affecting one's ability to reason or inhibit impulsive action. Similar stories could be told with insults and orgasms.

Such responses, however, only deliver covariation between harm and affect in their *polarity*, not *amount*. Since stress causes physiological harm only if experienced over long periods, each stressor is only weakly apt to harm/benefit, but the negative affect caused by an insult or seeing one's offspring injured can be quite intense. It is also doubtful that whatever modest physiological/psychological

[34] This would be in the spirit of C&T's teleological construal of harmfulness; see fn. 8.

benefits an orgasm causes measures up to the extraordinary positive affect that it involves. The same seems true in the case of the strongly negative affective experience involved in receiving an insult or seeing one's offspring injured.

So it is questionable whether these responses establish covariance between affect and aptness to harm/benefit. Of course, this lack of covariance alone does not disqualify harmfulness from being what affect represents. Something stressed in Sect. 2 was that a property might only weakly covary with affective experience and yet still be what it represents, since all that is needed is covariation *under optimal conditions*. Keeping this in mind, we can still ask: Given the kind of covariance fitness effects and harmfulness each display with affect, which of the two is more naturally construed as being that which affective experience covaries with under optimal conditions, and therefore represents? I submit that it is fitness effects.

There is, in addition, another problem for Harm: It seems to exclude, from being a cause of an affective experience, whatever event-type it is most plausibly a response to. In the case of an orgasm, for instance, the Harm theory says that the affective experience is a response to, not the event-type *potential procreation*, but rather those event-types constituting whatever physiological/psychological benefits the orgasm produces. But there are certainly much stronger selective forces at work for building covariation with *potential procreation* than there are with those event-types constituting the potential physiological/psychological benefits. Similarly, in the insult case, Harm must deny that the affective experience is caused by the event-type constituting one's descent in the social hierarchy, despite the evolutionary importance of this. Finally, in the injury to offspring case, Harm has to say that the affective experience is not caused by there having been a sharp drop in the fitness of someone sharing half of the subject's genes. In each scenario, Harm gets the explanation wrong, in that it excludes, from being what an affective experience is responding to, the event-type that natural selection would have worked the hardest to have affect covary with.

Overall, it seems that there are good reasons to prefer Fitness Effects to Harm, not just because fitness effects covary with affect in a way more strongly suggesting that they are what affect represents, but also because Harm seems to have to exclude the appropriate event-types from being that which causes an affective experience. I conclude that fitness effects are a more promising candidate for what affect represents.[35]

[35] One might argue that the theory offered in this article is not a competitor to Cutter and Tye's, but rather just one way in which their abstractly specified theory could be filled out. This seems incorrect, however, since Cutter and Tye take the affective component of pain experience to (a) represent something about a "bodily disturbance," where (b) this bodily disturbance is something occurring in the subject herself. The Fitness Effects theory, however, hypothesizes neither of these things, which is why it explains scenarios that Harm cannot. Of course, Harm could be modified so that affect is a response to something beyond the physiological/psychological effects that an event has, in which case the theory might avoid the untoward results described above. However, it seems that the logical next step would be to widen the theory so that affect is a response to the evolutionary consequences of the situation causing the affect, in which case the theory collapses into the theory of affect offered in this article.

5 Conclusion

Reductive representationalists have had difficulty finding some property that affective experience covaries with, casting doubt on the idea that it is representational. In this article, I argued that, because of affect's connection with motivation, there is selective pressure for it to covary with the average fitness effects of event-types, and that this selective pressure makes fitness effects a promising candidate for what affect represents. As seen above, the theory offered does not succumb to an objection Aydede and Fulkerson recently raised against representational accounts of affect, and does not suffer from the kinds of difficulty facing Cutter and Tye's proposal that affect represents harmfulness. Overall, there is ample reason to think that, in fitness effects, we finally have a plausible candidate for what affective experience represents.

References

Armstrong, D. (1962). *Bodily sensations*. London: Routledge.
Aydede, M. (2003). Is introspection inferential? In B. Gertler (Ed.), *Privileged access: Philosophical accounts of self-knowledge* (pp. 55–64). Burlington: Ashgate Publishing.
Aydede, M. (2005). The main difficulty with pain. In M. Aydede (Ed.), *Pain: New essays on its nature and the methodology of its study* (pp. 123–136). Cambridge: MIT Press.
Aydede, M. (2009). Is feeling pain the perception of something? *Journal of Philosophy, 106*(10), 531.
Aydede, M., & Fulkerson, M. (1962). Perceptual affect: A critique and a proposal. In D. Armstrong (Ed.), *Bodily sensations*. London: Routledge.
Aydede, M., & Fulkerson, M. (2014). Affect: Representationalists' headache. *Philosophical Studies, 170,* 175–198.
Bain, D. (2013). What makes pains unpleasant? *Philosophical Studies, 166*(1), 69–89.
Berridge, K. C. (2004). Motivation concepts in behavioral neuroscience. *Physiology & Behavior, 81*(2), 179–209.
Block, N. (1990). Inverted earth. *Philosophical Perspectives, 4,* 53–79.
Block, N. (1996). Mental paint and mental latex. *Philosophical Issues, 7,* 19–49.
Block, N. (2005). Bodily sensations as an obstacle to representationism. In M. Aydede (Ed.), *Pain: New essays on its nature and the methodology of its study* (pp. 123–136). Cambridge: MIT Press.
Charland, L. C. (2005). The heat of emotion: Valence and the demarcation problem. *Journal of Consciousness Studies, 12*(8–9), 82–102.
Clark, A. (2000). *A theory of sentience*. Oxford: Clarendon Press.
Cohen, J., & Fulkerson, M. (2014). Affect, rationalization, and motivation. *Review of Philosophy and Psychology, 5*(1), 103–118.
Colombetti, G. (2005). Appraising valence. *Journal of Consciousness Studies, 12*(8–9), 103–126.
Corns, J. (2014). Unpleasantness, motivational oomph, and painfulness. *Mind and Language, 29*(2), 238–254.
Cosmides, L. & Tooby, J. (1997). *Evolutionary psychology: A primer*. Available at: http://www.cep.ucsb.edu/primer.html.
Cosmides, L., & Tooby, J. (2000). Evolutionary psychology and the emotions. *Handbook of Emotions, 2*(2), 91–115.
Cutter, B., & Tye, M. (2011). Tracking representationalism and the painfulness of pain. *Philosophical Issues, 21*(1), 90–109.
Damasio, A. (1994). *Descartes' error*. New York: Avon Books.
Damasio, A. (1999). *The feeling of what happens*. New York: Harvest.
Darwin, C. (1859). *On the origin of species by means of natural selection*. London: John Murray.
Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge: MIT Press.
Dretske, F. (1995). *Naturalizing the mind*. Cambridge: MIT Press.

Eisenberger, N. I., & Lieberman, M. D. (2004). Why rejection hurts: A common neural alarm system for physical and social pain. *Trends in Cognitive Sciences, 8*(7), 294–300.
Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science, 302*(5643), 290–292.
Fodor, J. A. (1990). A theory of content, II: The theory. In *A theory of content and other essays*. Cambridge, MA: MIT Press.
Gould, S. (1980). The evolutionary biology of constraint. *Daedalus, 109,* 39–52.
Grafen, A. (2006). Optimization of inclusive fitness. *Journal of Theoretical Biology, 238*(3), 541–563.
Gray, R. (1992). Death of the gene: Developmental systems strike back. In L. Griffiths (Ed.), *Trees of life: Essays in the philosophy of biology* (pp. 165–209). Dordrecht: Kluwer.
Haake, P., Krueger, T. H., Goebel, M. U., Heberling, K. M., Hartmann, U., & Schedlowski, M. (2004). Effects of sexual arousal on lymphocyte subset circulation and cytokine production in man. *NeuroImmunoModulation, 11*(5), 293–298.
Hall, R. (2008). If it itches, scratch! *Australasian Journal of Philosophy, 86*(4), 525–535.
Hamilton, W. (1964a). The genetical evolution of social behaviour. I. *Journal of Theoretical Biology, 7*(1), 1–16.
Hamilton, W. (1964b). The genetical evolution of social behaviour. II. *Journal of Theoretical Biology, 7*(1), 17–52.
Harman, G. (1990). The intrinsic quality of experience. *Philosophical Perspectives, 4,* 31–52.
Helm, B. W. (2002). Felt evaluations: A theory of pleasure and pain. *American Philosophical Quarterly, 39*(1), 13–30.
Hofbauer, R., et al. (2001). Cortical representation of the sensory dimension of pain. *Journal of Neurophysiology, 86*(1), 402–411.
Jacob, F. (1977). Evolution and tinkering. *Science, 196,* 1161–1166.
Kerr, N. L., & Levine, J. M. (2008). The detection of social exclusion: Evolution and beyond. *Group Dynamics: Theory, Research, and Practice, 12*(1), 39.
Klein, C. (2007). An imperative theory of pain. *The Journal of Philosophy, 104,* 517–532.
Klein, C. (2015). *What the body commands: The imperative theory of pain*. Cambridge: MIT Press.
Kripke, S. (1972). *Naming and necessity*. Dordrecht: Springer.
Kupers, R., et al. (1991). Morphine differentially affects the sensory and affective pain ratings in neurogenic and idiopathic forms of pain. *Pain, 47,* 5–12.
Leknes, S., & Tracey, I. (2008). A common neurobiology for pain and pleasure. *Nature Reviews Neuroscience, 9*(4), 314–320.
Leventhal, H. (1993). The pain system: A multilevel model for the study of motivation and emotion. *Motivation and Emotion, 17*(3), 139–146.
Lycan, W. G. (2001). The case for phenomenal externalism. *Noûs, 35*(s15), 17–35.
Martínez, M. (2011). Imperative content and the painfulness of pain. *Phenomenology and the Cognitive Sciences, 10*(1), 67–90.
McGinn, C. (1982). *The character of mind*. Oxford: Oxford University Press.
Melzack, R., & Casey, K. (1968). Sensory, motivational, and central control. Determinants of pain. In *The skin senses* (pp. 423–439). Springfield: Thomas.
Millikan, R. G. (1989). Biosemantics. *The Journal of Philosophy, 86*(6), 281–297.
Millikan, R. G. (1995). Pushmi-pullyu representations. *Philosophical Perspectives, 9,* 185–200.
Mills, S., & Beatty, J. (1979). The propensity interpretation of fitness. *Philosophy of Science, 46,* 263–286.
Nesse, R. M. (1990). Evolutionary explanations of emotions. *Human Nature, 1*(3), 261–289.
Nesse, R. M., & Ellsworth, P. C. (2009). Evolution, emotions, and emotional disorders. *American Psychologist, 64*(2), 129.
Nowak, M., Tarnita, C., & Wilson, E. (2010). The evolution of eusociality. *Nature, 466*(7310), 1057–1062.
O'Sullivan, B., & Robert, S. (2012). Painful reasons: Representationalism as a theory of pain. *The Philosophical Quarterly, 62*(249), 737–758.
Peciña, S., Cagniard, B., Berridge, K. C., Aldridge, J. W., & Zhuang, X. (2003). Hyperdopaminergic mutant mice have higher "wanting" but not "liking" for sweet rewards. *The Journal of neuroscience, 23*(28), 9395–9402.
Pitcher, G. (1970). The awfulness of pain. *Journal of Philosophy, 48,* 481–492.
Prinz, J. (2010). For valence. *Emotion Review, 2*(1), 5–13.
Quine, W. (1981). *Theories and things*. Cambridge: Harvard University Press.

Rainville, P., et al. (1997). Pain affect encoded in human anterior cingulate but not somatosensory cortex. *Science, 277*(5328), 968–971.
Rainville, P., et al. (1999). Dissociation of sensory and affective dimensions of pain using hypnotic modulation. *Pain, 82*(2), 159–171.
Rousset, F., & Lion, S. (2011). Much ado about nothing: Nowak et al charge against inclusive fitness theory. *Journal of Evolutionary Biology, 24*(6), 1386–1392.
Russell, J. A., & Carroll, J. M. (1999). On the bipolarity of positive and negative affect. *Psychological Bulletin, 125*(1), 3.
Siewart, C. (1998). *The significance of consciousness*. Princeton: Princeton University Press.
Sober, E., & Wilson, D. S. (Eds.). (1999). *Unto others: The evolution and psychology of unselfish behavior (no. 218)*. Cambridge: Harvard University Press.
Stalnaker, R. (1984). *Inquiry*. Cambridge: MIT Press.
Stampe, D. W. (1977). Toward a causal theory of linguistic representation. *Midwest Studies in Philosophy, 2*(1), 42–63.
Thorndike, E. (1913). *Educational psychology*. New York: Teachers College Press.
Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology, 46*(1), 35–57.
Tye, M. (1995). *10 Problems of consciousness: A representational theory of the phenomenal mind*. Cambridge: MIT press.
Tye, M. (1996). Orgasms again. *Philosophical Issues, 7,* 51–54.
Tye, M. (1997). A representational theory of pains and their phenomenal character. *Philosophical Perspectives, 9,* 223–239.
Tye, M. (2000). *Consciousness, color, and content*. Cambridge: MIT Press.
Tye, M. (2002). Representationalism and the transparency of experience. *Noûs, 36*(1), 137–151.
Tye, M. (2005). Another look at representationalism about pain. In M. Aydede (Ed.), *Pain: New essays on its nature and the methodology of its study* (pp. 123–136). Cambridge: MIT Press.
Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin, 98*(2), 219.
Whipple, B., & Komisaruk, B. R. (1988). Analgesia produced in women by genital self-stimulation. *Journal of Sex Research, 24*(1), 130–140.