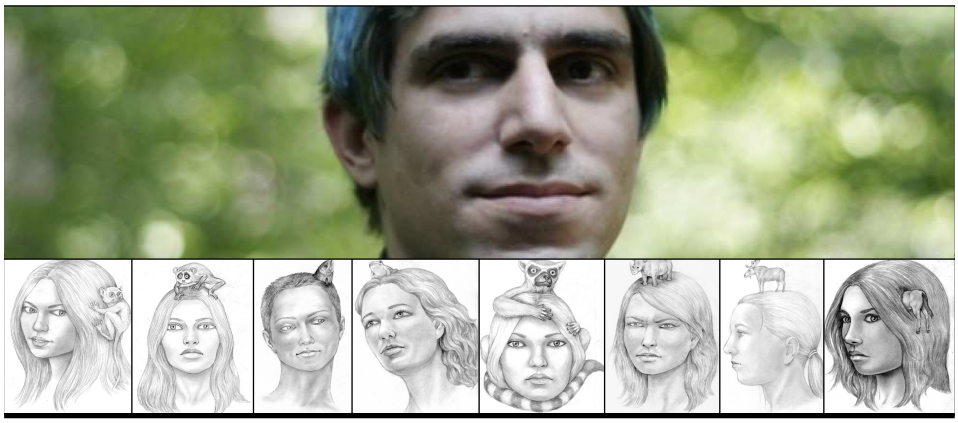
XXII SIUCC - SEFA Workshop on Jesse Prinz

Donostia – San Sebastián, September 6-8, 2012



Thursday 6

[09:15 – 09:30] Presentation
[09:30 – 11:00] **Jesse Prinz** (CUNY Graduate Center):
"The AIR Theory of Consciousness"
[11:00 – 11:30] Coffee break
[11:30 – 12:30] **Marta Jorba** (Logos; Universitat de Barcelona):
"The limits of sensory phenomenology: Jesse Prinz on conscious thought"
[12:30 – 13:30] **David Pineda** (Logos; Universitat de Girona):
"Are Emotions Valent Embodied Appraisals?"
[13:30 – 15:00] Lunch
[15:00 – 16:00] **Stefano Cossara** (Paris-Sorbonne University):
"Naturalism, Pluralism and Location Problems. Reflections on Jesse Prinz's Naturalistic Commitments"
[16:00 – 16:30] Coffee break
[16:30 – 17:30] **Xabier Barandiaran** (IAS Research; UPV/EHU):
"Pushing the accelerator on enactive perception: how sensorimotor dynamics constitute minds"

Friday 7

[09:30 – 11:00] **Jesse Prinz**:
"Sensing the World: Concepts, Perception, and Reality"
[11:00 – 11:30] Coffee break
[11:30 – 12:30] **Andrea Onofri** (Arché; University of St Andrews):
"Two Constraints on a Theory of Concepts"
[12:30 – 13:30] **Mark Cain** (Oxford Brookes University):
"The Proxytype Theory of Concepts"
[13:30 – 15:00] Lunch
[15:00 – 16:00] **Marc Artiga** (Logos; Universitat de Girona):
"Prinz's Theory of Conceptual Content"
[16:00 – 16:30] Coffee break
[16:30 – 17:30] **José Manuel Palma** (Universidad de Granada):
"The Missing Intentionality in Prinz's Theory of Emotion: (Historical) Reflections from Solomon"
[20:30] Workshop dinner

Saturday 8

[10:00 – 11:00] **Jake Davis** (CUNY Graduate Center):
"What Feels Right, Objectively: A Sentimentalist Rebuttal to Prinz's Sentimentalist Moral Relativism"
[11:00 – 12:00] **Alejandro Rosas** (Universidad Nacional de Colombia):
"Morality and the pro-social emotions: a nativist view"
[12:00 – 12:30] Coffee break
[12:30 – 14:00] **Jesse Prinz:**
"The Moral Self"

Organizing institutions
SEFA Sociedad Española de Filosofía Analítica
IAS-Research Centre for Life, Mind and Society
More information & registration:
http://siucc2012.ias-research.net/

Scientific Committee
Agustín Arrieta, Ángeles Eraña, Antoni Gomila, David Pineda, David Teira, Fernando Martínez Manrique, Jordi Fernández, Juan José Acero, Manuel García-Carpintero, Pepa Toribio.
Organizing Committee
Agustin Vicente, Antonio Casado da Rocha, Arantza Etxeberria, Xabier Barandiaran

Universidad del País Vasco · Euskal Herriko Unibertsitatea · SEFA sociedad española de filosofía analítica · MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD · ikerbasque Basque Foundation for Science

Table of Contents

Timetable

Thursday 6

[09:15 – 09:30] Presentation
[09:30 – 11:00] Jesse Prinz (CUNY Graduate Center): "The AIR Theory of Consciousness"
[11:00 – 11:30] Coffee break
[11:30 – 12:30] Marta Jorba (Logos; Universitat de Barcelona): "The limits of sensory phenomenology: Jesse Prinz on conscious thought"
[12:30 – 13:30] David Pineda (Logos; Universitat de Girona): "Are Emotions Valent Embodied Appraisals?"
[13:30 – 15:00] Lunch
[15:00 – 16:00] Stefano Cossara (Paris-Sorbonne University): "Naturalism, Pluralism and Location Problems. Reflections on Jesse Prinz's Naturalistic Commitments"
[16:00 – 16:30] Coffee break
[16:30 – 17:30] Xabier Barandiarán (IAS Research; UPV/EHU): "Pushing the accelerator on enactive perception: how sensorimotor dynamics constitute minds"

Friday 7

[09:30 – 11:00] Jesse Prinz: "Sensing the World: Concepts, Perception, and Reality"
[11:00 – 11:30] Coffee break
[11:30 – 12:30] Andrea Onofri (Arché; University of St Andrews): "Two Constraints on a Theory of Concepts"
[12:30 – 13:30] Mark Cain (Oxford Brookes University): "The Proxytype Theory of Concepts"
[13:30 – 15:00] Lunch
[15:00 – 16:00] Marc Artiga (Logos; Universitat de Girona): "Prinz's Theory of Conceptual Content"
[16:00 – 16:30] Coffee break
[16:30 – 17:30] José Manuel Palma (Universidad de Granada): "The Missing Intentionality in Prinz's Theory of Emotion: (Historical) Reflections from Solomon"
[20:30] Workshop dinner

Saturday 8

[10:00 – 11:00] Jake Davis (CUNY Graduate Center): "What Feels Right, Objectively: A Sentimentalist Rebuttal to Prinz's Sentimentalist Moral Relativism"
[11:00 – 12:00] Alejandro Rosas (Universidad Nacional de Colombia): "Morality and the pro-social emotions: a nativist view"
[12:00 – 12:30] Coffee break
[12:30 – 14:00] Jesse Prinz: "The Moral Self"

The limits of sensory phenomenology: Jesse Prinz on conscious thought

Marta Jorba (Logos; Universitat de Barcelona)

Extended abstract: 3967 words.

When we consciously think a thought or entertain a certain proposition we undergo a certain experience. We can say that the phenomenal character involved in such an episode is an instance of cognitive phenomenology, at least in the sense that there is *some* phenomenal character in the episode of conscious thought. Some agreement is reached concerning the *existence* of cognitive phenomenology, when this thesis is not filled in with more substantial claims (see Smithies, forthcoming). The mere existence of an experience of consciously thinking is not problematic *per se*, but controversies arise with respect to its *nature.* A question we need to answer is whether cognitive phenomenology is *specifically cognitive* or it is *reducible* to more familiar kinds of phenomenologies, such as the sensory or emotional one[1].

The answer to the nature of cognitive phenomenology implies, among other things, a view on the *reach* of phenomenal consciousness: proponents of a specific cognitive phenomenology defend that phenomenal consciousness includes cognition or thought, while proponents of non-cognitive phenomenologies think that cognition is not under the reach of consciousness (Bayne, 2009). As general views regarding the extension of phenomenally conscious mental episodes, we can thus distinguish between *expansionist* versus *restrictivist* views (Prinz, 2011).[2] The view according to which cognitive phenomenology is reducible to other non-cognitive phenomenologies would be among the restrictivist ones, whereas the defense of a specific phenomenal character would constitute an expansionist one. Notice that the expansionist/restrictivist dichotomy includes other positions regarding

[1] For an overview of the problem, see Bayne, T. & Montague, M. (2011).

[2] Terminology varies a lot here: Bayne (2009) labels both positions 'phenomenal conservatives' versus 'phenomenal liberals', Kriegel (2011) prefers 'phenomenological inflationists' versus 'phenomenological deflationists', and Siewert (2011), talks about 'inclusivism' versus 'exclusivism' (Siewert, 2011).

high-level perceptual properties, emotional episodes, etc., so it is a distinction that serves as an umbrella for many different theories regarding phenomenal consciousness. For the purposes of this paper, I will talk of reductionism and restrictivism as interchangeable labels. The reach of phenomenal consciousness can be cashed out in terms of which kind of mental episodes are thought to be phenomenally conscious and which are not. If one has an answer to these related questions, namely, the reach of phenomenal consciousness and the nature of cognitive phenomenology, one can evaluate whether certain theories of phenomenal consciousness can accommodate these results or not. In other words, if consciousness includes cognition and there are cognitive specific phenomenal properties, then *any* theory of phenomenal consciousness that denies specific cognitive phenomenology is undermined.

I think Jesse Prinz's work on consciousness is illuminating in this respect: he has extensively argued for a general theory of consciousness (Prinz, 2002, 2007, 2012) according to which consciousness arises at the intermediate level of perceptual systems, where feature integration takes place and attention mechanisms are involved[3], that is, with attended intermediate-level representations or AIRs (Prinz, 2005). It is an intermediate level between the low-level stage that responds to local stimulus features without integration and the high level perceptual stage that abstracts away details form the previous one. According to this theory, the neural correlates of perceptual consciousness are thus restricted to brain areas that implement those perceptual processing. The strategy of this account, like many others, is to think that an account of perceptual experiences will give a general account of consciousness, so that the following conclusion serves as an slogan for the view: *all consciousness is perceptual consciousness*[4]. Once we frame the question of consciousness in these terms, the issue of the nature of cognitive phenomenology demands a quite straightforward and direct answer: whatever phenomenal character we are to find in conscious thought, this will have to be perception-like, so we end up with some form of restrictivism or reductionism to the perceptual[5]. This makes us consider whether opposition to a specific cognitive phenomenology or to expansionism in this sense is somehow "theory-biased" in the first place, so that direct denials are provided only when certain theories of

[3] For Prinz, intermediate-level mechanism is necessary but not sufficient for consciousness: attention is needed for consciousness to arise.

[4] Prinz argues for the particular claim that all *phenomenal* consciousness is perceptual phenomenal consciousness, and he believes other uses and forms of consciousness are parasitic on phenomenal consciousness, and thus this more general claim can be defended (see Prinz, 2007, p. 336). This view constrast, for example, with Peacocke's (), according to which conscious thought is a special case of another kind of conciousness, namely, action consciousness. Action awareness is the other case apart from thought than can provide objections to Prinz's view (Prinz, 2007, p. 341).

[5] His intermediate-level view on consciousness has had some objections regarding cases in which non-cognitive phenomenology can be said to outstrip this intermediate-level: high-level perceptual representations, perceptual constancies, the experience of presence in absence, motor actions and emotions. In Prinz (2011), he provides answers to them, but here I am going to focus on the case of thought.

phenomenal consciousness are already accepted. Prinz (2007), however, appeals to parsimony, arguing that "having a single unified theory is, all things being equal, better than having a family of different theories for each kind of phenomenal state that we experience" (Prinz, 2007, p. 337). One assumption of the parsimony argument is that the reduction works.

My aim in this paper is to focus on some paradigmatic cases for specific cognitive phenomenology, then present the main elements of Prinz's reductive account of them and argue that they do not provide a satisfying view. I will not discuss all the subtle ideas and arguments of Prinz's work, but just provide what I take to be three important and new objections to his project. First, a problem related to the phenomenology of inner speech; second, what I call the 'phenomenological adequacy' problem and third, what I take to be 'the limits of sensory variation'. I then sketch an account that does not have these problems, thus making the case for acknowledging the cognitively specific phenomenal character of conscious thought, and thereby undermining Prinz's general theory for consciousness.

Arguments regarding cases for specific cognitive phenomenology[6] normally have the form of phenomenal contrast argument: they present two scenarios where there is a phenomenal change from one to another and nevertheless the non-cognitive components (mainly sensory and perceptual aspects) remain the same. Since, it is argued, the only difference between both scenarios is cognitive, the phenomenal contrast is to be accounted for by appealing to a specific cognitive phenomenology. The paradigmatic case is that of *understanding* vs. *not understanding* some written or heard sentence, for example (Strawson, 1994/2010). Or cases of the experience of comparing the price of some item in a familiar currency versus comparing it in a foreign one.[7] There are also single cases: the example of *imageless thought*, when thoughts are conscious but nevertheless lack images[8] or the cases of *languageless thought,* where there is some consciousness of the thought without any sentence being experienced, like sudden thoughts that occur without time for language (Siewert, 1998).

Given Prinz's theory of perceptual consciousness, the question to be asked regarding cognitive phenomenology is whether consciousness outstrips perception or the senses (Prinz, 2011, p. 174). Prinz's (2011) argumentative strategy regarding specific cognitive phenomenology is mainly negative,

[6] Provided by Husserl, (1900/1901); Siewert (1998); Strawson (1994/2010); Peacocke (1998); Kriegel (2011).

[7] Thanks to Manuel García-Carpintero for this example.

[8] The discussion on imageless thought goes back to introspective psychology and the debate whether imageless thoughts are possible. The experiments where based in people reporting whether they had images or not when asked certain questions: what substances are more costly than gold, etc?. These experiments and introspectionist psychology was after that highly dismissed and regarded as a failure of method: subjects can simply be wrong about their own mental states (see Prinz, 2011, for an overview of the debate).

as he tries to account for the cases in favor of cognitive phenomenology with the resources of his perceptual view of consciousness and then his positive stance consists in giving a diagnosis of the intuitions that guide expansionists in terms of introspective illusions. Before giving the main positions, he makes an important distinction between the vehicle, the content and the quality of mental episodes. The vehicle is a token particular that have representational content: in a sentence, the orthographic marks on the page, or the mental representations in the head. The content is what the vehicle represents: vehicles in the visual system represent shapes and colors, etc.[9] And the quality is how it feels when it is conscious, the phenomenal character. With this distinctions, the main positions in the debate are defined as follows: *restrictivism* is true if, and only if, for every vehicle with qualitative character there could be a qualitatively identical vehicle that has only sensory content; and *expansionism* is true if, and only if, some vehicles with qualitative character are distinguishable from every vehicle that has only sensory content. A content of a vehicle is sensory just in case that vehicle represents some aspect of appearance and a content is non-sensory if it transcends appearance – if there are two indistinguishable things by the senses, one of which has the property and the other not. The point is that the introduction of non-sensory content does not also introduce non-sensory phenomenal qualities, so that the content that goes beyond appearance does not have an impact on quality or experience. Restrictivists like Prinz, then, allow for *conscious* thoughts as long as there are no qualities over and above the sensory ones. He accepts that conscious thought "feels like" something (there is a phenomenology), but not that it feels differently than sensory activity (all phenomenology is reducible to sensory one).

He then tries to accommodate the phenomenal contrast of understanding and similar cases with differences in sensory elements, such as different associated mental images or inner speech or differences in the focus of attention. Briefly, Prinz's (2011, p. 189) conclusion is the following: for cases of imageless thought, verbal imagery is at place and explains the phenomenology and for cases of languageless thought, non-verbal imagery is at place and explains the phenomenology. Cases in which both are absent are difficult to find and think of.

I would like to put pressure on this view of cognitive phenomenology in three ways. The first problem is related to appeals to inner speech without paying enough attention to its phenomenology. I think inner speech or verbal imagery cannot play the role Prinz wants them to play. Notice that for his

[9] Prinz endorses an empiricist view, according to which the vehicles in thought are copies of the ones used in perception and besides shapes, colors, etc., visual vehicles can also represent objects, natural kinds, or more abstract properties like numbers.

account to work, it is the *sensory* component of inner speech the responsible for the changes in phenomenology. There is the distinction between the *sensory* and the *semantic* aspects of inner speech, the first containing the syntactic, phonologic elements, etc., and the second all the other non-sensory aspects like the interpretation of the sounds, etc. This can certainly be supported by the mechanics of inner speech: psychologists normally distinguish between a production system and the perceptual/comprehension system of inner speech, in a similar way as outer speech[10]. For example, McGuire, PK et al (1996) show that the brain areas which are activated in inner speech and imagining speech differ with respect to perception areas, while having the same brain area for speech production. There are also some studies that show that both elements are separable, so people born without the ability to make use of the speech apparatus and people born without the ability to hear may develop forms of inner speech (see Bishop (1985); Bishop (1988); Bishop and Robson (1989)). The semantic element thus is the responsible of the meaning of the string of words, whereas 'sensory' refers to all the other non-semantic elements present in inner speech: syntactic and phonologic elements, etc. The crucial question here is whether the *mechanics* and the separation in these two systems entail a *phenomenological* difference between sensory elements and semantic ones. Prinz seems committed to answering yes, but this does not seem what in fact occurs when we experience inner speech: we do not experience a string of sounds and afterwards an interpretation of them, but rather the *unity* of both components in the string of inner speech. There are at least two reasons to think this: (i) the interval of time for going from one system to the other is too small to be phenomenologically significant and (ii) restricting phenomenology to the sensory aspect of inner speech would amount to equating cases in which one repeats phrases or words without any sense, purely sensory streams of inner speech, with standard cases of inner speech in conscious thought. If the entailment from the mechanics of inner speech to its phenomenology is not warranted, as I have argued, any appeal to inner speech as a candidate for sensory reduction cannot succeed.

Second, I want to consider the problem from what I call '*phenomenological adequacy*'. It should be obvious that the phenomenal character of a certain mental episode “shows us” or “gives us” what it is like to undergo this episode, precisely because phenomenology is a matter of feeling a certain way and the most accepted definition we have so far of phenomenal character, though uninformative and controversial[11], is the what it is likeness for the subject (Nagel, 1974; Block, 1995). The point is then

[10] See Vicente, A. and Martinez Manrique, F. (2010) for the claim that inner speech shares fundamental properties with outer speech.

[11] See Kriegel (forthcoming) for an overview of the problems of this characterization.

that the *explanation* of cognitive phenomenology in sensory terms does not correctly *describe* what it is like to think, what is to have an experience of a cognitive episode. This is a general phenomenological point that receives support from the following idea. Which is the phenomenal element that marks the experience as one of thinking and not seeing or hearing? When we consciously consider a thought, for example, or understand something, or ask ourselves whether we know something, we seem to be able, at least, to distinguish that very episode from our current visual perception *solely* on the basis of experience. Just by way of undergoing the cognitive episode, we can at least pick up the episode as one of thinking and not just hearing or seeing something. Contrary to this, reductionist or restrictivist proposals are not in a position to distinguish thinking experiences from sensory or emotional ones on the basis of experience, because the sensory phenomenology associated with a cognitive episode and with a visual perception can be the same: if we see an ice-cream and if we consider whether this ice-cream is too expensive, and in both cases we have the same image of an ice-cream, the mere experiential character cannot differentiate between both mental episodes, according to the restrictivist.[12] One could resist the objection and argue that there still might be differences in both images that can account for the phenomenal difference between both episodes, but then notice that the most we have are some *sensory differences* between a visual perception and the mere entertainment of a proposition so that we are left with nothing else that makes us aware of undergoing a visual experience (and not thinking about it) or considering whether the ice-cream is expensive (we can imagine this situation in the absence of the visual perception of the ice-cream). Prinz could answer that in the case of considering a thought but not in the visual experience we might have some verbal imagery that explains the difference. The point is, again, that this verbal imagery is not in a position to tell us that we are considering and not just seeing, or desiring or remembering – where we could have the same verbal imagery. The dilemma his restrictivism is pushed towards is the following: either he accepts that we cannot differentiate between kinds of mental episodes on the basis of experience or that sensory phenomenology is typified in a way that can do the job, so that the sensory elements of cognitive phenomenology would be somehow "special" or sensory* (meaning: sensory of the kind involved in thought). Both horns of the dilemma do not seem to find support. In contrast, expansionism or views

[12] Notice that this argument resembles Pitt's (2004) epistemological argument for cognitive phenomenology, based on the premise that we know the content of our thoughts and we can distinguish one thought from another and from non-cognitive mental states. Besides showing a specific cognitive phenomenology, Pitt argues that this argument shows that there is a *distinctive* phenomenology (between the thought that *p* and the thought that *q*) and an *individuative* one (what determines the content of the thought that *p)* for thought. My point against restrictivism only assumes the capacity for distinguishing cognitive from non-cognitive episodes on the basis of experience, without the other more demanding requirements.

defending a specifically phenomenal character can accommodate this phenomenological fact by appealing to a cognitive phenomenal character specific for cognition or thoughts.

The third problem I see in Prinz's restrictivism can be called *'the limits of sensory variation'.* It has two sub-points: one the one hand, I will argue that sensory differences cannot account for the phenomenal *similarities* we find between thoughts of the same kind and, on the other hand, sensory differences cannot account for *all* the phenomenal differences we find in conscious thought. There is a motivation for this way of specifying cognitive phenomenology: talk of similarities and differences in phenomenal character has been important for progress in perceptual consciousness up to the point that for some authors, picking out similarities and differences in phenomenology is an essential condition for talking of experiential kinds and recognising a certain kind of phenomenal character (Martin (ms); Georgalis (2005)).

Consider the following example. You and me are standing in front of a field and we both see a flower and think the proposition that the flower is beautiful. What are the experiential *commonalities* and *differences* between me and you in the cognitive episode? One main intuition that the restrictivist cannot explain away is the idea that phenomenal variation in cognitive experiences is not always different but there seems to be some commonalities between my entertaining the thought that the flower is beautiful and your entertaining the same kind of thought. Restrictivist positions seem forced to say that sensory phenomenology cannot provide anything more than phenomenal differences, given the kind of variation that characterizes sensory imagery. But I think there is a sense in which we experientially have the *same* kind of thought, and this can be explained by the fact that our thoughts are *about* the same thing, namely, the flower that stands in front of us. A familiar picture of why this is so is the fodorian view of concepts as concrete mental particulars in the head that are constitutively linked to the world, and thus externally individuated. If these world-tied aspects of concepts have any contribution to phenomenology, I think they provide us with *commonalities* in cognitive phenomenal character, or at least it is not *in virtue of* the world-tied aspect of concepts that we have different cognitive experiential mental states. The differences in phenomenal character are provided by differences in *inner speech* (yours being a certain kind of tone and speed and mine another), *images* associated with this thought and possible feelings associated with it – and let's assume that they are reducible to perceptual phenomenology, as in Prinz's (2004) view. If we suppose, for the sake of the argument, that all these elements are the same in you and me, as are the concepts FLOWER and BEAUTIFUL externally individuated, can we still say that you and me have different experiential or

phenomenal character while entertaining the thought that the flower is beautiful?

Against restrictivism, I think the answer is yes. The sensory, perceptual and emotional elements in the episode of conscious thought do not suffice to explain the differences between you and me in the phenomenal character of the conscious thought, precisely because of the *network* in which our thoughts are embedded in our mental economies. My proposal is that the connections and the "situation" of the concept in our cognitive mental lives form a kind of network, that is needed to account for further cognitive experiential differences, over and above the elements mentioned. This network is constituted by the *background knowledge* one possesses about a certain concept and that we both may differently carry. The idea is that the more knowledge one has over a certain subject, the bigger the network is, and more differences in cognitive phenomenology we can find or the richer it is. The connections of this network are clearly not differences in sensory phenomenology, so if my proposal is sound, Prinz's restrictivist position is in trouble.[13]

One might object that we are talking about *occurrent* conscious thought, and the network is a set of dispositional concepts that cannot be experientially present when we entertain the thought, so they are elements that cannot account for the cognitive phenomenal difference between you and me. This proposal has to be presented in more detail, but the idea and response to this objection would be that the network is somehow felt with the occurrent concept you are thinking about, just in the sense in which one can say that there is phenomenal consciousness in the peripheral areas of the visual field that are not the focus of our attention.

The sketched proposal I have offered is a view on cognitive phenomenology that takes into account sensory variation (like restrictivism) but also explains the intuitions of commonality and further differences that are not just sensory, thus providing a specification of the nature of cognitive phenomenology that is not available to the restrictivist. If there is motivation to look for similarities and differences in phenomenal character in cognitive phenomenology and my account is sound, it shows the limitations of Prinz's restrictivism in both directions: in a nutshell, there are more phenomenal variations in cognitive phenomenology than restrictivism recognizes – in particular, cognitive experiential differences – and there are also commonalities that restrictivism *per se* cannot account for.

The limits of sensory variation, together with the problem with the phenomenology of inner

[13] Strawson (2011) thinks we should postulate an identical cognitive phenomenal character over an above the one determined by the world-bound aspect and the internal economy, so that we can account for the idea that, by hypothesis, me and my Twins (my Twin in Perfect Twin Earth where 'water' refers to XYZ, my Instant Twin which has just popped into existence and my Brain in a Vat Twin which has no external connection to the world) have the same qualitative character.

speech and the phenomenological adequacy problem discussed above constitute, to my mind, an important obstacle to Prinz's account of cognitive phenomenology. If his reductionist account of cognitive phenomenology cannot solve these problems, then his general theory of phenomenal consciousness is undermined, given the "all consciousness is perceptual" claim. One moral of this paper is that before giving arguments from unity and parsimony for a theory of consciousness, we could try to specify the nature of cognitive phenomenology and work from there on. I have suggested a way of doing so through phenomenological similarities and differences and have argued that it gives evidence for the defender of a specific cognitive phenomenology view.

* * *

References

Bishop, D.V.M. (1985). "Spelling ability in congenital dysarthria: Evidence against articulatory coding in translating between graphemes and phonemes". *Cognitive Neuropsychology*, 2:229§251.

Bishop, D.V.M. (1988). "Language development in children with abnormal structure or function of the speech apparatusÍ, in D. Bishop and K. Mogford (eds). *Language Development in Exceptional Circumstances*. Edinburgh: Churchill-Livingstone.

Bishop, D. V. M and Robson, J. (1989). "Unimpaired short-term memory and rhyme judgment in c ongenitally speechless individuals: Implications for the notion of articulatory coding." *Quarterly Journal of Experimental Psychology*., 41A:123§141.

Block, N. (1995). "On a confusion about a function of consciousness", *Behavioral and Brain Sciences,* 18: 227-87.

Bayne, T. (2009). "Perception and the reach of phenomenal content", *Philosophical Quarterly,* 59: 385-404.

Bayne, T. and Montague, M. (eds.)(2011). "Cognitive Phenomenology: and Introduction", in Bayne, T. and Montague, M. (2011). *Cognitive Phenomenology.* New York*:* OUP.

Georgalis, N. (2005). T*he primacy of the Subjective: Foundations for a Unified Theory of Mind and Language.* Chapter 3. MIT Press.

Husserl, E. (1900-1901/1970). *Logical Investigations*. London and New York: Routledge.

Kriegel, U. (forthcoming). *The Varieties of Consciousness. Studies in non-sensory phenomenology.*

McGuire, PK et al (1996), "Functional neuroanatomy of verbal self-monitoring". *Brain*, 119:907§917.

Martin, M. (manuscript). *Uncovering Appearances.*

Nagel, T. (1974). "What is it like to be a bat?", in *Philosophical Review,* LXXXIII: 435-50.

Peacocke, Ch. (1998). "Conscious attitudes, attention, and self-knowledge", in Wright, C., Smith, B. and MacDonald (eds.). *Knowing our own minds.* OUP.

Pitt, D. (2004). "The phenomenology of cognition, or, what is it like to think that p? *Philosophy and Phenomenological Research*, LXIX, No. 1, 2004

Prinz, J. (2002). *Furnishing the Mind: Concepts and their Perceptual Basis.* Cambridge, MA: MIT

Press.
Prinz, J. (2004). *Gut reactions: A Perceptual Theory of Emotion.* New York: OUP.
Prinz, J. (2005). "A neurofunctional theory of consciousness", in A. Brook and K. Akins (eds), *Cognition and the Brain: Philosophy and Neuroscience Movement.* Cambridge: Cambridge University Press, pp. 381-96.
Prinz, J. (2007). "All consciousness is perceptual", in B. McLaughin and J. Cohen (eds), *Contemporary Debates in Philosophy of Mind.* Oxford: Blackwell, p. 335-57.
Prinz, J. (2011). "The Sensory Basis of Cognitive Phenomenology", in Bayne, T. and Montague, M. (eds)(2011). *Cognitive Phenomenology.* New York*:* OUP.
Prinz, J, (2012). *The Conscious Brain.* OUP.
Siewert, Ch. (1998). *The Significance of Consciousness*. Princeton University Press.
Siewert, Ch. (2011). "Phenomenal Thought", in Bayne, T. and Montague, M. (eds)(2011). *Cognitive Phenomenology.* New York*:* OUP.
Smithies, D. (forthcoming). *"*The nature of cognitive phenomenology", in *Philosophy Compass.*
Strawson, G. (1994/2010). *Mental Reality.* 2nd ed. Cambridge, MA: MIT Press.
Strawson, G. (2011). "Cognitive phenomenology: Real life", in Bayne, T., and Montague, M (eds). *Cognitive Phenomenology*. New York: OUP.
Tye, M., and Wright, B. (2011). "Is there a phenomenology of thought?", Bayne, T., and Montague, M (eds). *Cognitive Phenomenology*. New York: OUP.
Vicente, A. and Martinez Manrique, F. (2010). What the...? The role of inner speech in conscious thought. *Journal of Consciousness Studies,* 17, issue 9-10, pp. 141-167.

Are Emotions Valent Embodied Appraisals?

David Pineda (Logos; Universitat de Girona)

The ultimate aim of my contribution is to argue that appraisal theories of emotion (Lazarus 1991, Scherer et al. 2001, Ellsworth and Scherer 2003) fare overall better as accounts of emotions than the perceptual embodied appraisal theory as defended in Prinz 2004, 2007 and 2008. I will try to make my case by relying on three argumentative strategies. First, I will point to certain problems that are inherent to Prinz's view. Second, I will discuss certain crucial issues about emotions and argue that they can be better accommodated by appraisal theories than by Prinz's alternative. Third, I will try to defend appraisal theories from Prinz's main criticisms.

In this paper, I will first present the bare bones of the two theoretical contenders: Prinz's perceptual theory, on the one side, and appraisal theories, on the other (I will be especially brief regarding appraisal theories). Secondly, I will mention, and briefly discuss, some of the arguments I'm going to use corresponding to the three strategies just announced.

1. Prinz's theory.

In his book *Gut Reactions* (Prinz 2004), Jesse Prinz has defended an original perceptual theory of emotion. Part of the originality and interest of the theory lies in the fact that it cleverly integrates elements of so-called cognitive theories of emotions (Solomon 1976, 2003, Nussbaum 2001) together with elements of the James-Lange theory (James 1884), two main theoretical approaches to emotion often thought to be antagonistic. The theory is also compelling because it is argued for after a careful examination both of philosophical considerations and extant empirical studies about emotions.

The first component in Prinz's theory is definitely Jamesian. Prinz argues that (at least in basic cases, more on this later) bodily changes frequently associated with emotions (facial expressions, vocal and musculo-skeletical changes, and changes in the Autonomous Nervous System and the Endocrine System) actually precede emotion rather than following it. James was then right to hold that bodily changes are causes of emotion and not effects thereof. Actually Prinz claims that emotions are perceptions of bodily changes.

There seem to be three main reasons for this Jamesian first conclusion. First, Prinz simply accepts James' "subtraction argument". According to James, if we fancy a strong emotion and abstract from it all feelings of bodily disturbances what we are left with is definitely not an emotion. Prinz reads this thought experiment as showing that the phenomenology of emotion is exhausted by feelings of bodily changes. This is of course a conclusion which squares perfectly well with the view that emotions are perceptions of bodily changes but can hardly be accounted for by cognitive theories according to which an emotion consists of some sort of evaluative judgment or appraisal of the stimulus and as such is a direct cause of the corresponding bodily changes.

Second, he accepts Robert Zajonc's view that emotion and cognition involve separate neuroanatomical structures. In *Gut Reactions*, Prinz mentions Joseph LeDoux's findings about fear (LeDoux 1996) as providing good evidence for this conclusion. LeDoux found out that fear responses to snake-like objects are entirely processed through subcortical regions of the brain. It seems that when the thalamus has the information that the stimulus might be a snake (the thalamus cannot make fine discriminations, the primary visual cortex is required for that task) it sends a signal not just to the primary visual cortex but to the amygdala as well. The amygdala is another very important subcortical structure which is known to orchestrate all by itself the sort of bodily changes typically involved in episodes of fear (Damasio 2010). As the amygdala gets activated the aforementioned changes ensue and usually a typical withdrawal behavior follows quite quickly, before or just when the signal reaches the primary visual cortex. This is why one can sometimes find himself stepping back from a coiled object at the same time one realizes it is not an snake but, say, a house pipe. Now assuming that subcortical brain regions do not implement tasks which require the use of concepts, LeDoux's evidence would then show that some fear responses occur without the mediation of cognitive states such as those that would be required for the sort of appraisals and evaluations postulated by cognitive theories. Of course, LeDoux's evidence, by contrast, is entirely consistent with the Jamesian view that a state of fear is just the perception of the bodily changes orchestrated by the amygdala.

Thirdly, Prinz also endorses the claim, which was already put forward by Karl Lange, that emotions can arise by direct physical induction. The administration of certain drugs seems to have the power to change the emotional state. Consider for instance the ingestion of alcohol and its emotional effects. There seems to be also some evidence to the effect that voluntary acquisition of facial expressions characteristic of the expression of certain emotions actually gives rise to the corresponding emotion (Zajonc et al. 1989). This is again something which a Jamesian theory can perfectly account for. The explanation would be, for instance, that certain drugs have the power to provoke the sort of bodily changes the perception of which is the emotion.

Yet, although for these three reasons Prinz thinks that emotions follow bodily changes and are actually perceptions of these changes, his view is not entirely Jamesian. He claims that emotions are perceptions of bodily changes but they do not represent the bodily changes they perceive. He introduces a subtle distinction between registering and representing to make that crucial point clearer. A mental state, he says, registers that which reliably causes it to be activated. Yet representing is defined drawing on ideas of Dretske and Millikan: a mental state represents that which it has the function to carry information about. Or to put it in the concise terms that Prinz likes to use: a mental state represents that which it is set up to be set off by.

Now according to Prinz, emotions are definitely not set up to carry information about bodily changes. This view, he thinks, cannot adequately explain why emotions were naturally selected as they conferred some sort of survival advantage. He argues that emotions are used to promote certain specific behaviors which are unintelligible if we

assume that they represent bodily changes. For instance, in many cases fear compels us to run away from the eliciting stimulus, but to say that we run away because we feel that certain changes are taking place in our body makes little sense. In an interesting twist in the discussion between cognitivist and Jamesians he claims that emotions represent *core relational themes*. This is, surprisingly enough, a technical term directly borrowed from Richard Lazarus' cognitivist appraisal theory of emotion. Lazarus famously argued that an emotional episode starts when the stimulus is appraised by the organism according to several appraisal dimensions (more in this in the next section, where I introduce appraisal theories). Moreover, an emotion type is individuated by the results of the appraisal process in each of these dimensions. A core relational theme is introduced then by Lazarus as a sort of summary of the results for each appraisal dimension. Therefore, for each emotion type (fear, anger, guilt, etc.) we will find one individuating core relational theme. For instance, in the case of fear, the core relational theme, according to Lazarus, is "facing a danger" (Lazarus 1991, p. 122).[1]

As Prinz rightly stresses, core relational themes gloss the bearing of a stimulus on the well-being of the organism. To mention other prominent examples, offense would be the core relational theme of anger and loss that of sadness. If, as Prinz wants, an instance of an emotion type, say a case of fear, represents its core relational theme, then we can understand what kind of survival advantages did emotions confer on our ancestors. Fear would then be a mechanism set up (by evolution) to detect dangers. And we can also understand why emotions compel us to act in certain ways. For instance, it is most reasonable to fly away from an impending danger.

Prinz's central claim is then that emotions represent core relational themes by registering bodily changes. This distinction is further elucidated by Prinz's previous theory about natural kind concepts (Prinz 2002). Consider the concept of dog. This concept applies correctly to something X only if X has a certain complex biological property, a certain genome. Yet, we humans do not have genome detectors. How does our concept then manage to track this biological property? Prinz's answer is that we actually register certain apparent properties (being four-legged, barking, etc.) which are actually caused by the genome in question. To the extent that there is some sort of robust relation between the appearances and the referent we may then develop concepts which track this referent simply by directly registering the relevant appearances. Prinz calls the appearances the nominal content of the concept and the referent the real content. A similar story could be told for the concept water. The nominal content being in this case something as being liquid at certain temperatures, falling from the sky under certain atmospheric conditions, etc., and the real content the property of being H2O. The idea is then that, for each emotion, the real content is its core relational theme while the nominal content would be the relevant bodily changes. This is why he speaks of emotions as being "embodied appraisals", since they are supposed to represent relations that bear on the well-being of the organism by registering bodily changes.

[1] To be precise, Lazarus distinguishes between anxiety and fright, but we can ignore this complication for the moment. I will return to the issue about types of fear when discussing my arguments.

2. Appraisal theories of emotion.

Appraisal theories constitute one of the main approaches to emotion in contemporary psychology. Although they can be traced back to the ancient Stoics, psychologists often mention the work of Magda Arnold as the seminal source of the view (Arnold 1960). The leading idea is that the emotion undergone by an organism depends crucially on how the organism interprets the stimulus rather than on the nature of the stimulus. This is supposed to explain well-known facts about emotions, namely, that the same stimulus may elicit different emotions in different subjects, or even in the same subject at different times, and that a stimulus may or not elicit an emotion depending on the organism facing it.

Prinz treats appraisal theories in psychology as belonging to the same class, for the purposes of his main argument in *Gut Reactions*, as classical cognitive theories defended by philosophers such as Solomon or Nussbaum. This move assumes that appraisals or interpretations of a stimulus involve some sort of cognitive states and the deployment of axiological concepts. Although Richard Lazarus, probably the most influential appraisal theorist among contemporary psychologists, can perhaps be interpreted as espousing such a view the fact is that some of the current appraisal theorists do not follow him in this respect (see for instance Scherer 2009). This will be of importance later on.

One aspect which is present in most versions of the appraisal theory and is crucially different from such theories as Solomon's and Nussbaum's is that emotions are taken to be causal processes rather than simple states. So, strictly speaking, we should refer to emotional responses as emotional episodes rather than emotional states. According to this view, an emotional episode starts when the organism makes an appraisal of the stimulus.[2] As happened with Lazarus, current versions of this theory also hold that the appraisal is itself a process which involves the evaluation of the stimulus along a series of parameters or appraisal dimensions. Versions of the general view differ as to exactly which and how many dimensions to count in (see Scherer et al. 2001 for a survey of appraisal theories). By way of illustration, most of them typically include as appraisal dimensions: goal relevance (whether the stimulus, or an aspect of it, bears on some goal or need of the organism); goal conduciveness (whether the stimulus helps to promote some goal or need or it rather obstructs it); coping potential (an estimation of the capacity of the organism to change or modify or conveniently deal in some other way with the stimulus or its relevant consequences, --something which turns out to be crucial

[2] There is a characteristic hesitation among appraisal theorists on whether to count this appraisal as the initial component of the emotional episode or rather as its triggering cause. For reasons I cannot dwell into here, I think the first option is better than the second. But nothing of consequence follows from this for the purposes of this paper.

when the stimulus has been appraised as obstructive). According to some models, these appraisal dimensions are then processed sequentially. This is all the more reasonable, since some appraisal dimensions seem to require the result of others in order to start on. For instance, an estimation of the coping potential seems to require an output result for the dimension of goal conduciveness as one of its inputs.

The upshot of this appraisal process is the bringing about of a set of characteristic effects. Most models count among them physiological responses (for instance changes in the endocrine system), motor expressions (for instance, facial expressions) and action tendencies (for instance, a tendency to approach or withdraw from the stimulus). This array of effects is supposed to occur more or less at the same time and together constitute the second main stage in the emotional process or episode.

This second stage causes in its turn the third one, which would consist in a representation of most of the elements involved in the previous stages. This representation is thought to be phenomenally conscious.[3] This would then in sum amount to the feeling component of the emotional episode, according to these models. The function attributed to this stage ranges from being required for the purposes of communication to being a monitoring device of the whole process which improves its accuracy, efficiency and flexibility. Some models mention also that the modification or suppression of certain emotional behaviors (which might be due to some personal strategic reason or to social norms and pressures) also requires that most elements of the emotional episode be adequately represented by the mind. Finally, some models add also a fourth final stage of verbalization, but we do not need to consider it here.

3. Arguments.

Now that both contenders –Prinz's embodied appraisals view and appraisal theories— have been summarily presented, let me then mention some of the arguments which build my case that appraisal theories are to be preferred to the embodied appraisals view. In the introduction I mentioned three argumentative strategies. I will therefore start with the first one: problems inherent to Prinz's view.

3.a Problems of articulation of Prinz's view.

Perhaps the most important in this score is that there seems to be a serious tension between Prinz's view of basic emotional responses as having an evolutionary origin and his Jamesian view that bodily changes precede and actually cause emotions. Let me explain.

Let's go back to the registering / representing distinction and to the central claim that a core relational theme is the real content of an emotion type whereas a certain syndrome of bodily changes is its nominal content. The emotion tracks a core relational theme by

[3] According to some models, however, not all elements represented need to be consciously represented (see for instance Scherer 2004). I will not dwell into these niceties here.

registering a certain syndrome of bodily changes. Given how the nominal / real content distinction is explicated this thesis can only be sustained if the bodily changes in question are reliable indicators of the given core relational theme. Compare: the dog concept works only if the appearance properties of dogs trough which we track the dog's genome are reliable indicators of the dog's genome. And this is supposed to be so since, in fact, these appearance properties are caused to be instantiated (in normal conditions) by the dog's genome.

So why are the bodily changes reliable indicators of the core relational theme, say, danger, in the case of fear? Prinz's answer is that in the basic cases (which are supposed to be basic because all the rest of cases will be explained in terms of them, more on this later on) there is an evolutionary explanation for this reliable connection:

"Evolution has undoubtedly endowed us with distinctive physiological responses to various situations that our ancestors encountered. The heart is predisposed to race (along with several other physiological responses) when we see looming objects, snakes, crawling insects or shadows at night" (Prinz 2004, p. 69).

So the general idea is that snakes, crawling insects or shadows at night are innate themes for fear.[4] Evolution has so designed our central nervous system, by way of the natural selection process, that when we perceive any of these themes then a whole syndrome of bodily changes follow. It is evolution then what guarantees that this syndrome is a reliable indicator of any of these themes and a fortiori of danger. But why is such a connection adaptive in the first place? The answer is that this syndrome of changes is known to prepare the organism for an appropriate response to a danger or threat. A racing heart, for instance, enables us to run away. And it is of course highly adaptive to have been provided with a mechanism which enables us to escape from dangers in our (evolutionary) environment.

So far so good, but this does not tell us anything about the natural selection of emotions themselves. Recall that, according to Prinz's view, the emotion is the perception of the syndrome of bodily changes. According to what has been said so far, emotions are idle aspects of the process going from the perception of a theme of fear (say, a snake) to the performance of some behavior adequate to the challenge posed (say, running away). Traits of organisms are naturally selected in virtue of certain effects they have which turn out to be highly adaptive in a given environment. One would like to say (and this is what most people thinking that at least some emotions have an evolutionary origin usually say) that an emotion was selected because it somehow enabled or promoted behaviors which were appropriate for dealing with the sort of challenge posed to the organism by a given stimulus type. For instance, one would hypothesize that fear was selected as a promoter of appropriate responses to dangers. Yet, on Prinz's view the bodily changes which actually enable this sort of adaptive behaviors occur *before* and not *after* the emotion of fear, and they are *causes* of fear rather than *effects* of fear.

[4] I borrow the terminology here from Ekman (2003).

One may think that this is not a fatal objection to Prinz's general view. One might retort that the previous reasoning only shows that emotions, as perceptions of bodily changes, were not selected as causes of these changes, according to Prinz, but as they had some other effects. Perhaps, just as the appraisal theorist tends to think of the feeling component of the emotional process, Prinz thinks that the effect of emotion which turned out to be adaptive was that of monitoring the bodily changes and therefore of rendering the so called emotional behaviors more flexible and amenable to modification, an effect which would not be in place if the bodily changes were not represented through emotion.

This is not, however, what Prinz seems to have in mind. As I pointed out in the first section, emotions are not supposed to have the function, in the evolutionary sense, to monitor bodily changes and whatever effects derive from this monitoring function. Prinz says that emotions do not represent bodily changes. He says that emotions do not have the function of carrying information about bodily changes; they instead have the function of carrying information about core relational themes. And recall that for Prinz, to carry information about something is just to be reliably caused by this something. So, an emotion like fear is supposed to have the function of being reliably caused by dangers. This is however a bit unfortunate. Again, traits in an organism are not naturally selected for what causes them but actually for some of their causal effects. So fear could not have been selected for being caused by dangers. If one wants to elucidate the notion of representation in terms of functions in the teleological sense and one wants to say that fear represents danger, as definitely Prinz wants to do, the natural move to make is to argue that fear was selected because it promoted behavior which was somehow suitable for dangers. And then we stumble again upon the same problem: the bodily changes enabling the behavior in question are actually the cause of fear, according to Prinz, and not its effect. It looks as if Prinz is at this critical juncture putting the cart before the horses.

This is then the essential tension I mentioned at the beginning of this section. There are two claims which together build up Prinz's characteristic account of emotions: 1) An emotion represents its core relational theme; 2) An emotion is caused by a syndrome of bodily changes which prepare the organism for an adequate response to the instantiation of its core relational theme. The problem is that these two central claims pull in opposite directions if representation is spelled out in terms of functions in the teleological sense, just as Prinz intends to do.

I see also another problem of articulation, which complements with the one just discussed, when Prinz tries to account for non-basic cases of emotion. In Prinz's theory there are two non-basic cases to consider: one brand consists in cases in which a basic emotion like fear is elicited by a stimulus other than an innate theme; the other concerns non-basic emotions. According to Prinz, there is only a limited pool of basic emotions, fear among them, and the rest are derived from the basic ones through two different processes: blending and calibration.

I will make two points about Prinz's treatment of non-basic cases: i) his account of how learnt elicitors of basic emotions arise is doubtful; ii) all non-basic cases are explained as cases in which a syndrome of bodily changes consciously represented is caused by an appraisal of the stimulus. If (ii) is correct, then Prinz's theory looks as something pretty close to what appraisal theorists defend. The crucial difference, of course, is that for Prinz the non-basic cases depend on the basic ones, and actually arise out of them through the sort of mechanisms I'm about to discuss. But as I have just argued, Prinz's account of basic cases involves a tension which renders his account less than compelling.

Let's therefore begin with the first case of non-basicness: elicitation of a basic emotion by a stimulus which is not an innate theme. Consider, to take one of Prinz's examples, being afraid of an exam. It is of course absurd to construe these cases as ones in which evolution has secured some causal link between the mental categorization of a situation as being an exam and the syndrome SF of bodily changes characteristic of fear. Prinz suggestion then is that the thought 'this (the exam) is dangerous (or threatening)' becomes a reliable cause of SF (which in its turn, as usual, causes the state of fear) through a process of associative learning which draws on the basic cases of fear elicited by some of its innate themes.

The problem is that Prinz does not spell out in detail how this learning mechanism is supposed to work. At some point (Prinz 2004, p. 76), he hypothesizes a possible "developmental sequence":

"At some point, while experiencing fear in a darkened room, we entertain the verbally mediated thought that we are facing a dangerous situation. This happens on a number of subsequent occasions. At first, the thought "I'm in danger" is an effect of fear (…) But, through associate learning, that thought becomes a trigger for fear as well. Eventually, the explicit thought "I'm in danger" becomes capable of initiating fear responses in situations that lack the physical features that are predisposed to upset us as a function of biology".

So the general idea is that at our early stages of development we experience states of fear as a result of frequent encounters with fear's innate themes (darkness, snakes, crawling insects). We then develop a concept of danger as a consequence of experiencing all these states. This is supposed to be a concept which "captures the features unifying" these themes (íbid.). As a result of this process, the concept of danger becomes strongly associated with experiences of fear in such a way that the mere application of the concept to a stimulus which is utterly different in nature to any innate theme becomes a triggering cause of an experience of fear.

Now this account makes a number of assumptions about how the concept of danger is acquired which should in any case be empirically confirmed (and the same goes, of course, for the other concepts involved in the rest of basic emotions (such as offense or loss, for instance). On the face of it, a number of questions spring to mind. Is it really required, in order to acquire the concept of danger, that young children experience

relatively frequent cases of fear as a result of encounters with innate themes of fear? What if a young child is lucky enough not to encounter, or at least not frequently enough, such themes as snakes or shadows at night? Will she then not develop a concept of danger? Or will she then acquire a concept of danger which won't become associated with states of fear? We'll have to wait and see whether future research in developmental psychology helps to answer these questions in the way required by Prinz's theory.

Moreover, Prinz claims that danger is introduced as a concept which denotes the unifying features of the innate themes of fear. But, on the face of them, these themes are utterly different regarding their physical nature. Snakes and darkness, for instance, have little in common in this respect. Furthermore, there are perhaps countless ways of grouping together snakes and darkness, but many of them will not group them together with guns or exams (to name a few fear elicitors which are not innate).

Let's now move on to the second type of non-basic cases: Prinz's account of non-basic emotions. Non-basic emotions are supposed to arise out of basic ones by the effect of two different mechanisms: blending and calibration. Blending is just the combination of two basic emotions. For instance, Prinz conjectures that contempt may be a blend of anger and disgust (Prinz 2004, p. 144). This would of course entail, following Prinz's general theory, that contempt consists in the simultaneous perception of the syndrome of bodily changes SA of anger and the syndrome SD of disgust. This is, as far as I know, an empirical consequence that remains to be tested. Be that as it may, Prinz's claims about blending appear very tentative and one has the impression that the mechanism which does more theoretical work is calibration.

Prinz draws again on an idea of Dretske to make clear what is meant here by calibration. According to Dretske, evolved representations can be sometimes put to new uses. Prinz comments on an example outside of the mental realm and derives the intended implication for his theory of emotion:

"Coughing has the evolved function of clearing the throat. But a spy might also use a cough as a secret code in communicating with an accomplice. A spy's cough might represent the fact that the microfilm has been delivered. Likewise, an embodied appraisal that usually represents a demeaning offense (anger) may represent an infidelity (jealousy) when used under the direction of the right judgment. We can recalibrate our embodied appraisals to occur under conditions that are somewhat different than those for which they were initially evolved" (Prinz 2004, p. 99).

In this explanation, it is assumed that anger is a basic emotion which represents offenses and consists in the perception of a given syndrome of bodily changes SA. Then jealousy, a non-basic emotion, is supposed to arise as the bodily changes SA, which are originally calibrated to be caused in general by offenses, get recalibrated so as to be caused also by thoughts of infidelity.

Again this mechanism of calibration raises a number of questions but I will concentrate here on two. Firstly, one may quite naturally ask why these recalibration processes do

occur in the first place. Why are the changes SA recalibrated to be caused by infidelity thoughts? There must be, it seems, some non-accidental connection between the core relational theme represented by anger –offense—and that represented by jealousy – infidelity. This is what Prinz actually suggests. We must recognize, he says, that infidelity involves an offense (see Prinz 2004, p. 148).

The second concern is more serious. Given the answer to our first question there appears to be a serious problem in the way that Prinz individuates non-basic emotions which arise out of a calibration process. If jealousy consists in the perception of the same SA bodily changes as anger and it is also required that the subject judges that infidelity situations are offensive situations, then, given the fact that offense is the core relational theme of anger, one wonders why jealousy is reckoned as a distinct, albeit non-basic, emotion instead of the same old basic emotion of anger.

It looks as if Prinz is individuating here jealousy by a given appraisal of the eliciting situation –namely as one in which some infidelity is involved— and by the fact that this appraisal causes some syndrome of bodily changes and the mental perception of them. Once again, this is exactly what the appraisal theorist does in general. The problem is not merely that recalibrated non-basic emotions turn out to be something quite close to what the appraisal theorist has in mind. The deep problem here is that Prinz seems to be using the appraisal theorist's way of individuating emotions and disregarding his own.

Prinz tries to evade this problem when he claims that the judgment that one's lover has been unfaithful need not always be the cause of jealousy:

"Jealousy can be triggered by the judgment that one's lover has been unfaithful, but it can also be triggered by other judgments, such as the judgment that one's lover has been staying unusually late at work. Jealousy can even be triggered by perceptual states, such as the smell of an unfamiliar perfume on a lover's clothes" (Prinz 2004, p. 101).

This is not, however, a good way of evading the problem. Of course, there are countless judgments or perceptions which may give rise to jealousy, as they are countless many others that may give rise to fear, anger or sadness. But of course not anything goes. Only when the organism appraises any of these ways as involving infidelity will jealousy ensue. Otherwise it will not. So this appraisal is unavoidably crucial, it appears, to individuate jealousy.

So the upshot of this section is as follows. The explanation of basic cases of emotion in Prinz's theory is unstable, since its two characteristic claims –that emotions represent core relational themes and that emotions are caused by bodily changes—pull in opposite directions. On the other hand, the explanation of non-basic cases is problematic on two counts: first, the mechanisms by which non-basic cases are derived are unclear and problematic; second, some of them at least seem to involve ways of individuating emotions which are those of the appraisal theorist and certainly not those which would follow from Prinz's embodied appraisal account.

3.b Issues better explained by appraisal theories.

Let me now mention two prominent issues about emotions which I think appraisal theories are better equipped to account for than Prinz's theory.

Some of them have to do with what I would like to call the "complexities of the emotional response". In his *Gut Reactions*, Prinz complains that appraisal theorists have mistakenly build the sort of complex properties represented by emotions –danger, offenses, losses, etc.- into the structure of emotions themselves. He is relying here again on another Drestkean point. Consider this example. A simple, unstructured beep emitted from a fuzz buster represents the presence of a police radar. The beep itself is not decomposable in one part meaning "police", another meaning "radar", etc. So the moral is that the complexity of the property represented need not be reflected in the structure of the thing representing it. Appraisal theorists, according to Prinz, have overlooked this possibility in thinking that emotions must be constituted by a complex array of appraisals because of the sheer fact that emotions represent properties of a similar complexity.

But I do not think that this complaint is fair enough. The reason why appraisal theorists take emotions to be highly structured processes is not that they are wrongly assuming that representations must have as much structure as those things they represent. It is rather the fact that the complexities of emotional response and behavior can be hardly explained unless emotional episodes are richly structured. Let me elaborate a bit on that.

Consider once again the case of fear. Not all fear episodes lead, or orient, to the same sort of behavior. When we are afraid of something, we do not always run away from it. In us, as in many other animals, also a fight response is preferred in some cases. And there are still cases in which fear causes a characteristic freezing behavior (which can be conjectured to attempt at camouflage or perhaps to deceive the predator). The appraisal theorist, precisely thanks to the rich structure of the appraisal process which according to her sets off the whole emotional episode, has ways to account for this differential response in different fear episodes. For instance, she can say that a fight, flight or freezing response may critically depend on the result of the appraisal dimension of coping potential.

Prinz is aware of the differential response in different fear episodes. What he says is that there are different types of fear, each one of these types being the perception of a different syndrome of bodily changes. He is clearly assuming that there is a general syndrome of bodily changes which are common or central to every fear episode and then there are some differences which allow us to distinguish between types of fear. This is reasonable, but it shows that Prinz is himself honoring my point: the complexities of emotional response can only be explained by attributing to emotion a more or less rich structure. In his case, this structure is to be discerned in the

relationships between the different syndromes of bodily changes which according to his view individuate emotions.

Yet this may not be sufficient. There are further complexities involved in emotional behavior which are not that easily explained by an account such as Prinz, but can be nicely explained in the framework of appraisal theorists. For instance, it has been suggested that subjects who, for whatever reason, have a low self-esteem will tend to deliver low results in the appraisal dimension of coping potential. As a result of that, appraisal theories predict that such subjects will tend to experience more frequently emotions which are characterized by an appraisal of low coping potential, like sadness (Van Reekum and Scherer 1997). It has also been argued that the optimism-pessimism personality dimension provides predictable systematic biases for the appraisal dimension of goal conduciveness (Scheier & Carver 1985). It's hard to see how the incidence of personality traits, or temporary personal conditions, in the emotional life of an organism can be accounted for on Prinz's account.

The second issue I want to bring up has to do with the connection between emotion and motivation. Embodied appraisals are supposed to be states which track core relational themes by registering changes in the body. They are therefore doxastic states with a mind-to-world direction of fit (Searle 1983). They are states characterized by their tracking, detecting, representing states of the external world and conditions of the body. Yet emotions are commonly thought to be strong motivators for action. Emotions play a crucial role in motivation. This is surely why emotion is a chief topic in psychology since motivation seems to be the key to understand behavior. If emotions were embodied appraisals the role of emotions in motivation would be overshadowed.

Prinz is perfectly aware of this and his final proposal is that emotions are not mere embodied appraisals but embodied appraisals coupled with a valence marker. A valence marker is taken to be a mental state with an imperative content. There are supposed to be two valence markers with the imperative content "More of this!" or "No more of this" corresponding to a positive or negative valence respectively. All emotions have either positive or negative valence (for instance, sadness, anger or fear have negative valence; pride or joy are positive) and Prinz claims that this is so because all emotions include either a positive or a negative marker.[5] Thus Prinz's final and complete theory is that an emotion is a valent embodied appraisal.

There are many points of interest in this construal. For instance, the idea of mental states with an imperative content is particularly intriguing, but I must leave discussion of this for a better occasion. The point I want a make now is that Prinz's construal of emotion as a conjunction of two utterly distinct types of mental states faces the usual problem with such conjunctive theories: nothing in principle prevents one of the conjuncts to be instantiated in the absence of the other. There might be cases, if Prinz's theory were right, in which the sort of embodied appraisal belonging to fear or sadness

[5] There might be emotions with a mixed valence, for instance nostalgia, but we can put these cases aside for the purposes of this paper.

is instantiated in the absence of its negative marker or, worse still, together with the positive marker "More of this!". One can perhaps make sense of exceptional cases in which sadness or fear enjoy a positive valence, I very much doubt this, but let's assume that this is so. The problem, however, is that since embodied appraisals and valence markers are completely different mental states, with anything interesting in common (the former having indicative content and the latter imperative content), I can't see nothing in Prinz's theory that prevents this sort of cases from being the norm rather than the exception. And this is surely an implausible consequence.

Appraisal theories, I think, fare again better in this regard. Appraisal theorists usually explain valence in terms of the net result of some appraisal dimension. For instance, it has been explained as the result of the appraisal dimension of "goal conduciveness", positive valence arising out of an estimation of goal congruence and negative valence out of an estimation of goal obstruction (Lazarus 1991). Other theorists have instead suggested an appraisal dimension of "intrinsic pleasantness", one of the first dimensions to be processed, which would estimate whether the stimulus is expected to produce pleasure or pain, broadly construed (Scherer 2001).

Prinz's dismisses this sort of explanations as being too "overly cognitive" (Prinz 2004, p. 168), but his judgment seems to be based on the premise that appraisals always involve sophisticated, cognitively demanding mental states, a premise simply unshared by most appraisal theorists (more on this in the next section).

The point I want to stress is that appraisal theories seem to have the resources to explain valence as one natural consequence of the very appraisal process and therefore it can be sustained that the valence of a mental episode is inherent to it. This I think is a more promising account.

3.c Reply to Prinz's Jamesian arguments.

So far I have raised a number of objections to Prinz's account of emotion and I have mentioned some crucial issues which I think are better explained by appraisal theories. Yet, as I said at the beginning of this paper, Prinz has offered three main reasons for his Jamesian conclusion that bodily changes precede and actually cause emotions rather than the other way around. This runs not only against common sense, as James knew perfectly well when he claimed that his view was contrary to the "natural way of thinking", but it is definitely incompatible with the central claim of appraisal theories. For, according to these theories, bodily changes involved in an emotional episode are mainly efferent effects produced by the result of different appraisal dimensions. I will then conclude by examining Prinz's three reasons. My conclusion will be that they do not overall constitute a case against appraisal theories, although some of them may point at matters of concern which need further development and refinement.

The first reason is James' subtraction argument according to which the phenomenology of an emotional episode is exhausted by feelings of bodily changes. This is a crucial premise for Prinz's theory. If the phenomenology of an emotional episode were not

exhausted by feelings of bodily changes, then it couldn't be sustained that an emotion is (leaving aside valence) a perception of bodily changes. Yet I expect many defendants of the existence of cognitive phenomenology to be unimpressed by James' argument. These thinkers claim that there is something it is like to think that something is the case (Pitt 2004). Some of them have also carried over this general view into the specific case of emotion. Thus, according to some thinkers, the phenomenology of a sadness episode includes not just feelings of bodily changes but also, among other things, an experience of loss (Goldie 2002, Kriegel 2011).

The claims that there exists in general a cognitive phenomenology and that the phenomenology of an emotional episode includes a cognitive part are controversial and I myself would like to suspend my judgment about them. In any case, I expect Prinz to reject them, specially the second one, which is overtly incompatible with his view.

Even so, and that is the main point I want to make about this first reason, appraisal theories, as they are commonly conceived, have I think little to fear from it. For even if we were to endorse, by accepting James' subtraction argument, that the phenomenology of emotion is exhausted by feelings of bodily changes, we could still hold that this is the phenomenological content of the feeling component of an emotional episode, the third stage of the emotional process devised by appraisal theorists which I discussed in the second section. There is nothing, I think, that prevents appraisal theorists from claiming that an emotional episode becomes phenomenally conscious only when it reaches its third stage, the "feeling component", and that when it does so what we feel are just these bodily changes.

The point is that Prinz seems to be arguing here abductively, using an inference to the best explanation. His argument seems to be this: first, James is right, the phenomenology of emotion is exhausted by sensations of bodily changes; second, the only explanation of this is that bodily changes come first and emotions are caused by them. But the second premise can be challenged. An explanation of James' finding might also be that an emotional episode is a process which starts with an appraisal then it is followed by a syndrome of bodily changes and finally reaches a state of phenomenal consciousness of these changes.

Let's then move on to the second Jamesian reason. Prinz backs Robert Zajonc in his dispute with Richard Lazarus and endorses the claim that emotion and cognition involve two distinct neuroanatomical structures. Of course, this claim is more or less plausible depending on how the elusive notion of cognition is spelled out. Prinz has his own way of doing this which I will not discuss here, but the intended view is that basic cases of emotion like for instance those unveiled by LeDoux's research with fear responses to snake-like objects do not involve cognitive states at all. Prinz, as Zajonc, draws from this the conclusion that they do not involve appraisals of any kind.

This is of course something that appraisal theorists simply deny, and expectedly so since they of course attribute to other animals and new-born infants the capacity to emote. And furthermore they think that the human emotive system has an evolutionary origin,

it was naturally selected as a way of discerning in the environment matters of importance for the organism and of helping it to deal with them in appropriate ways. This being the case, the existence of pre-wired emotional responses, or predispositions to respond emotively to certain stimuli allegedly prominent in our evolutionary past, is to be expected.

Consequently, most models in the appraisal approach have it that some appraisal dimensions, especially those processed first, are frequently run in an automated fashion and in many cases do not involve any cognitive states in the sense of 'cognitive' which bothers Zajonc or Prinz. In fact, some theorists argue that the last appraisal dimensions to be processed in an emotional episode, which would involve an estimation of the bearing of the stimulus or its consequences with respect to social norms or the subject's ego ideal, are in principle only present in humans (and to a limited extent perhaps also in some other primates) and require considerable cognitive effort.

The general idea is that the human emotive system is built upon an evolutionary basis and the cognitive sophistication of the human brain allows it to be more subtle in the sort of appraisals performed and more flexible in the sort of responses given as output. This of course only adds adaptability and efficacy to the whole system. The main idea of appraisal theories is that the emotional response is determined by the subjective evaluation of the stimulus along several appraisal dimensions. This may be hard-wired for the most part, that is to say, the causal connection between a certain result of a certain appraisal dimension and its efferent effects may be hard-wired. This would explain why, for instance, an episode of fear elicited by a snake and an episode of fear elicited by a job interview may largely involve the same efferent responses. But, of course, the processing of each appraisal dimension, the sort of process leading to an output result for a particular dimension, may be more or less cognitively mediated, depending on the cases and the dimensions.

Of course, the interplay of pre-wired and learnt elements in the delivery of the emotional response is at present still little understood. Some theorists claim that more cognitively demanding processes are only called into action when the simpler automated processes cannot solve by themselves the problem posed by the stimulus (Scherer 2001). I'm not persuaded by this view since in many cases a stimulus will typically be felt as posing a "problem" only when it is appraised in certain ways with the help of higher cognitive states. For instance, as far as automated processes are concerned, there is nothing wrong with a gun as opposed, say, to a snake. No problem then to solve and no need to worry about guns as far as these pre-wired mechanisms are concerned. It seems clear that things do not work this way.

Most appraisal models include a dimension of "urgency" which would estimate the need of a fast response to the stimulus. It might then be that when an innate theme for a given emotion is perceived, say a snake in the case of fear, this appraisal dimension delivers the highest degree of output as a result of a fully automated process and that our emotive system is pre-wired in such a way that when the urgency dimension delivers an

output response at its highest degree, or above a given threshold, it causes an appropriate behavior without waiting for the intervention of more sophisticated cognitive processes. This is only a bald conjecture on my part with no empirical support as far as I know, but it is only intended to show that there is room within the framework of appraisal theories to accommodate such as empirical results as LeDoux's.

It remains only to be considered the last reason afforded by Prinz for thinking that bodily changes precede emotion. This is the alleged existence of cases of direct physical induction of emotion. Administering certain drugs or even adopting certain characteristic facial expressions is told to provoke an emotional response. Of course this is on the face of it a fact which seems more easily explained by Prinz's theory than by the appraisal theory. Indeed a plausible explanation of what goes on in those cases is that drugs may cause the bodily changes the registration of which constitutes the emotion according to Prinz. It seems on the other hand harder to defend that drugs may affect the sort of appraisals which according to appraisal theorists give rise to emotions.

It is interesting to note that most appraisal theorists claim that this sort of cases is not covered by the theory. This may be one of the reasons why psychologists tend to speak of appraisals as being the usual cause of an emotional episode rather than its first and triggering component. The idea is that they are the usual way in which emotional episodes arise, but there are also other ways and direct physical induction would be one of them.[6]

I think on the contrary that this reaction may be too quick and that it actually underestimates the resources available to appraisal theories. I will conclude then suggesting some ways in which cases of physical induction can be accounted for by them. As was the case with the previous objection, my aim here is not primarily to argue for any of these ways but rather to show that these cases need not be seen as posing an insurmountable problem to the appraisal approach.

One consideration is this. We have seen that feelings of bodily changes constitute the phenomenology of an emotional episode at least to a great extent. Moreover, it is reasonable to assume that the part of an emotional episode which is amenable to verbal report and conscious recognition is precisely that part which is phenomenally conscious. There can be little wonder then if in cases of physical induction a subject reports feeling something very similar to standard cases of emotional episodes, if not plainly the same. This being said, it is quite a different matter if what is felt in these cases is the genuine thing. Some elements of the emotional episode are probably present –a syndrome of bodily changes, a conscious representation of them—but other crucial ones –the appraisal process—are missing. This is perhaps why drinking alcohol in order to change sadness feelings into joy feelings is not in the long run satisfying at all.

A second consideration allows us to go a bit further. Little is yet known about mental architecture as implemented by the human brain. But some of the best known processes

[6] Another case would be listening to instrumental music (Scherer 2001). Some philosophers who favor the appraisal approach to emotions also hold this "pessimistic" view, see for instance Scarantino (2010).

show that the brain uses recurrent networks. In recurrent networks information does not travel in one direction only, always forward till the last stage of processing is reached. Instead the output of a certain stage of the process feeds back and influences the output of a previous stage which is processed again and again until the general process stops (Damasio 2010, chapter 3). Some appraisal theorists favor this sort of architecture for appraisal processes involved in emotion and even for the whole emotional episode. According to this, appraisal dimensions which are processed earlier receive input from the results obtained by the processing of later dimensions. The stimulus is appraised and reappraised along the different dimensions and the results obtained in one dimension are influenced by those obtained in the others until the process stops, and this probably occurs when results remain stable and unaltered for some period of time or when the urgency dimension recommends action. Likewise, it is thought that the connections linking the processing of appraisal dimensions with efferent effects are also two-way, with signals travelling forward and backward (Scherer et al. 2001). If this is indeed so then it can be sustained that certain efferent effects brought about by direct physical induction may affect the appraisal process inherent to emotion in such a way that crucial brain sites for emotion may reach a pattern of activation closely resembling that of a standard emotional episode.

REFERENCES.

Arnold, M. (1960): *Emotion and Personality*. Columbia University Press.

Damasio, A. (2010): *Self Comes to Mind. Constructing the Conscious Brain*. William Heinemann.

Ekman, P. (2003): *Emotions Revealed. Recognizing Faces and Feelings to Improve Communication and Emotional Life*. Henry Holt and Company.

Ellsworth, P. & Scherer, K.R. (2003): "Appraisal Processes in Emotion", in Davidson, R., Scherer, K., & Goldsmith H. (eds.): *Handbook of Affective Sciences*. Oxford U.P., pp. 572-595.

Goldie, P. (2002). "Emotions, Feelings and Intentionality", *Phenomenology and the Cognitive Sciences* 1: 235-254.

James, W. (1884): "What is an Emotion?", *Mind* 9: 188-205.

Kriegel, U. (2011): "Towards a New Feeling Theory of Emotion", *European Journal of Philosophy* (forthcoming).

Lazarus, R. (1991): *Emotion and Adaptation*. Oxford U.P.

LeDoux, J. (1996): *The Emotional Brain*. Simon and Schuster.

Nussbaum. M. (2001): *Upheavals of Thought. The Intelligence of the Emotions*. Cambridge U.P.

Pitt, D. (2004): "The Phenomenology of Cognition; or What is it Like to Think that P", *Philosophy and Phenomenological Research* 69: 1-36.

Prinz, J. (2002): *Furnishing the Mind: Concepts and their Perceptual Basis*. MIT Press.

Prinz, J. (2004): *Gut Reactions. A Perceptual Theory of Emotion*. Oxford U.P.

Prinz, J. (2007): *The Emotional Construction of Morals*. Oxford U.P.

Prinz, J. (2008): "Précis of Gut Reactions", *Philosophy and Phenomenological Research* 76/3: 707-711.

Scarantino, A. (2010): "Insights and Blindspots of the Cognitivist Theory of Emotions", *British Journal for the Philosophy of Science* 61: 729-768.

Searle, J. (1983): *Intentionality*. Cambridge U.P.

Scheier, M.F. & Carver, C.S. (1985): "Optimism, Coping, and Health: Assessment and Implications of Generalized Outcome Expectancies", *Health Psychology* 4. 219-247.

Scherer, K.M. (2001): "Appraisal Considered as a Process of Multilevel Sequential Checking", in Scherer et al. (2001), pp. 92-120.

Scherer, K. R. (2004): "Feelings Integrate the Central Representation of Appraisal-Driven Response Organization in Emotion", in Manstead, A., Frijda, N., & Fischer, A. (eds.): *Feelings and Emotions: The Amsterdam Symposium*. Cambridge U.P., pp. 136-157.

Scherer, K.R. (2009): "The Dynamic Architecture of Emotion: Evidence for the Component Process Model", *Cognition and Emotion* 23/7: 1307-1351.

Scherer. K. R., Schorr, A. & Johnstone, T. (eds.) (2001): *Appraisal Processes in Emotion: Theory, Methods, Research*. Oxford U.P.

Solomon, R.C. (1976): *The Passions*. Doubleday.

Solomon, R.C. (2003): *Not Passion's Slave*. Oxford U.P.

Van Reekum, C.M. & Scherer, K.M. (1997): "Levels of Processing in Emotion-Antecedent Appraisal", in Matthews, G. (ed.): *Cognitive Science Perspectives on Personality and Emotion*. Elsevier Science, pp. 259-300.

Zajonc, R., Murphy, S., Inglehart, M. (1989): "Feeling and Facial Efference: Implications of the Vascular Theory of Emotion", *Psychological Review 96*: 395-416.

NATURALISM, PLURALISM AND LOCATION PROBLEMS. REFLECTIONS ON JESSE PRINZ'S NATURALISTIC COMMITMENTS

Stefano Cossara (Paris-Sorbonne University)

Jesse Prinz's work encompasses an impressive variety of themes: from aesthetics to concept theory, from metaethics to the psychology of emotion. However, this admirable variety of topics does not prevent from identifying at least two overarching commitments: one to Hume's empiricism, which has sometimes led Prinz to read his own work (for sure too modestly) as a series of footnotes to Hume's; the latter to a throughout naturalism, whose implications are ontological as well as epistemological. This paper focuses on the latter aspect.

Prinz makes his own naturalistic commitments explicit in the preamble to *The Emotional construction of Morals* (2007), where he endorses four varieties of naturalism. The first is *metaphysical naturalism*, which Prinz reads as a denial of supernaturalism:

> Our world is limited by the postulates and laws of the natural sciences. Nothing can exist that violates these laws, and all entities that exist must, in some sense, be composed of the entities that our best scientific theories require (Prinz 2007, p. 2)

Existence can be granted only to entities that are required by our best scientific theories. Spirits and fairies are not included in those theories, hence they cannot exist. Metaphysical naturalism is connected to the so called location problems, concerning the attempt to find a place in the world for those facts and entities that do not seem to be included in our best scientific theories. In *The Emotional Construction of Morals* Prinz faces the problem of locating moral facts, so as to avoid that they must be considered non-existent like fairies and ghosts.

Metaphysical naturalism is by no means the only variety of naturalism that Prinz endorses. He actually takes metaphysical naturalism to entail a sort of *explanatory naturalism*: all that exists and is not described in the language of science, must in the end be describable in those terms. Prinz hastens to add that his position does not amount to reductionism: one need not think that lower-level explanations are the only genuine explanations, and that higher-level explanations must be *deduced* from the former. However, higher levels must be tied to the lower levels by some kind of systematic correspondence.

Prinz also endorses a kind of *methodological naturalism*, which he takes to come from Quine: if all the facts are in a sense natural facts, those facts must ne investigable by methods suitable to the investigations of natural facts. Prinz also subscribes to a fourth and less popular variety of naturalism, which he again takes to derive from Quine: *transformation naturalism*, according to which we always operate from within our theories of the world; we cannot step outside and adopt a transcendental position, for we cannot think of the world independently of our theories.

In *The Emotional Construction of Morals*, Jesse is very clear in pledging his alliance with naturalism, but he does not argue for the theses he endorses. In this paper I maintain that those theses do actually require an explicit and systematic defence. In section 1 I suggest that naturalism cannot be taken for granted, because of the strength of the theses it entails and of their being significantly controversial. In section 2 I try to cast light on what seems to me a tension within Prinz's naturalism, one that is related his methodological pluralism. In section 3 I briefly sketch an alternative approach to location problems, one that is still naturalistic but that avoids some of the problems of classical naturalism.

1. Why should we be naturalists?

Prinz can hardly be criticised for not providing an explicit defence of his naturalistic commitments, for most contemporary partisans of naturalism seem to take it for granted. However, if naturalism is to be more than a self-justifying dogma, or an intellectual fashion, a defence seems necessary, especially considering that the theses he endorses are strong and by no means uncontroversial. As an ontological thesis, Prinz's naturalism boils down to the idea that no part of the existent can lie beyond the world described by natural sciences: all facts are in a sense natural facts. As an epistemological[1] thesis, it states that all genuine knowledge is scientific knowledge, or can be traced back – in one way or another – to scientific knowledge. As a consequence, there is no genuine knowledge out of science; a conclusion that would certainly strike most people as quite strong.

Not only naturalism leads to strong conclusions; those conclusions, however popular in some philosophical circles, are by no means universally accepted. Paul Horwich, for example, does not confine himself to raising doubts about the tenability of naturalism; he seems to suggest that naturalism is entirely unwarranted, and evidently so. On Horwich's view, all we need to get rid of the idea that it is necessary to 'locate' apparently non-naturalistic facts in a natural world is a *superficial* explanation of naturalism's initial appeal:

> a) Naturalism rests on the impression that any non-natural facts would be intolerably weird.
> b) That impression stems from a combination of three factors: first, the singular practical and explanatory importance of naturalistic facts; second, the very broad scope of the naturalistic – the striking range and diversity of the facts that it demonstrably encompasses; and third, the feeling that reality must 'surely' be fundamentally uniform – so all facts must be naturalistic.
> c) This final feeling is based upon a misguided overextension of scientific norms: in particular, the norm of theoretical simplicity. For it is pretty clear (i) that the metaphysical and epistemological variety of possible facts corresponds exactly to the variety of possible meanings (i.e. of possible regularities of word-use); (ii) that the latter will certainly include many that are *non*-naturalistic; and (iii) that many of those will be socially useful and will therefore be deployed.
>
> In order to undercut the sense of 'weirdness' that can stem from our failure to naturalistically 'locate' a given phenomenon it suffices to acknowledge the evident plausibility of this diagnosis. (Horwich 2010, p. 157)

If Horwich's diagnosis is correct, the naturalist is wrong in his ontological and, by extension, in his epistemological claims. Maybe to say that the plausibility of the diagnosis is evident is too optimistic; perhaps, Horwich might be criticised for begging the question on the central point of the issue: the naturalistic assumption that reality is sufficiently uniform as to contain naturalistic facts only. It might be said that, while not evidently true, this thesis is not evidently false either: it is not so clear, in other words, that the many philosophers (Prinz included) who make this assumption are entirely wrong. Thus, reading a conclusive refutation of naturalism in Horwich's text might be too optimistic for the anti-naturalist. However, it seems to me that this argument still does some important work to the effect of showing what I am arguing for: that to the extent that it rests on premises that are by no means uncontroversial, naturalism requires a defence. More precisely, I want to leave the following question open for Jesse: why should we think that a substantive uniformity of reality – which, as Horwich has it, is a necessary precondition for metaphysical naturalism to be tenable – actually obtains? To think so seems to be at odds with evidence from the senses.

The credibility of naturalism is sometimes upheld by presenting it as an alternative to supernaturalism, where supernaturalism is read as entailing the existence of entities we uncontroversial know to not exist. This informal argument runs more or less like this:

A) Thinking that naturalism is false is tantamount to admitting that fairies and ghosts exist
B) Fairies and ghosts do not exist

[1] It seems to me that Prinz's explanatory and methodological naturalism can be combined in order to yield this strong epistemological conclusion.

hence

C) Naturalism is correct

It should be evident, however, that this argument only becomes interesting if combined to a conflation of supernaturalism (the thesis that supernatural entities such as fairies and spirits exist) and non-naturalism (the thesis that non-natural properties such as moral properties exist without supervening on natural properties). This conflation, however, is unwarranted: one may well think that the latter exist, without necessarily granting existence to the former. The non-naturalist's justification for a similar attitude might be that while science can be granted authority over facts concerning ghosts and fairies (entities that, if existent, would be 'out there' just like cats and birds), it is not up to the task of adjudicating on the existence of moral facts, which concerns properties of a totally different kind. Apparently, this boils down to just a legitimate delimitation of the scope and authority of science. Maybe a similar delimitation is in the end incorrect, but again, the burden of the proof seems to rest on the naturalist to show that her position is the good one.

2. A tension in Prinz's naturalism

In the previous section I have confined myself to showing that naturalism cannot be taken for granted. I will now raise a problem for the variety of naturalism endorsed by Jesse. As should by now be clear, on Prinz's accounts natural facts mark the borders of what really exist. But what are natural facts? Talk about nature risks being empty in the absence of further specification. Of course, Prinz provides this specification by appeal to a the following principle: natural facts are those that can be investigated using the methods of natural sciences, and can be expressed by statements employing the vocabulary of natural science only. What unifies the universe of natural facts is their privileged relationship with the methods and the vocabularies of the natural sciences. But is this principle strong enough to really grant unification? This is a serious question. For if it were shown that the set of the so called natural facts (which for Prinz are the only genuine facts) were indeed internally heterogeneous, it might be suspected that its borders are established in a purely arbitrary fashion. If there is no substantive uniformity within the set of the facts thus identified as natural, why should not the set be open to, e.g., moral facts?

Of course, in order to establish whether the criterion employed by Prinz to indentify natural facts yields or not a substantially uniform set, it is necessary first of all to establish what counts as natural science. On the one hand, it is possible to identify the natural sciences with the hard sciences. In this case, one will probably find in the methodological and explanatory procedures of those sciences a sufficient degree of uniformity as to grant the conclusion that the facts described by those sciences comprise a substantially uniform set. At any rate, one will be able to find a similar degree of uniformity within physics, that is often thought of by empirically-minded metaphysicians as the best source of insights concerning the fundamental nature of reality. But Prinz's methodological naturalism encompasses a far broader array of disciplines, and extends to sciences such as history, that are often thought of as human or social rather than natural sciences. The problem is not whether cultural history deserves or not the label of a natural science, for that is simply a linguistic matter. The crucial point is whether a substantially uniform set of 'natural' facts can include at the same time the facts described by sciences that are so different in the methods and the vocabularies they employ: physics on one side, history and anthropology on the other. In order to be allowed to draw metaphysical conclusions from his sources of empirical evidence, Prinz needs to answer affirmatively. But then the question is: if the universe of natural facts is so heterogeneous as to include at the same time the facts of physics and those of cultural history, why could not that set also include facts about morals, causation, or modality? Prinz's pluralism helps us to recognize that there is no uniformity in reality; but acknowledging that seems to undercut the vary rationale behind talk of location problems and the naturalistic commitments that underpin them.

3. Maximizing pluralism: a subject naturalistic perspective on location problems

It seems to me that pluralism is the key to a more satisfactory treatment of location problems. However, it is necessary to distinguish two kinds of pluralism. One variety of pluralism is *horizontal pluralism*, which concerns a plurality of ways of doing the same thing – of performing, as it were, the same linguistic task. According to Quine's principle of ontological relativity, for example, there exists a plurality of alternative scientific worldviews, "each empirically adequate to more or less the same degree, and none, even in principle, having privileged claim to provide a truer description of the world" (Price 1992, p. 389). As a moral relativist, Prinz is a horizontal pluralist: he admits the existence of a range of equally coherent moral viewpoints, none objectively superior to any other. However, the important move towards a novel treatment of location problems consists in adopting a further variety of pluralism, which I will call *discourse pluralism.* Discourse pluralism consists in recognizing that philosophy deals with an irreducible plurality of kinds of discourses, of games of language: for example, the moral as well as the scientific. When it comes to morals, the discourse pluralist will agree with Prinz that moral facts cannot be reduced to non-moral facts; however, she will resist his suggestion that in order to take them to exist, one needs to ground them on non-moral facts. She will reject the very idea that different domains of discourse need to be unified, and that there is one single universe of facts that exhausts the scope of reality. I do not have the time to spell out this alternative view in the details here. I hasten to say, however, that it need not lead to any bizarre form of naturalism. To the contrary, the approach it yields is fully naturalistic, even though the variety of naturalism it exemplifies is different from the one endorsed by Prinz. While Prinz, as most contemporary naturalists, is interested in the objects and properties that can be deemed really existent, I am more concerned with the different functions and roles that language can play in the life of natural creatures like us human beings. The concern is, as it were, with the subject rather than the object. As a consequence, this different approach has been labelled *subject naturalistic* (Price 2004).

REFERENCES

Horwich, P., 2010. "Rorty's Wittgenstein", in *Wittgenstein's Philosophical Investigations : A Critical Guide*, Arif Ahmed, ed., Oxford: Oxford University Press, pp. 145-161.

Price, H., 1992. "Metaphysical Pluralism", *Journal of Philosophy* 89 (8), pp. 387-409.

Price, H., 2004. "Naturalism without representationalism", in *Naturalism in Question,* David Macarthur and Mario de Caro, eds., Harvard University Press, 2004, pp. 71–88.

Prinz, J., 2007. *The emotional Construction of Morals,* Oxford : Oxford University Press.

(2454 words, notes excluded)

Pushing the accelerator on enactive perception

How sensorimotor dynamics can constitute minds

Xabier E. Barandiaran

IAS-Research Centre for Life, Mind and, Society
Dept. of Logic and Philosophy of Science
UPV/EHU, University of the Basque Country

http://xabier.barandiaran.net
xabier.academic@barandiaran.net

Enactivism (Noë, 2005; Stewart, Gapenne, & Paolo, 2011; Thompson, 2007; Varela, Thompson, & Rosch, 1991) is far from the fashionable-new-hype following "Noë's siren-call" that Prinz (2006) makes us believe. It follows the tradition of those that overcame the empiricist school that Prinz (2002) so enthusiastically vindicates: pragmatists. John Dewey summarized one of the central claims of enactivism over 100 years ago:

> Upon analysis, we find that we begin not with a sensory stimulus, but with a *sensorimotor coordination* (...) and that in a certain sense it is the movement which is primary, and the sensation which is secondary, the movement of the body, head and eye muscles determining the quality of what is experienced. (…) the real beginning is with the act of seeing; it is looking, and not a sensation of light. (Dewey, 1896, pp. 358–359, italics added).

Ever since, psychologists, neuroscientists and philosophers alike have tried to deepen into the sensorimotor nature of mind and experience. To some extent these attempts were reduced (by the epistemological demands of behaviourism) to a statistical notion of stimulus-response correlations, to be latter substituted by computational representationalism. Conceptual, mathematical and experimental constraints (that we have just started to unlock) were partly responsible for the limited scientific development of the early insights on the sensorimotor nature of experience made by phenomenology (Heidegger, 1991; Merleau-Ponty, 1942, 1944) and pragmatism (Dewey, 1896, 1925). Things have changed recently, but not enough. To put it in terms of neurodynamic researcher Walter Freeman:

> What allows us a fresh start now is our ability to image brain activity during normal behavior and to model our findings with the tools of nonlinear dynamics. However, these new data are being acquired under preconceptions embodied in old experimental designs, and we have to reinterpret them as they bring new concepts to light. (Freeman, 2001, p. 12)

These "preconceptions embodied in old experimental design" are still alive. Jesse Prinz (2000, 2002, 2006) has become one of the youngest and strongest supporters of some of them (with certain contributions of his own), "joining the front-lines" to defend the

boundaries between perception and action against enactive and sensorimotor approaches to cognitive science. In "Putting the Brakes on Enactive Perception" (2006) Prinz makes a fierce attack on Noë's "Action in Perception" (2005), questioning dozens of Noë's arguments.

There is no room here to reply to each of the Prinz's critiques to Noë's book. I shall instead concentrate on three main subjects. First, I address some empirical neuroscientific issues that are central to Prinz's resistance to the enactive view and his neglect of motor function for perceptual awareness. The second aspect I will discuss combines both conceptual and neurodynamic aspects. I will propose a simulation model that illustrates a notion of dynamic coordination that is richer than the kind of causal-metaphysical assumptions underlying Prinz's work; both at the level of agent-environment interaction and at the neurodynamic level. Finally I shall identify the real challenge that some of the strongest position in enactivism have to face: the relationship between virtuality and sensorimotor coupling.

I

The first of Prinz's central claims I want to discuss is of an empirical nature: "[N]euroscience provides an overwhelming case for the view that perception is not essentially linked to action" (Prinz, 2006, p.11). Contrary to Prinz's clame, I will summarize some neuroscientific evidence showing that perception *is* "essentially" (more on this term latter) linked to sensorimotor dynamics at developmental, anatomical, and functional scales.

One of the most cited supporting evidence for enactive development comes from Held and Hein's experiments. Two kittens where reared by holding one immobile and attached to the other, so that both received the same sensory stimulation, yet only one had freedom to control movement. After a period of rearing kittens where tested in different perceptual tasks, where behavioural consequences should be able to asses whether the kitten was capable of correct visual discrimination. In one of these experiments the immobile kitten was put in front of a cliff (protected by a transparent glass on the floor) and walked through without noticing. Prinz dismisses Held and Hein's experiments (Held & Hein, 1963) by interpreting that the immobile kitten just "did not have enough experience walking on edges". To be fair, neither did the freely moving kitten (during rearing it did not confront walking on edges), yet it showed no incapacity to perceive the cliff and avoid it. However, I will concede to Prinz that some of these experiments might not be able to disambiguate with sufficient accuracy between perceptual dysfunction and perception-action coordination problems. Unfortunately for Prinz, and his categorical assertion that "the Held and Hein study was never replicated", latter studies have supported the perceptual dysfunction interpretation with further evidence coming from lesion studies on ocular muscles on kitten, identifying selective neuronal blindness for visual features orthogonal to the movements made by the occluded muscles (Buisseret, Gary-Bobo, & Imbert, 1978; Buisseret, Gary-Bobo, & Milleret, 1988).

Prinz wants to strengthen his claim against the developmental role of action for perception by claiming that "studies of human infants with muscle atrophy show that when humans are prevented from moving in early development, there is no decrement in the

visual comprehension of space" (Prinz, 2006, p. 10). And yet, examples of spinal muscular atrophy do not invalidate sensorimotor accounts of perceptual development simply because head and saccadic eye movement are perfectly intact in those cases. The enactivist claim is not that all forms of motor function need be intact in order to develop "normal" perception. What is required is that the developing organisms have access to the way in which perspectival changes and movements affects sensory stimulation. However, developmental facts are not decisive. No matter how development occurs, current perceptual experience might not necessarily depend on motor activity. Simply put, physiological conditions for correct development are often different from the conditions necessary to correctly carry out physiological functions.

Perhaps the strongest of Prinz's claims regarding the actual lack of evidence for action in perception is the following:

> If the brain areas that are known (because of their behavioral consequences) to encode the motor consequences of visual stimuli are not implicated in visual consciousness, then there is no reason to think Noë's theory of consciousness is correct. Noë is forced to say that representations in the ventral visual stream are also involved in the coordination of action, but there is absolutely no evidence for this conjecture. All evidence implicates the dorsal stream. (Prinz, 2006, p.10)

Prinz is here referring to the "two stream theory of vision" (Goodale & Milner, 1992) which states that there are two distinct visual pathways: the dorsal stream (also referred as "vision for action") and and the ventral stream (or "vision for perception"). First, it is important to remark Noë's insistence on the fact that "the enactive approach is not committed to the idea that vision is for the guidance of action, so neither the fact that some visual processing is for the guidance of action, nor the fact that some visual processing is not, has any direct bearing on the enactive approach" (Noë, 2005, p.19). And yet, there is evidence for action in perception along the ventral stream (vision for perception). In a recent review of the two stream theory (Milner & Goodale, 2008) the authors of the theory remind us that "there is complementary evidence that supports a ventral-stream role in the planning of action" (p.776) and that "in most normal circumstances, our actions will be visually co-determined by complementary processing in both dorsal and ventral streams" (p.776). More importantly, they also take for experimentally confirmed that the ventral stream is used to coordinate sensorimotor tasks when the movements are awkward or not automatized. Visual illusions, that are processed only by the areas involved in the perceptual stream, have consequences for reaching and grasping when subjects are asked to do so with the left hand or in non-automatized situations (Gonzalez, Ganel, Whitwell, Morrissey, & Goodale, 2008). Milner and Goodale conclude that "only highly practiced actions with the right hand operating in real time and directed at visible targets presented in the context of high-level illusions are likely to escape the intrusion of ventral-stream perceptual control" (Milner & Goodale, 2008, p. 780). Thus it turns out that the vision-for-perception stream is actually involved in precisely those aspects of movement planning and execution that require conscious control. The empirical facts are far from Prinz's bold claim that there is "absolutely no evidence" for ventral stream involved in the coordination of action.

Part of Prinz's dificulty to include a role for action in perception is that he conceives a rather one-directional "object → eye → V1 → … → V4*" causation sequence, with some kind of "elusive marking" at * coming from attentional processes, that makes neural activity conscious, but ignoring any possible role of motor and pre-motor activity (Prinz, 2000). However, neurological evidence suggests early involvement of thalamo-cortical loops (LGN projecting directly to V1) on the emergence of perceptual experience, bringing together sensory and (pre-)motor dynamics into the constitution of neurodynamic correlates of perceptual awareness. The modulatory effect of LGN activity on visual perception is nowadays widely proven (Briggs & Usrey, 2011; Kastner, Schneider, & Wunderlich, 2006; Royal, Sáry, Schall, & Casagrande, 2005). Moreover, effects of saccadic eye movements on LGN alter "not only response strength but also the temporal and chromatic properties of the receptive field " (Reppas, Usrey, & Reid, 2002, p. 961) and motor planing has being shown to influence LGN activity (Royal et al., 2005). It is therefore untenable to claim that "V1 is a primary source of inputs to another region in which consciousness can rightfully be said to reside" (Prinz, 2000, p. 246) without even considering LGN as a proxy for motor influences on V1 and, consequently, on visual awareness.

II

The second aspect of Prinz's position and resistance to enactivism has to do with an impoverished conception of neurodynamic organization and agent-environment dynamics. "Every aspect of experience, from illusory contours to motion illusions, from phantom limbs to diffuse pains, can be correlated with some neuronal *response*." (Prinz, 2006, p.17, italics added). The term "response" is crucial at this point. Prinz offers no analysis of this term and it is reasonable to assume that he conceives this "response" as some kind of local state or activity. There is no consideration of large scale transient synchronization or any other kind of mesoscopic dynamic structure of brain activity and its sensorimotor coordination with the environment. It seems like the underlying conception of causation is a linear, sequential and atomic one. Prinz's neglect of the complex neurodynamics of brain and sensorimotor functioning could be further illustrated with the following statement: "functional organization is mirrored by the organization of the nervous system; functional components are anatomically distinct" (Prinz, 2000, p, 256). A linear one-to-one mapping between anatomical and functional structures, seems to be uncritically assumed. On what follows I will first consider interactive aspects of the dynamic constitution of experience and then move to internal (or strictly neuronal) aspects; showing how Prinz's underlaying metaphysics of causation falls short to make justice to the complexity of the interactive and neurodynamic processes that underlie experience.

Here is where Prinz's criticism to enactivism connects with a wider philosophical debate around extended cognition[1]. It has being argued that, despite the abundance of examples, proponents of sensorimotor coupling as constituting/causing cognition make very mild claims about the exact kind of coupling involved (Aizawa, 2010). In order to contribute to the ongoing debate around the causal vs constitutive role of sensorimotor dynamics for cognition (Adams & Aizawa, 2009; Aizawa, 2007; Block, 2005; Clark,

1 Note that conditions for perceptual awareness are stronger than those that might be imposed for extended cognition. Perceptual awareness does not supervene on all forms of distributed cognition.

2006; Lenay & Steiner, 2010), I will introduce a conceptual synthetic model of sensorimotor coordination: the *situated HKB* model (Aguilera, Bedia, Santos, & Barandiaran, in preparation; Santos, Barandiaran, Husbands, Aguilera, & Bedia, submitted). The model shows a minimal agent capable of performing phototaxis in a 2D environment by means of internal metastable regimes of the HKB equation's single variable φ. The crucial experiment is one in which the input of a freely-behaving agent is recorded and then played back into an identical but immobile agent. The "brain" dynamics of the immobile agent are qualitatively different from that of its freely-behaving twin, even if the structure of the sensory input is identical. The model illustrates and makes a proof of concept for the case that *neuronal metastable transients that are functional at the behavioural/cognitive scale might emerge from fine grained micro-dynamic sensorimotor compensations and coordinations.* There is no "state" or "response" of φ to a sensory perturbation that can be said (itself) to correlate with any particular functional contribution to phototaxis. It is through sensorimotor coupling that transient dynamics become functionally relevant. Prinz's critique to enactivisms rests on a narrow conception of sensorimotor dynamics as is apparent when he responds to Noë's account of visual stabilization by stating that:

> If this were true [that perception depends on sensimotor contact with the environment], it would show only that the world is a causal precondition for having some phenomenal experiences. It would *show only that the brain is incapable of entering certain configurations without external stimulation.* (Prinz, 2006, p. 16, italics added)

The situated HKB model shows a clear illustration of how sensorimotor coupling can go beyond the "entering a certain configuration without external stimulation". Again, stimulation is not a cause of brain dynamics. It is the environment and the sensorimotor embodiment that might become essential for the kind of sensorimotor↔brain coordination that characterize the neurodynamic patterns that correlate with perceptual awareness.

If we further consider that a) brains are in a continuous state of metastability in highly interconnected holistic dynamic cores comprising sensorimotor, emotional and higher-order centers (Chialvo, 2004; Rabinovich, Huerta, Varona, & Afraimovich, 2008; Tognoli & Kelso, 2009; Werner, 2007) and b) that transient coordinations correlate with perceptual awareness (Freeman, 2001; Llinas, 2001; Tononi, Sporns, & Edelman, 1994; Varela, Lachaux, Rodriguez, & Martinerie, 2001) then, it is reasonable to assume that neurodynamic coordination is the characteristic form of constitution of experience. Varela (1995) made the following calculation: at a spike travelling speed of 10m/s a spike wave would take about 40ms to make a return trip between both hemispheres (25cm travel). One such cycle will thus involved a frequency of 1000/40 = 25Hz. The gamma band (25-40Hz) is just above the minimum frequency required to synchronize the activity of the full brain (or, at least, the cortex). We need to add that the formation of a visual percept takes up to 100-200ms thus allowing for 3-5 cycles of whole brain reciprocal influence or coordination to take place. It is no coincidence that conscious experience and attentional phenomena (an essential part of Prinz's AIR theory of consciousness) have been systematically related to the gamma band activity (Crick & Koch, 1990; Jensen, Kaiser, & Lachaux, 2007).

We can now go back to Prinz's statement that "every aspect of experience can be correlated with some neuronal *response*" (and his defence of a direct mapping between anatomical → functional/representational → phenomenological units) to see how it falls short to capture the kind of interactive neurodynamics at stake.

III

Prinz states that "to support wide supervenience, Noë should show that, when we keep the brain fixed and change the environment, there can be changes in experience. He attempts no argument of this kind" (Prinz, 2006, p. 16) By now it should be evident that it simply makes no sense "to keep the brain fixed". That would amount to mental death. So does long term sensorimotor deprivation (Ebert & Dyck, 2004; Grassian & Friedman, 1986). Perhaps a more reasonable formulation of Prinz's concern is whether perceptual experience *always, necessarily and systematically* depends on *direct* sensorimotor coordination with the environment. At this point Prinz is right to suggest we should push the break on enactivism or at least slow it down. It is here where enactivism might have to re-negotiate the most radical forms of externalism and wide supervenience. But first it is important to remind ourselves that Noë's enactivism includes the notion of virtuality (and not only direct sensorimotor exercise with environmentally accessible features) as constitutive of perception:

> As a matter of phenomenology, the detail is present not as *represented*, but as *accessible*. Experience has content as a potentiality. In this sense, the detail is present perceptually in my experience virtually. Thanks to my possession of sensorimotor and cognitive skills, I have access to nearby detail. (…) [V]irtual presence is a kind of presence, not a kind of non-presence or illusory presence. (…) Qualities are available in experience as possibilities, as potentialities, but not as completed givens. Experience is a dynamic process of navigating the pathways of these possibilities. Experience depends on the skills needed to make one's way. (Noë, 2005, pp. 215–217)

Unfortunately, Noë remains mostly silent about the neural basis of this virtuality and its "mediation by knowledge of sensorimotor contingencies". I think this gap can be perfectly filled in in terms of the neurodynamic integration of virtual sensorimotor loops (i.e. those comprising coordination dynamics between premotor areas and sensory afferents). But this demands that we *internalize* virtuality (something that I doubt Noë will be ready to accept). The very notion of "external virtual presence" is a metaphysical and conceptual oxymoron. To claim that perception is a temporarily extended process doesn't save it, for Noë himself acknowledges that "experience is fractal, in this sense": no matter how much you directly engage with a specific feature of the environment, perceptual content will always include virtual aspects that are not directly in view. So, no matter how long you engage in sensorimotor interaction with the environment you will never fill all the gaps. And if knowledge of these present, yet virtual (i.e. not current), sensorimotor contingencies is constitutive of perception, it must have some neural basis. If enactivism wants to push the accelerator again, it needs first to slowly fill the conceptual and empirical gaps that explain how brains integrate virtual sensorimotor knowledge and direct sensorimotor coupling to give rise to perceptual awareness.

References

Adams, F., & Aizawa, K. (2009). Why the mind is still in the head. *The Cambridge handbook of situated cognition*, 78–95.

Aguilera, M., Bedia, M., Santos, B. A., & Barandiaran, X. E. (in preparation). Situated-HKB model: Theoretical Exploration of the Sensorimotor Spatial Coupling of an oscillatory System.

Aizawa, K. (2007). Understanding the Embodiment of Perception. *Journal of Philosophy*, *104*(1), 5–25.

Aizawa, K. (2010). The coupling-constitution fallacy revisited. *Cognitive Systems Research*, *11*(4), 332–342. doi:10.1016/j.cogsys.2010.07.001

Block, N. (2005). Review of Alva Noë, Action in Perception.

Briggs, F., & Usrey, W. M. (2011). Corticogeniculate feedback and visual processing in the primate. *The Journal of Physiology*, *589*(1), 33–40. doi:10.1113/jphysiol.2010.193599

Buisseret, P., Gary-Bobo, E., & Imbert, M. (1978). Ocular motility and recovery of orientational properties of visual cortical neurones in dark-reared kittens. *, Published online: 27 April 1978; | doi:10.1038/272816a0*, *272*(5656), 816–817. doi:10.1038/272816a0

Buisseret, P., Gary-Bobo, E., & Milleret, C. (1988). Development of the kitten visual cortex depends on the relationship between the plane of eye movements and visual inputs. *Experimental Brain Research*, *72*(1), 83–94. doi:10.1007/BF00248503

Chialvo, R. (2004). Critical brain networks. *Physica A: Statistical Mechanics and its Applications*, *340*(4), 756–765.

Clark, A. (2006). Vision as Dance? Three Challenges for Sensorimotor Contingency Theory. *Psyche*, *12*(1), 1–10.

Crick, F., & Koch, C. (1990). Towards a Neurobiological Theory of Consciousness. *Seminars in the Neurosciences*, *2*, 263–275.

Dewey, J. (1896). The reflex arc concept in psychology. *Psychological Review*, *3*(4), 357–370.

Dewey, J. (1925). *Experience and nature*. Courier Dover Publications. Retrieved from http://www.archive.org/details/experienceandnat029343mbp

Ebert, A., & Dyck, M. J. (2004). The experience of mental death: The core feature of complex posttraumatic stress disorder. *Clinical Psychology Review*, *24*(6), 617–635. doi:10.1016/j.cpr.2004.06.002

Freeman, W. J. (2001). *How Brains Make Up Their Minds* (1st ed.). Columbia University Press.

Gonzalez, C. L. R., Ganel, T., Whitwell, R. L., Morrissey, B., & Goodale, M. A. (2008). Practice makes perfect, but only with the right hand: Sensitivity to perceptual illusions with awkward grasps decreases with practice in the right but not the left hand. *Neuropsychologia*, *46*(2), 624–631. doi:10.1016/j.neuropsychologia.2007.09.006

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20–25.

Grassian, S., & Friedman, N. (1986). Effects of sensory deprivation in psychiatric seclusion and solitary confinement. *International Journal of Law and Psychiatry*, *8*(1), 49–65. doi:10.1016/0160-2527(86)90083-X

Heidegger, M. (1991). *Being and Time*. John Wiley & Sons.

Held, R., & Hein, A. (1963). Movement-produced stimulation in the development of visually guided behavior. *Journal of Comparative and Physiological Psychology*, *56*, 872–876.

Jensen, O., Kaiser, J., & Lachaux, J.-P. (2007). Human gamma-frequency oscillations associated with attention and memory. *Trends in Neurosciences*, *30*(7), 317–324. doi:10.1016/j.tins.2007.05.001

Kastner, S., Schneider, K. A., & Wunderlich, K. (2006). Beyond a relay nucleus: neuroimaging views on the human LGN. *Progress in brain research*, *155*, 125–143.

Lenay, C., & Steiner, P. (2010). Beyond the internalism/externalism debate: the constitution of the space of perception. *Consciousness and Cognition*.

Llinas, R. R. (2001). *I of the Vortex: From Neurons to Self*. The MIT Press.

Merleau-Ponty, M. (1942). *The structure of behavior*. Beacon Press.

Merleau-Ponty, M. (1944). *Phenomenology of perception*. Routledge.

Milner, A. D., & Goodale, M. A. (2008). Two visual systems re-viewed. *Neuropsychologia*, *46*(3), 774–785. doi:10.1016/j.neuropsychologia.2007.10.005

Noë, A. (2005). *Action in Perception*. The MIT Press.

Prinz, J. (2000). A Neurofunctional Theory of Visual Consciousness. *Consciousness and Cognition*, *9*(2), 243–259. doi:10.1006/ccog.2000.0442
Prinz, J. (2002). *Furnishing the Mind* (2004th ed.). Cambridge, MA.: MIT Press.
Prinz, J. (2006). Putting the brakes on enactive perception. *Psyche*, *12*(1), 1–19.
Rabinovich, M. I., Huerta, R., Varona, P., & Afraimovich, V. S. (2008). Transient Cognitive Dynamics, Metastability, and Decision Making. *PLoS Comput Biol*, *4*(5), e1000072. doi:10.1371/journal.pcbi.1000072
Reppas, J. B., Usrey, W. M., & Reid, R. C. (2002). Saccadic eye movements modulate visual responses in the lateral geniculate nucleus. *Neuron*, *35*(5), 961–974.
Royal, D. W., Sáry, G., Schall, J. D., & Casagrande, V. A. (2005). Correlates of motor planning and postsaccadic fixation in the macaque monkey lateral geniculate nucleus. *Experimental Brain Research*, *168*(1-2), 62–75. doi:10.1007/s00221-005-0093-z
Santos, B. A., Barandiaran, X. E., Husbands, P., Aguilera, M., & Bedia, M. G. (submitted). Sensorimotor Coordination and Metastability in a Situated HKB Model. *Connection Science*.
Stewart, J. R., Gapenne, O., & Paolo, E. A. D. (2011). *Enaction: Toward a New Paradigm for Cognitive Science*. MIT Press.
Thompson, E. (2007). *Mind in Life Biology, Phenomenology and the Sciences of Mind* (1st ed.). Harvard University Press.
Tognoli, E., & Kelso, J. A. (2009). Brain coordination dynamics: true and false faces of phase synchrony and metastability. *Progress in neurobiology*, *87*(1), 31–40.
Tononi, G., Sporns, O., & Edelman, G. M. (1994). A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(11), 5033 –5037.
Varela, F. J. (1995). Resonant cell assemblies: a new approach to cognitive functions and neuronal synchrony. *Biological Research*, *28*(1), 81–95.
Varela, F. J., Lachaux, J. P., Rodriguez, E., & Martinerie, J. (2001). The brainweb: phase synchronization and large-scale integration. *Nature Reviews Neuroscience*, *2*(4), 229–239.
Varela, F. J., Thompson, E., & Rosch, E. (1991). *The Embodied Mind*. Cambridge, Mass: MIT Press.
Werner, G. (2007). Metastability, criticality and phase transitions in brain and its models. *Biosystems*, *90*(2), 496–508. doi:10.1016/j.biosystems.2006.12.001

Acknowledgedments

Xabier E. Barandiaran currently holds a postdoctoral position funded by FP7 project eSMC IST-270212 (EU 7th Framework through "ICT: Cognitive Systems and Robotics") and also hold a Postdoc with the FECYT foundation (funded by Programa Nacional de Movilidad de Recursos Humanos del MEC-MICINN, Plan I-D+I 2008-2011, Spain) during the development of this work. XEB also acknowledges funding from "Subvención General a Grupos de Investigación del sistema universitario vasco. Grupo Filosofía de la Biología" from Gobierno Vasco IT 505-10.

Two Constraints on a Theory of Concepts

Andrea Onofri (Arche; University of St Andrews)

Self-censored by Copyright

The Proxytype Theory of Concepts

Mark Cain (Oxford Brookes University)

Introduction

Concepts play an important role in our cognitive lives as we employ concepts whenever we have a thought, engage in reasoning or categorise an object. Without concepts we wouldn't be fully-fledged thinkers and the stock of concepts that an individual has limits the thoughts that she is capable of thinking. In the light of this it should come as no surprise that the question as to the nature of concepts has been very prominent within cognitive science and the philosophy of mind in recent years. One of the most significant recent additions to this literature has been made by Jesse Prinz who, in his book *Furnishing the Mind*, develops a new theory of concepts that he dubs 'the proxytype theory'. Prinz firmly places his theory in the empiricist tradition and claims particular inspiration from John Locke and the contemporary psychologist Lawrence Barsalou. In this paper my aim is to evaluate the proxytype theory. Although I have profound admiration for Prinz's work in this area I will offer a number of criticisms.

The Proxytype Theory

The proxytype theory emerges as a result of an examination of the strengths and weaknesses of a number of competing theories of concepts that dominate the contemporary landscape. To describe and evaluate the proxytype theory it will be helpful to begin with an account of one of its competitors, namely, Jerry Fodor's informational atomism (Fodor, 1987, 1990, 1998).[1] Fodor is committed to the existence of a Language of Thought (LOT) (Fodor, 1975, 2008). Although LOT is not a public language such as English, Italian or Japanese, it shares key features of such languages. In particular, it has a battery of meaningful primitive symbols and syntactic rules for combining those symbols to form complex structures such as phrases and sentences. And the meaning of any such complex is determined by the meaning of its primitive components and the way they are put together (that is, the syntactic structure of the complex). Symbols can be realised in the brain. That is, just as a symbol of English can be physically embodied by means of a sound or a mark, a symbol of LOT can be physically embodied by means of a state of the brain. LOT is the vehicle of thought in that whenever an individual tokens a belief, desire or any other propositional attitude she will token a physically embodied sentence of LOT in her brain that has the appropriate content. For Fodor, concepts are symbols of LOT. To have the concept DOG then, is to have a symbol in one's LOT that has the content *dog*. This raises the question of the basis of the content of LOT symbols: why does the LOT analogue of 'dog' have the content *dog* rather than some other content or no content at all? It is Fodor's answer to this question that makes his theory a version of informational atomism.To a first approximation , he thinks that the content of a LOT symbol is matter of what reliably causes it to be tokened. So for example, the LOT symbol DOG has the content *dog* because its tokenings are caused dogs and only dogs. Or more precisely, because it is a law that dogs cause the tokening of DOG.

[1] Prinz himself adopts this tactic in his paper 'The Return of Concept Empiricism' (Prinz, 2005).

Fodor recognises that as it stands this won't do as tokenings of DOG are often caused by things that aren't dogs as when one mistakes a fox on a dark night for a dog or one thinks about dogs as a result of thinking about cats. So, one might ask, why doesn't DOG have the content DOG-OR-FOX-ON-A-DARK-NIGHT or DOG-OR-THOGUHT-ABOUT-A-CAT?[2] Fodor's answer is that there the dog-DOG causal relation is more basic than the other causal relations that DOG enters into in that the latter asymmetrically depend on the former. That is, were it not the case that dogs caused tokenings of DOG then it wouldn't be the case that foxes on a dark night (or thoughts about cats) caused tokenings of DOG, but not vice versa.

This theory is atomistic in that it rejects the thesis that the content of a concept is determined by its relations to other concepts so that, at least in principle, one could have the content DOG without having the concept CAT, ANIMAL or any other particular concept. Thus, for Fodor, concepts are certainly not theories. However, it is important to note that Fodor is happy to allow that complex mental structures such as beliefs and theories (encoded by means of LOT sentences) to mediate the content determining causal relations between concepts and what they represent. It is just that the content of those beliefs and theories doesn't enter into the content of the concepts in question. This explains why you and I could have quite different theories or beliefs about dogs yet still share the concept DOG.

Fodor's approach provides a helpful point of access to Prinz's proxytype theory. Prinz draws a distinction between long term and working memory. Thoughts are occurrent states as opposed to states that exist in the mind for lengthy periods of time. Thus thoughts reside in working memory. And as having a thought involves deploying a concept then concepts also exist in working memory. However, there is a close relationship between working and long-term memory in that items occurring in the former are often constructed from resources stored in the latter. Indeed, such a relationship exists in the case of concepts. With respect to concepts what exists in long-term memory are complex networks of representations. What binds together the elements of these networks are causal connections. The elements are causally connected in that activation of any one element of the network (an activation that involves its tokening in working memory) will typically cause the activation of some other element.

These networks stored in long-term memory correspond to categories of things in the outside world. For example, there is a network corresponding to dogs. Such a network was constructed over time on the basis of perceptual interactions with dogs. Moreover, the network is constructed out of representational primitives that are utilised by our various senses and so represent the kind of properties that we perceive objects to have. For example, these primitives have contents such as *red*, *edge*, *round*, and so on, where their content is a matter of what they casually covary with. Given that their basic representational elements come from a variety of sensory systems, the networks are multi-modal representations.

Prinz doesn't quite want to identify such networks with concepts for the reason alluded to above: concepts are involved in occurrent mental states that are located in working memory. When one employs a concept an element of a relevant network is activated. That is to say, an element is tokened in working memory. When this happens an element of the network goes proxy for the category the network in working memory. For example, whenever you employ the concept DOG in thought an element of a complex network stored in your long-term memory will be tokened in

---

[2] This is the so-called disjunction problem.

your short term memory. On different occasions and in different contexts you might token different elements of the complex. On all such occasions you are thinking a thought involving the concept DOG because the representation you token is drawn from one and the same complex, a complex that was constructed on the basis of interactions with dogs.

I began by stating that Prinz identifies concepts with proxytypes. We are now in a position to understand what this claim comes to. A proxytype is any element of a complex representational network stored in long-term memory corresponding to a particular category that could tokened in working memory to go proxy for that category. I also began by stating that Prinz's theory is an empiricist theory and we are now in a position to see what that claim comes to. In the context of concepts empiricism is often characterised as the view that all our concepts are learned as opposed to being innate. Now Prinz does think that the networks that proxytpes belong to are constructed on the basis of experience and so are not part of our innate endowment. However, the representational primitives out of which they are constructed are innate. What makes Prinz's theory empiricist is that these primitives are perceptual representations so that concepts are constructed out of perceptual resources. In other words, Prinz is endorsing Locke's (and ultimately Aquinas's) slogan that nothing is in the mind unless it was first in the senses.

There are several further features of Prinz's account that are worth bringing out. First, in virtue of the fact that different proxytypes are utilised on different occasions when thinking thoughts involving the concept DOG, we don't have a single concept DOG; rather we have many DOG concepts. However, Prinz points out, there is a likely to be a default proxytype that is employed when there is not sufficient context to result in the tokening of a more specific proxytype. Second, Prinz is committed to an atomist view of content. What gives a given proxytype its content is a matter of the content of the complex network that it is drawn form and the content of that network is a matter of the identity of the things that it was constructed on the basis of perceiving. For example, a DOG proxytype is an element of a network that was constructed on the basis of perceptual interactions with dogs.

A third additional feature of the account relates to Prinz's emphasis on the importance of concepts for categorisation and inference. When one categorises something as a dog what happens is a match is found between a current perceptual state and one of one's DOG proxytypes. And when one infers from this that the animal so categorised barks, the proxytype tokened in categorisation causes the tokening of another proxytpe belonging to the network that represents the barking aspect dog behaviour. This second proxytype will have been added to the network as a result of hearing dogs bark.

At this point it should be clear that there are considerable differences between Prinz's prototype theory and Fodor's theory, nothwithstanding the fact that both are committed to an atomistic view of the content of concepts. First, for Fodor concepts are amodal representations. That is to say, they are arbitrary symbols that do not take the form of any representations involved in perception. Prinz, on the other hand views concepts as being built from perceptual representations that are associated with a range of modalities and so that concepts are multi-modal representations. Second, Fodor regards most lexical concepts (that is concepts expressed by a morphologically simple words) as being simple representations whereas for Prinz such concepts are complex representations. Fodor doesn't deny that there are complex representational structures associated with concepts expressed by means of simple symbols of LOT. Consider DOG for example. For Fodor the fact that dogs reliably cause the tokening

of this LOT symbol – thereby playing a role in fixing its content – could depend upon complex structures that represent various properties of dogs including those that are readily perceivable. Such structures would serve as mechanisms that mediate the casual connection between dogs and DOG but they are not to be identified with the concept DOG.

Evaluating the Proxytype Theory

I now turn to the task of evaluating the proxytype theory. One interesting objection is implied by Edouard Machery (2009) as part of a general examination of work on concepts in both philosophy and psychology. Machery argues that psychological and philosophical work on concepts has quite different explanatory ambitions and so cannot be evaluated by the same criteria. Psychologists are primarily concerned with the mechanisms involved in categorisation, concept acquisition and inference (particularly inductive inference). Philosophers, on the other hand, focus on how it is possible for us to have thoughts, that is to say, propositional attitudes such as beliefs and desires. A core element of this project involves explaining how our thoughts manage to be about what they are about. Fodor would be a clear-cut example of someone whose work on concepts addresses a philosophical agenda. An example of a theory of concepts engaging with a psychological agenda would be any version of the prototype theory emanating from the work of Eleanor Rosch. The upshot of this is that it doesn't count against a psychological theory of concepts if it doesn't solve a problem of concern to a philosopher and vice versa.

The objection that this line of thought generates against Prinz is as follows. In motivating the proxytype theory Prinz examines an number of alternative theories developed by both philosophers and psychologists. He judges that all of these are ultimately unsatisfactory in virtue of failing to explain at least one important feature of concepts. Thus, a new theory is needed and the proxytypes theory constitutes this by explaining all the required features. Some of these features belong to what Machery would regard as a philosophical agenda and some to a psychological agenda. But if these agendas are independent of one another it is not incumbent on any theory to engage with both of them. Hence, the proxytype theory is designed to achieve a misconceived goal and the failure of competitor theories to fulfill that goal hardly counts against them.

I'm not convinced by this objection. For it to go through it would have to be the case that psychologists and philosophers were talking about quite different things when they used the term 'concept'. Indeed, Machery seems to be suggesting that this is the case as he says that 'concepts in psychology' are 'bodies of knowledge that are used by default in the processes underlying the higher cognitive capacities) (2009: 7) whereas 'concepts in philosophy' are 'capacities for having propositional attitudes' (2009: 31). I don't deny that there are differences in the aims, emphases and methods employed by, respectively, psychologists and philosophers yet Machery overstates the extent and significance of these differences. Historically philosophers interested in concepts have been concerned with how we acquire concepts, how we use them to categorise and how we make inferences involving them. The British empiricist philosophers Locke and Hume stand out in this regard. Moreover, it is difficult to see how psychologists couldn't be concerned with our capacity for thought. For isn't categorizing something as a dog a matter of thinking or believing that it is a dog? And isn't inducing from one's experience of several dogs barking that all dogs bark a matter of forming one belief on the basis of another? Of course a psychological theory of concepts doesn't have to explain every property of concepts. But a given theory is

problematic if it implies that concepts don't or couldn't have a property that we have independent reason to believe that they have. And it is this thought that lies at the heart of Fodor's objection.

I now turn to objections to the prototype theory that I regard as being more decisive. The first such objection emanates from a response to a criticism that Prinz (200?) directs at Fodor. Here Prinz argues that identifying concepts with amodal symbols fails to explain how we categorise the things we interact with and that this is a major failing given that categorisation is one of the primary functions of concepts. Consequently, in order to make sense of categorisation Fodor also needs to postulate complex representational structures that mediate the causal connection between concepts and the items that fall under them. In the case of DOG, this complex structure will represent the perceivable properties that dogs typically have. But, Prinz continues, the upshot of this is that his account should be preferred on grounds of simplicity. For, by identifying concepts with the kinds of structures that Fodor regards as mediating mechanisms he abandons any need to postulate additional amodal symbols.

A problem with this objection is that it overlooks the chief motivations for postulating the existence of a language of thought made up of amodal symbols. For Prinz categoristion involves the activation of a component of a complex network stored in long-term memory. For example, suppose I am confronted by a dog. A match is found between the perceptual state that the dog causes and a component of the network built on the basis of perceptual interactions with dogs. Thus, that proxtype is activated, an event that constitutes my categorising the animal before me as a dog. Suppose that the dog is silent when I perceive it but that I go on to infer that it barks. This will involve the proxytype I token causing the activation of another element of the network. This element will be a proxytype that was added to the network on the basis of experiences of dogs barking. The kind of reasoning portrayed here is based upon associative learning and involves the tokening of quite simple thoughts. Thus, on seeing a dog I think DOG (or IT'S A DOG) and go on to conclude BARKS (or IT BARKS). Now perhaps the proxytype theory can handle this kind of reasoning. But much of our reasoning is far more complex than this in the respect that it involves many steps, drawing upon information from a range of very different domains, making connections which outstrip one's experience, and tokening thoughts containing many concepts. Consider an example. Suppose that I have to collect my children from school by 6.00 p.m. at the latest. I'm running late as it is 5.00 p.m. and I've just come out of a meeting on a campus 30 miles away. Following my normal route home takes me 50 minutes but I don't automatically select this route as I reason that given the current time that route may well be subject to traffic congestion that would slow me down considerably. So I begin reflecting in order to work out if there any alternative routes that will get me home on time. In doing this I take into account a range of factors such as route lengths, speed limits, the number of roundabouts and junctions, the proximity of the routes to large residential areas, the amount of fuel I have in my tank, and so on. I eventually settle on a route different to my normal one and arrive with five minutes to spare. This is an example of everyday reasoning but it does seem quite distant from the kind that the proxytype theory seems well suited to handle. The relevant point in this context is that it is the kind of reasoning that has a logical character and so is readily explained in terms of the employment of logical rules or principles. But employing such rules involves applying them to representations that have an appropriate logical form. Now the simple symbols of LOT that Fodor postulates belong to a language that has syntactic rules for combining

those symbols to create more complex structures. These complex structures do not merely include complex concepts such as BROWN DOG but thoughts such as THE BROWN DOG THAT LIVES NEXT DOOR INVARIABLY BARKS WHEN THE POSTMAN DELIVERS A LETTER. In other words, they include thoughts that have precisely the kind of logical forms that enable them to be figure in processes of logical inference, processes that involve the application of logical rules and principles. In short then, an important motivation for postulating amodal symbols and identifying them with concepts is to make sense of our complex reasoning capacities. Prinz does think that proytypes can be combined but the kinds of examples he focuses upon involve the combination of two concepts like BROWN and DOG to form the complex BROWN DOG. But what he needs to show is that the proxytype theory can make sense of how we combine our concepts to create the kind of thoughts that we routinely have and that the resultant structures have a form that enables them to figure in processes of logical reasoning.

In a nutshell I have objected that Prinz focuses on simple inferences that, perhaps, can be handled by the proxytype theory, but overlooks the more complex thought processes that Fodor's approach is designed to handle. For what a theory of concepts needs to do is explain both how our concepts can be combined to form the complex thoughts that we are capable of having and do so in such a way that explains how such thoughts could figure in the reasoning processes that we routinely.

A second objection to the proxytype theory relates to Prinz's account of how proxytyes get their content. Prinz argues that the DOG proxytypes have the content they have because they are drawn from a complex network that was built on the basis of interactions with dogs. This readily accounts for misrepresentation for if, say, a fox causes the tokening of a proxytype from this network the fox will have been misrepresented as a dog in virtue of the historical origins of the proxytype. However, Prinz also argues that the networks are constructed over time at any point in their history new elements can be added to them. For example, if I encounter a pomeranian for the first time I may well add more to the DOG network in order to reflect what is distinctive about Pomeranians. But this generates a problem for it is highly likely that at some point interactions with non-dogs has led to additions to the putative DOG network implying that that network was constructed on the basis of interactions with a category of creatures broader than that of dogs with the implication that proxytypes drawn from that network have a content broader than *dog*.

A third objection once more relates to the content of our concepts. Since Putnam's (1975) classic article The Meaning of "Meaning"' externalism has become the orthodoxy in the philosophy of mind. According to such a view the protagonists in Putnam's Twin Earth thought experiment express different concepts by means of the word 'water' (and, therefore, different thoughts by means of sentences featuring that word). This is the case despite the fact that they are molecule for molecule duplicates. Earth dwelling Oscar expresses the concept WATER by means of 'water' in virtue of the fact that the local odourless, colourless liquid that he interacts with is water (that is, H_2O). Twin Oscar, on the other hand, expresses the concept TWIN-WATER in virtue of the fact that the local odourless, colourless liquid that he interacts with is twin-water (that is, XYZ).

The problem for the proxotype theory is this: how can it account for this divergence in content between the respective concepts of the twins and, therefore, the fact that they express different concepts by means of 'water'? Given that proxotypes are ultimately constructed out of perceptual representations the upshot would appear

to be that the twins have exactly the same proxotypes and, therefore, exactly the same concepts.

Prinz is alive to this problem and in addressing it he employs Locke's distinction between real and nominal essences. The real essence of water (that is, the colourless, odourless liquid found here on Earth) is a matter of its microphysical constitution. The nominal essence of water is a matter of the perceivable properties characteristic of water on the basis of which we typically identify a sample of water as such. Corresponding to this distinction is that between real and nominal content. The real content of the respective concepts expressed by means of 'water' by Oscar and Twin Oscar differ. This is because the stuff falling under Oscar's concept has the real essence of being H_2O whilst the stuff falling under Twin Oscar's concept has the real essence of being XYZ. On the other hand, their concepts have the same nominal content as the perceptual representations that figure in the proxytypes that constitute their respective contents are identical. This distinction between real and nominal content corresponds to the familiar one between broad and narrow content. In effect, what Prinz is saying is that the real content of a particular concept possessed by an individual is a matter of the essence of the items that the individual causally interacted with in constructing that concept. As Oscar interacted with H_2O in constructing his concept, that concept has the real content *water*. Whereas, Twin Oscar's corresponding concept has the real content *twin water* as it was constructed on the basis of casual interactions with Twin Water. This way of dealing with the problem posed by Putnam's thought experiment clearly echoes Prinz's approach to dealing with misrepresentation described above.

However, what I have said so far leaves out a crucial aspect of Prinz's line of thought and this has to do with his endorsement of a view that has become known as psychological essentialism. Psychological essentialism is a view that emanates from developmental psychology.[3] According to this doctrine children are innately essentialist about many of the categories for which they have concepts. That is to say, that children think that the items that belong to a particular category are bound together by having a common essence. An essence is a collection of properties that something must have to belong to the category in question and which are the underlying hidden causes of the readily perceivable properties of the category members. Thus, if a child were an essentialist with respect to the category corresponding to the concept WATER she would think that anything falling under that concept did so in virtue of having the relevant hidden properties, properties that are causally responsible for surface properties relating to its appearance and behaviour.

There is considerable empirical evidence in favour of psychological essentialism. To get a flavour of this evidence consider Frank Keil's (1989) classic experiment. Keil showed children and adults a picture of a racoon. When asked these subjects answered that the picture was of a racoon. They were then told that the pictured animal underwent a series of changes including changes to its appearance (through fur-dying its fur and plastic surgery), the insertion of a smell sac, and modifications to its behaviour. They were then presented with a picture of an animal resembling and skunk and told that it was of the original animal post-modification. When asked about the identity of the animal at this stage children over the age of seven and adults systematically answered that it was a racoon despite its appearance indicating that for

[3] Prominent champions of psychological essentialism include Keil (1989), Gelamn (2003) and Bloom (2004).

them something's being a racoon is a matter of its origins and/or hidden nature rather than its observable properties. Typically, psychological essentialists regard children as holding a placeholder conception of essence; that is, children do not usually have any substantial views as to the precise nature of the categories they adopt an essentialist attitude towards (Medin and Ortony, 1989).

Prinz endorses psychological essentialism. Thus, with respect to Oscar he would say that he thinks of the stuff falling under his concept WATER as having a particular essence (the nature of which he may well think himself ignorant) that is the causal basis of the perceivable properties in virtue of which he typically identifies a sample of water as such (that is, the properties that are represented by the relevant proxytype). Thus, Prinz accounts for the real content of Oscar (and our) concept WATER on the basis of Oscar's (and our) essentialist commitments along with the fact that that concept was constructed on the basis of causal interactions with H_2O. Without such an essentialist commitment the concept Oscar and we express by means of 'water' would have a content such as to apply to anything with an appearance like that of water. Thus, it would apply to XYZ as much as to H_2O.

What I will now argue is that that way of dealing with the problem of accounting for the content of our concepts in the light of Putnam's Twin Earth thought experiment is problematic with the upshot that Prinz cannot explain how Oscar and Twin Oscar can diverge in their concepts.

Essences of types of stuff do not always take the same form. Water has a microphysical essence. However, the same is not true of milk as can be seen by considering the following thought experiment. On an arid planet a team of super-intelligent robots who have never previously encountered water, synthesise a collection of H_2O molecules that they store in a beaker in their laboratory. These molecules form a colourless liquid that any visiting human would be unable to distinguish from water. Would this stuff be water? I contend that it would even though it has different origins from the water here on Earth and even though it doesn't play anything like the same role in the life of its home planet that water does here. For example, it doesn't fall as rain, fill any lakes or rivers or help sustain the life of any living creature. This is a simple consequence of water's having a microsphysical essence.

Now suppose that the robots take the water they have manufactured and mix it with a range of vitamins, minerals and fats that they have also synthesized so as to make something that is identical at the physico-chemical level to the glass of milk that I have just poured from a plastic bottle in my fridge. They don't drink this liquid and if they did it would certainly not provide them with any nourishment. Neither did they make it with the intention to provide nourishment for any other things. In fact, they are not in contact with any living things that would be nourished by the liquid. Question: is the liquid they have made milk? My answer is that it is not as what makes milk milk is not its physico-chemical properties per se. Rather, the essence of milk has to do with its origins and function; that it is manufactured in the body of a living creature with the function of sustaining and nourishing its young offspring. In short, the milk-like liquid the robots manufacture doesn't have the relevant origins and function to be milk.

Now consider Twin Earth where the liquid that they call milk – a liquid that is produced in the bodies of the creatures they call 'mammals' and is made and used to provide nourishment for the young offspring of those creatures – is largely made up of XYZ. Question: is this liquid milk? I would deliver an affirmative answer on the basis that it has a relevant origin and function.

In sum then, a sample of liquid can fail to be milk whilst being identical at the physico-chemical level to the milk in my glass and something can be milk whilst being very different at the physico-chemical level to that milk. What this implies is not that milk doesn't have an essence but that its essence isn't microsphysical or chemico-physical; rather it is functional or bio-functional.

Now suppose a child resident on Earth constructs a concept that she comes to express by means of the word 'milk' on the basis of interactions with samples of milk. Will that concept be the concept MILK, will it have the real content *milk*? Prinz would answer affirmatively. Now of course the samples of milk the child interacted with would all fall under the concept MILK. But they would also fall under a distinct physico-chemical concept due to the contingent fact that all milk here on Earth has the same basic physico-chemical makeup (for example, it is all largely made up of H_2O). The child's twin on Twin Earth would also be interacting with milk but the samples there would fall under a different physico-chemical concept as they were made up largely of XYZ. This raises the question of why the child here on Earth constructs the concept MILK rather than a distinct but locally co-extensive physico-chemical concept? Now Prinz needs to provide an answer to this question otherwise his proxytype theory will makes it a mystery how someone could acquire the concept MILK and imply that the concept most people express by 'milk' has an indeterminate content. It won't do to appeal to the child's essentialist commitments. Such commitments will only help if the child's essentialism takes the form of an idea as to the specific nature of the essence of 'milk'. In other words, the child will need to think that the concept she is constructing binds together samples of stuff not on the basis of their physico-chemical nature but on the basis of their bio-functional nature. Now one could coherently attribute to children such a precise essentialist commitment but it is difficult to see how Prinz could countenance such a view for the following reason. It is difficult to see how a typical child could arrive at such a view without explicit instruction or without it's being part of her innate endowment. The first option is hardly plausible for, as Paul Bloom (2000) points out, even educated Westerners don't talk to their children about essences. The second option hardly fits with Prinz's empiricism and his accompanying desire to restrict attributions of innate items to general learning mechanisms and perceptual representations.

This problem doesn't just apply to the concept MILK but also to the more familiar philosophical example of WATER. Every sample of water will fall under a concept that binds together samples of liquid that have a common origin, 'lifestyle' and role in human life and life in general. One might describe this as the concept of a liquid that fills rivers and streams, falls as rain, comes out of taps, and is fundamental to the survival of most living things. I argued that MILK is a bio-functional concept. With respect to the concept I am now describing, it might be described as a functional concept. Call this concept FWATER. Despite the fact that everything here on Earth that falls under the concept WATER also falls under the concept FWATER, and vice versa, the two concepts are not co-extensive as the XYZ on Twin Earth falls under FWATER though it is not water. And the H_2O synthesized by the super-intelligent robots described above falls under WATER but not FWATER.

So the problem for Prinz is to explain how we construct the concept WATER on the basis of our interactions with water rather than the concept FWATER whilst still making sense of how we construct the concept of MILK on the basis of our interactions with milk. A commitment to an unarticulated notion of essence will hardly work given that essences come in different forms and the concept FWATER is just as subject to essentialist analysis as that of WATER. What the child needs is an

articulated notion of essence distinct from that that she employs in constructing the concept MILK, one that enables her to represent the items falling under the target concept as being bound together by having a microphysical (rather than, say, a functional) essence. Once again, the question arises as to how the child acquires such a notion of essence and none of the available answers appear to be open to Prinz in virtue of his empiricism and the implausibility that children receive explicit instruction as to the general form of the essence that water takes prior to having a full grasp of the concept WATER.

In sum then, the proxytype theory has major difficulties explaining how we could acquire concepts such as MILK and WATER in virtue of the fact that these types of stuff have quite different kinds of essence.

Conclusion

In this paper I have given an account of Jesse Prinz's proxytpe theory and argued that it is open to three substantial objections. First, it cannot make sense of reasoning processes that go beyond the simple cases of inferring that something barks from the thought that it is a dog. Second, it cannot deal with the problem of misrepresentation. Third, it cannot explain how such everyday concepts as WATER and MILK have the concepts that they have in the light of Twin Earth thought experiments and their ilk.

References

Bloom, P. 2000: *How Children Learn the Meaning of Words*. Cambridge, MA: MIT Press.

Bloom, P. 2004: *Descartes' Baby: How the Science of Child Development Explains What Makes us Human*. New York: Basic Books.

Fodor, J. 1975: *The Language of Thought.* Cambridge, Mass: Harvard University Press.

Fodor, J. 1987: *Psychosemantics*. Cambridge, Mass: MIT Press.

Fodor, J. 1990: *A Theory of Content and Other Essays*. Cambridge, Mass: MIT Press.

Fodor, J. 1998: *Concepts*. Oxford: Oxford University Press.

Fodor, J. 2008: *LOT 2: The Language of Thought Revisited.* Oxford: Oxford University Press.

Gelman, S. 2003: *The Essential Child*. New York: Oxford University Press.

Keil, F. 1989: *Concepts, Kinds, and Cognitive Development.* Cambridge, MA: MIT Press.

Machery, E. (2009) *Doing Without Concepts.* Oxford: Oxford University Press.

Medin, D. and Ortony, A. 1989: 'Psychological essentialism.' In S. Vosniadou (ed.) *Similarity and Analogical Reasoning.* New York: Cambridge University Press.

Prinz, J. (2002) *Furnishing the Mind: Concepts and their Perceptual Basis*. Cambridge, MA: MIT Press.

Prinz, J. (2005) 'The Return of Concept Empiricism.' In H. Cohen and C. Leferbvre (Eds.) *Categorization and Cognitive Science*. Amsterdam*:* Elsevier Science.

Putnam, H. 1975: 'The meaning of "meaning"'. In his *Mind, Language and Reality: Philosophical Papers Volume 2*. Cambridge: Cambridge University Press.

Prinz's Theory of Conceptual Content

Marc Artiga (Logos; Universitat de Girona)

May 1, 2012

Abstract

In many of his arguments, Prinz's has heavily relied on a naturalistic account of conceptual content, which he has put forward and defended in several works (Prinz, 2000, 2002, 2006). In this essay, I outline his account of conceptual content and raise certain objections that suggest that this account should be abandoned.

1 Introduction

In this paper, I would like to discuss Prinz's naturalistic account of conceptual content. This is an aspect of his theory that has not been much discussed in the literature, if even some of his main arguments heavily rely on it. For instance, when Prinz (2006) argues that we can perceive abstract entities, he supports his argument with a particular view of how conceptual content is determined. In this essay, I would like to show that his own theory of content determination falls prey to important difficulties.

More precisely, here I will focus on Prinz's account of *referential* content (that is, truth-conditions), which Prinz distinguishes from something he calls 'Nominal Content' (Prinz, 2000) or 'Cognitive Content' (2002). The are two main reasons for that preference: first of all, Prinz's theory of referential content is the one he uses in most arguments in which a theory of content is playing an important role. Secondly, an account of Nominal Content (which, in any case, Prinz has not developed in much detail- see Prinz, 2000) will probably ride piggyback on a theory of referential content, so I think some of the problems of the former will probably carry over to any theory of Nominal Content.

The main goal of Prinz's theory of conceptual content is to explain in virtue of what process conceptual states acquire their content. In other words, Prinz wants to describe the process by means of which certain mental states come to have certain meanings. Why does my concept DOG mean dog rather than *cat* or *Obama*? This is a deep problem in philosophy that has generated an extense philosophical literature. Here I would like to outline Prinz's contribution to this important topic.

2 Prinz's Account

As he admits, Prinz's (2000, 2002, 2006) account is intended to be a combination of Fodor's (1990) Asymmetric Dependence Theory and Dretske's (1981, 1986) Informational Theory. According to him, for a concept C to have X as its content (that is, for C to mean X) two conditions need to be met: (1) there has to be a *nomological covariance* between C and X and (2) X must be C's *incipient cause.* Let us define both notions in some detail.

First of all, Prinz appeals to the notion of causal covariance between the concept and its referent. The intuition that the reference relation is determined by some notion of covariance is a common claim that has lead different proposals (e.g. Dretske, 1981, 1986; Rupert, 2008). However, Prinz *'s* concept of *nomological covariance* differs from other proposals in not being based on a covariance within the actual world, but across possible worlds. That is, C does not covary with X in virtue of the fact that the presence of C increases the probability of X's occurrence, as it is usually assumed. Nomological covariance has to do with covariance in proximate worlds. According to Prinz (2002, p. 241):

> COVARIATION Xs *nomologically covary* with concept C when, ceteris paribus, Xs cause tokens of C in all proximate possible worlds where one possesses that concept.

That is, John's concept DOG means *dog* partially because in all proximate possible worlds where John has DOG, tokens of this concept have been caused by dogs.

By appealing to causal relations that would hold in counterfactual situations, Prinz intends to to solve the 'Swampman problem'. The 'Swampman problem' is an objection based on a thought experiment, that was originally raised against certain historical theories of mental content, such as Millikan's (1984) and Papineau's (1984). Suppose that a lightening bolt strikes a swamp and a creature is produced (a 'Swampman') that happens to be microphysically identical to a normal human. Now, many people have the intuition that Swampman has representational states; since he is microphysically identical to a normal human, it seems he would behave and even talk in the same way as we do. However, any theory of content that requires that in order for a state C to represent X, there must be a causal relation between X and C is committed to the denying that Swampman has representational states, because nothing has caused his brain states. That is an unwelcome result for causal and historical theories of mental content.

But notice that, while Swampman lacks causal history, it seems his brain states support the same counterfactuals as we do*,* since *ex hypothesi,* swampman is microphysically identical to normal humans and the truth of many counterfactuals seem to be grounded on internal properties of human beings. So Prinz's notion of covariation seems to be in position to attribute representational states (and concepts) to swampbeings.

Nevertheless, Prinz is well aware that COVARIATION alone is too weak a relation for grounding semantic relations because there are many things men-

tal states nomologically covary with. First, my concept WATER nomologically covaries with water (H_2O), but it also nomologically covaries with XYZ, if in proximate worlds the transparent and colorless liquid that fills oceans and ponds is XYZ. In other words, anything that sufficiently resembles WATER would be included in he content of John's concept WATER (in the actual world). That seems to make concepts highly disjunctive. It seems we need to narrow down the set of possible candidate for content.

For this reason, (2000) Prinz adds a second condition: C means X only if X has caused the origin of the concept, that is, only if X is what Prinz calls the 'incipient cause' of C. In that respect, Prinz was inspired by Dretske's appeal to a learning period (1981). In a similar fashion, Prinz claims that a concept's reference should be identified with the cause that originated the concept.

In short, Prinz's view (Prinz, 2002, p.251) is the following:

INCIPIENT X is the intentional content of C if:

1. Xs nomologically covary with tokens of C and, in accordance with COVARIATION
2. An X was the incipient cause of C.

Let me now argue why I think this account is unlikely to be satisfactory.

3 Discussion

First of all, notice that there is some tension between 1 and 2. While 1 was designed to attribute representational states to Swampman, 2 precludes this attribution. Since nothing has caused Swampman's thoughts, there is no incipient cause of their mental states, and hence they are not about anything. In other words, by including incipient causes within the definition we are undermining the main motivation for endorsing preferring COVARIATION. Of course, there is still the intuition that concepts somehow covary with their referents, but it is not clear that the kind of covariation that has intuitibe support is the one put forward by Prinz. Furthermore, by adding 2, not only fails one of the main motivations for the thory: it shows that Prinz's account falls prey to the Swampman problem.

Secondly, Prinz does not provide any theoretical or empirical motivation for 2: why should we think the incipient cause plays such an important role? Why should we think the first cause of a mental concept plays a crucial role in fixing content? It is not obvious that this claim has intuitive support (though I admit that my intuitions my be biased at that point). Thus, as a first approximation, it seems INCIPIENT is not sufficiently motivated.

Indeed, I will argue that, even if independent reasons for motivating INCIPIENT were put forward, I think it suffers from serious difficulties. In particular, let me present 4 objections to Prinz's view. The first two arguments involve condition 1, the third argument involves condition 2 and the final remark is a general worry about this approach.

3.1 Indeterminacy

First of all, even if INCIPIENT can avoid including entities that exist in other possible worlds and ressemble very much the entities in the actual world (such as H_2O and XYZ), there are still many sources of indeterminacy that he does not properly address. For instance, John's MONARCH concept nomologically covaries with monarchs, but also with butterflies, and also certain retinal imagines (in particular, the retina image that is produced when seeing a monarch) because all of these states also cause John's MONARCH concept in all proximate worlds where monarchs cause them.[1] This is what most people call the 'Indeterminacy Problem' (which Prinz also calls the 'qua and chain problem'). Prinz is well aware if this difficulty, but he thinks condition 1 can deal with it:

> The first clause solves the qua and chain problems and can be embellished with further detail about the nature of the nomological relations involved to solve the semantic-marker problem.(...). For example, nomological covariance determines that my MONARCH concept refers to monarchs and monarch mimics but not to butterflies or retinal images, (...).

The problem is that, as it stands, 1 does not solve the chain problem. As we said, not only monarchs covary with C, but also butterflies, certain activations in the retina, neuronal activity in the optic array, and so on.

Prinz has outlined an original solution to this problem (which, interestingly enough, go beyond INCIPIENT), but it is unsufficient. Prinz (2002, p. 242-3) claims that whether a concept refers to a natural kind, an individual or an appearance is determined by a further condition, which he calls a 'semantic marker'. If, had the appearance X changed, X would still cause tokenings of concept C in the most proximate worlds, then X refers to a kind. If, instead a change in the appearance had stopped X to cause C in the most proximate worlds, then C is a concept of X-looking things. Of course, there are two serious problems with this view: First of all, *monarchs, butterflies* and *insects* are all natural kinds. So semantic markers are not fine-grained enough for the task at hand. Secondly, Prinz is inverting the order of explanation; it seems that the conditionals stated are true precisely *because* what concept C means rather than establishing the conditions for a concept to means anything. This is a general problem for his view that will be discussed below (3.4).

Indeed, it seems that even if we exclude states in different levels of distality (e.g. neuronal firings) and general properties (e.g. being a butterfly, being an animal) Prinz cannot explain why my concept MONARCH refers to monarchs rather than things that in the actual world resemble monarchs (like many other butterflies) because nothing ensures that the first thing that cause my MONARCH concept was a monarch rather than a similar butterfly. This problem will be extended and several consequences will be considered in 3.3.

[1] Indeed, in some cases the connection is much stronger. If monarchs are butterflies *necessarily,* then in all metaphysically possible worlds where a monarch causes MONARCH, a butterflies does.

So, pace Prinz, it is not easy to see how nomological covariance and the incipient cause can solve any of the problems of indeterminacy that affect other prominent theories of content.

3.2 Method of cases

The second objection is that the notion of nomological covariance appealed to in condition 1 causes INCIPIENT to attribute the wrong content to some mental states. On the one hand, condition 1 can be satisfied by the wrong entity playing the role of X. Suppose that John lost part of his visual capacities due to an extremely unlucky traffic accident when he was a child. Due to this impairment, he fails to distinguish oranges from tangerines. He applies the same concept to all of them. I think we would intuitively claim that his concept means something like *orange or tangerine*. Nonetheless, 1 and 2 might still hold in respect to oranges; it might happen that his first tokening of the concept was (by chance) caused by an orange. Furthermore, in all proximal worlds he has not had a traffic accident (remember that in the actual world he was extremely unlucky), so in these worlds he can perfectly distinguish oranges from tangerines and token this mental state only when confronted with oranges. So, it follows from INCIPIENT that *in the actual world,* his mental state means orange. But that cannot be right of John's actual concept.

Secondly, there seems to be cases where a subject has a concept even if condition 1 is not satisfied. Suppose John won the lottery. For this reason, he cancels a trip to Morocco and travels to China, where he bumps into an exotic fruit. He wonders how people call this fruit, how they would cook it,.. so John develops a well-formed concept of this fruit. However, in all proximal possible worlds, John does not win the lottery, so he travels to Morocco where he finds a different exotic fruit and wonders how do people call it, how they cook it,... So, again, INCIPIENT has as a consequence that in the actual world John lacks the concept that refers to the fruit in China because condition 1 is not satisfied.

Now, I think there is a plausible reply available to Prinz in support of the necessity and sufficiency of INCIPIENT.[2] Prinz could respond that the concept in the actual world and the concept in the counterfactual condition are different; since, according to COVARIANCE, in order to assess whether there is nomological covariance between the concept and its referent we must consider the most proximal worlds where a subject has *the same concept,* these counterexamples can be dismissed (this answer seems to be suggested in Prinz, 2002, p. 253) So, on the first example I gave, the concept applied to oranges and tangerines in

[2]One could claim that the 'ceteris paribus' clause in INCIPIENT is supposed to deal with this sort of cases, but it not easy to see how this clause should be interpreted (indeed, in Prinz (2002, p.13) there is no mentioning of 'ceterius paribus'). If 'ceteris paribus' is supposed to mean something like 'in normal conditions', it is hard to assess whether in the scenarios I present normal or abnormal conditions hold (without begging the question, of course).

A more general worry is that 'ceteris paribus' clauses are usually not accepted in theories of content determination without explicit analysis for a very good reason: these clauses seem to be introducing what has to be shown, namely what are the *normal* conditions for content determination (Fodor, 1990; Neander, 2006; Millikan, 2004)

the actual world is different from the concept applied to oranges in the counterfactual condition. Secondly, the concept I apply to an exotic fruit in China and the concept I applied to an exotic fruit in Morocco in the counterfactual condition are different concepts. So it seems Prinz has a satisfactory reply to all the cases I just presented.

However, I think this reply is utterly flawed. First, we may reasonably ask what grounds the claim that they are different concepts. In order for the reply not to be ad hoc, Prinz is required to provide some justification this assertion. The only way I see he could justify the claim that they are different concepts is either by appealing to the fact that they have different prototypes, proxytypes or functional roles or to the fact that they have different contents.[3] For instance, taking the similarity of content as a criterion, he could argue that in the first example the concept in the actual world (let us call it 'A-concept') means *orange or tangerine* and the concept in the counterfactual situation (C-concept) means *orange*. Since the only counterfactual condition that matters for content determination according to INCIPIENT is the one where the same concept is involved (the reply runs), and in the counterexamples there are always different concepts involved because they have different content, this is not a valid counterexamples to INCIPIENT. Unfortunately, this reply will not do for obvious reasons: Prinz cannot merely assume that the content of the two concepts differs, since what we are trying to settle is what determines the content of A-concepts. So he cannot individuate concepts across possible worlds by appealing to their content (at least, not when assessing whether a given concept satisfies 1 of INCIPIENT).

On the other hand, appealing to functional roles is also unsatisfactory, since in all the counterexamples we can stipulate that A-concepts and C-concepts share functional role in the mental economy of the subject: he is supposed to make the same inferences, perform the same actions,....Indeed, that gives us a good reason for thinking that the A-concept and the C-concept are indeed the same concept.

Prinz could adopt a different strategy. He could reply that the functional roles he appeals to in order to individuate concepts include wide dispositions (Harman, 1990); so, while in the actual world John is disposed to apply A-concepts to orange and tangerines, in the counterfactual world, he is disposed to apply it only to oranges. Since there is a difference in wide dispositions, there is also a difference in the functional role of A-concepts and C-concepts, and hence it seems Prinz could appeal to these dispositions in order to justify the claim that A-concepts and C-concepts are different. The problem, however, is that dispositions do not distinguish between right applications and mistakes, since we are also disposed to make errors. So, we can merely stipulate that in the counterfactual world, while he prominently applies 'orange' to oranges and orange has been the incipient cause, once in a while he makes mistakes and applies a C-concept to tangerines. If we add this condition, then the wide

[3] Prinz would probably opt for identifying concepts across possible worlds by appealing to something like proxytpe, prototypes or functional role (Prinz, 2002, p.7, p. 270)

functional roles of A-concepts and C-concepts are identical, and there is not reason to believe they are different. So the objection still holds.

3.3 Vagueness

The third problem is that, while it is usually thought that concepts can progressively change their meaning, Prinz cannot accommodate this fact without abandoning the key insight of his theory.

First of all, notice that many contentful concepts fail to satisfy 2. As Papineau (2006) points out, requiring that the content of the mental state has to be the first cause of the mental state seems too strong. If, for instance, one of our concepts is systematically tokened by a certain item, it is plausible to think that at some point it will come to represent this item, no matter whether it was the incipient cause or not. For instance, if the first time I saw a caiman I tokened the same concept that for the rest of my life I have used when I wanted to think about crocodiles, it seems very plausible to claim that I have been using the concept CROCODILE. But INCIPIENT entails that if when I created the concept it was caused by a caiman, then it represents caimans, and so I have been using the concept wrongly all my life. To say the least, that looks very implausible. Again, in this case Prinz suggests that the concept originally used for caiman and the concept I use most of the time are different concepts (Prinz, 2000, p. 253). Since they are different concepts, he seems to be able to accommodate the intuition that the concept I have used all my live in order to refer to CROCODILES in fact refer to crocodiles. Furthermore, in this case the he is not appealing to counterfactual worlds, so the problems raised earlier in identifying concepts across possible worlds do not apply.

However, when we consider the details of such an account, some tensions appear. Consider again the example in which the concept I have always been applying to crocodiles was incipiently caused by a caiman. Suppose at t_1 my concept C is caused by a caiman and at t_2 it is caused by a crocodile. Does the concept at t_2 mean caiman (and hence, it is wrongly applied to a crocodile) or is it the first tokening of a new concept (and hence it is rightly applied to a *crocodile*?) How can we know whether a concept is wrongly applied to an entity or whether it actually means something different? There are only two replies available to Prinz and none of them seems to be satisfactory. Prinz faces a dilemma.

On the one hand, Prinz can argue at t_2 John is correctly applying a new concept. The problem, of course, is that this account fails to account for cases of misrepresentation: if any case where the concept applies to a different item, this item counts as its incipient cause and it is considered a new concept, there will be no case where a concept is wrongly applied to a certain entity.

On the other, he can argue that that at t_2 John is misapplying C to a crocodile. Similarly, we can imagine that at t_3 John is confronted with a crocodile as well, and at t_4, and so on. As we saw, Prinz's answer is that after many tokenings of the concept being caused by crocodiles, at some determinate time t_n a different concept arises. Hence, (assuming INCIPIENT), there must be

a time $t_n$such that at $t_{n-1}$the concept was wrongly applied to a crocodile, and at t_n it suddenly becomes a new concept, whose incipient cause is a crocodile. I think this claim is very implausible.

The standard reply to this sort of cases is that at t_2 John is wrongly applying C to crocodiles, and there is no determinate point at which a new concept is created. Instead, there is a gradual change of meaning and, after a large number of times John has used C to refer to crocodiles, C gradually comes to mean *crocodile*. Unfortunately, this reply is not available to Prinz, since it contradicts the main insight of INCIPIENT, namely the appeal to an incipient cause. In a nutshell, the objection I am trying to raise is that INCIPIENT cannot account for progressive change of meaning. So, if condition 2 was unmotivated, now we see that we also have some reasons for rejecting it.

A related problem is that, according to INCIPIENT, non-deferential concepts can never have ambiguous contents (for an account of deferential concepts- see Prinz (2000)). Following INCIPIENT, if my concept JADE had not been deferential, it would either mean jadeite or nephrite, depending on the entity that first caused it. That is an implausible result since, as a matter of fact, some of our concepts are ambiguous (Millikan, 2000). So neither vagueness nor ambiguity can be accommodated within the theory.

3.4 Circularity

Finally, I would like to raise a general worry concerning this sort approach. A striking problem with INCIPIENT is that (as Fodor's Asymmetric dependence theory) we lack a (non-intentional) justification of why 1 should hold. Of course, it is true of many of our concepts that in the most proximal worlds the referent still causes them, but this is usually explained by appealing to the fact that concepts mean why they mean. In other words, Why do monarchs in most proximal worlds cause my concept MONARCH? precisely because MONARCH means monarch. The intuition that 1 is on the right track, comes from the fact if MONARCH means monarch, it seems the former will usually covary with the latter.

The root of the problem is that the truth of counterfactual statements is usually thought to be grounded in relations that hold in the actual world. For instance, consider the following counterfactual: *If Obama had not won the elections in 2008, McCain would have been the U.S. president.* We think this counterfactual is true because of certain causal relations holding in our world. The general problem with counterfactual accounts of content is that there is always the worry that the truth of the counterfactuals might be grounded on the intentional relations they are trying to explain. So, in order to provide a full characterization of a concept and its content, one should specify in virtue of what non-intentional property this nomological relation holds. The fact that no such characterization is provided, I think lends support to the suspicion that these accounts are merely assuming what they are supposed to show.

References

[1] Dretske, F. (1981) *Knowledge and the Flow of Information.* MIT Press.

[2] Dretske, F. (1986). Misrepresentation. In R. Bogdan (ed.), Belief: Form, Content, and Function. Oxford University Press

[3] Fodor, J (1991) *A Theory of Content and Other Essays,* MIT Press.

[4] Millikan (2000) *On Clear and Confused Ideas,* MIT Press.

[5] Papineau, D. (2006). Phenomenal and Perceptual Concepts. In Torin Alter & Sven Walter (eds.), *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism.* Oxford University Press.

[6] Prinz, J. (2006) Beyond Appearances: The Content of Sensation and Perception. *In Tamar Gendler & John Hawthorne (eds.), Perceptual Experience. Oxford University Press*

[7] Prinz, J. (2002) *Furnishing the Mind: Concepts and Their Perceptual Basis.* MIT Press

[8] Rupert (2008) Causal Theories of Mental Content. *Philosophy Compass*, 3: 353–380

The Missing Intentionality in Prinz's Theory of Emotion: (Historical) Reflections from Solomon

José Manuel Palma (Universidad de Granada)

It can be recognized in the actual debate about emotion two main lines of thoughts. Authors on the **intentional stand** defend emotions as intentional states. Intentionality is the property of such emotions of being *about*, being *directed towards* specific objects and events of the world, or to particular aspects of it. Generally speaking, defining emotions as intentional states is a way of pointing to the world in order to give their identity conditions. If emotions are about the world, it is more than possible that we can find identity conditions, at least in part, in some aspects or properties of the world, which would be the responsible of the elicitation of emotional responses. So, emotions are then defined appealing to some intentional component. Intentionality is a very special and difficult notion characteristic of language and thought. So, from the beginning, in these theories emotions have a strong link with language and, let me say, they share something like the same "structure". In this way, many proposals of the intentionality stand identify emotions with some characterization of thoughts, beliefs or judgments as evaluative. That is, as intentional mental states that appraises the world (Stoics, Spinoza 1677, Solomon 1976). For this reason, these different models of interpreting emotions have being grouped together in what is called the "*propositional attitude" model*. There is a sense in which emotions share essential features of language that makes them engage with propositional activities (as thinking and speak). So, in some sense, they have to present some propositional or linguistic "form".

The second main way of thinking about emotions is the **feeling stand**. Emotions are defined as feelings and their proponents apply a *perceptual model* to understand them. An emotion (feeling), like sense perceptions, is a perception of something. We can take back to Descartes (1649) for discovering the main lines of this perspective. Particularly, Cartesian model of perception presents the content of sense perception as an idea (res cogitans) generated by some impressions in our senses (res extensa). An emotion, in Descartes proposal, is like a second order perception of that idea generated by the res extensa; it is the idea over the sense idea. From this perspective, Hume's definition of emotions as "impressions of impressions" represents the same strategy. However, there was an important modification of that model of sense perception by one of the co-founders of modern psychology: W. James (1884, 1890). James, because of different reasons, felt uncomfortable with this model that sees emotions as *second order* perceptions or impressions. He brings to the debate of emotions a definition that sees them more naturally, as first order perceptions. They are perceptions of bodily changes. They are feelings of our body, not of our soul o res cogitans. Emotions, as feelings, are neither perceptions of some other ideas nor impressions of impressions. They are the direct perception of some bodily changes. So long, emotions are interpreted using the perceptual model, but now they are not second order perceptions. Therefore, still inside the perceptual model, James may defend a search of identity conditions of emotions in the res extensa, the body, being possible now a scientific study of them, something that was impossible with a second order concept of perception, such as the Cartesian, that only would admit introspection as a proper way to access or know anything about emotion.

As it is well known, both theories present big problems. Roughly speaking, on one hand, the intentional stand has difficulties explaining feelings and, because of that, the explanation breaks the continuity of emotions between linguistic and non-linguistic creatures. On the other hand, feeling stance has problems explaining intentionality, the part of the emotion that is directed towards particular aspects of objects and events of the world. The lack of a strong link between emotion and language, that makes possible to refer to a shared and structured world, makes difficult to properly accommodate emotions in the cognitive dimension

of linguistic creatures, and therefore, to explain properly this aspect of their intentional content. In feelings theories, emotions are reflects of goings on in our body (or mind), they are mental episodes that cannot explain the complexity and different roles in cognitions that some emotions plays in linguistic creatures. In this sense they are powerless, epiphenomenal states. Choosing one tradition in the search of identity conditions of emotions seems to advocate defining emotions only partially. It is in this context in which Prinz's theory makes its contribution. It is a serious attempt to cover the intentional demand, he would say cognitive, from James' concept of emotion.

Prinz recognizes the intentional problem in the traditional jamesian model and he confronts it wisely: he changes the model that sustains all emotion explanation. Emotions are still perceptions of bodily changes, but he is not going to explain perceptions as sense impressions. His concept of perception is derived from Dretske conception of mental representation. Let me call this the *representational model*, an interpretation of the perceptual mode based on dretskean representations. These representations are functional and a good and quick image of them is that of a *marker*, like a bright sign (which would be the feeling) that indicates us the presence of the emotional property. So, this somatic marker would represent the property that elicits it. Briefly sketched, when Prinz defines emotions as perceptions of bodily changes he is saying that emotions *represent* those properties of the world that elicit them. A property of events and situations of the world causes some bodily reactions. The somatosensory system *registers* these bodily changes (*nominal content*) and, because they have been reliably caused by those properties and somatosensory has the function of detecting them, *represents* those properties (*real content*). His appealing to core relational themes (Lazaru's cognitive account of emotions) for explaining these representations and the relational properties of the world responsible of them helps to think in this strategy as the adequate one for incorporating all the demands a theory of emotion has. In a conception where perceiving is representing in this dretskean sense, the identity conditions of emotions depends on, let's say, "both" nominal and real content. Better, if we think in perceiving as representing and the representation as a somatic marker, similar to a feeling that represents, nominal and real content are the same content, but heuristically analyzed from different points of view. In this way, it can be thought that Prinz does justice to both stands. The intentionality, (the properties of) the world that worried intentional stand (are) is included among the identity conditions of emotions along with feelings. Like good solutions, it is intended to show that both poles of the dilemma are pretty much the same saw from different perspectives.

However, a carefully reading of Solomon can challenge this successful solution. From the very first book of Solomon (1976), and beyond the hackneyed use of Solomon`s thesis that emotions are judgments, we can recognize in his proposal a rich searching of what it is important about intentionality. It is important to notice, from a beginning, that in a similar way to Prinz's view of emotions as feelings (bodily perceptions) with intentionality, Solomon applied something very similar to a perceptual model to his intentional stand. The notion he is appealing to is that of "perceptual judgments", on latter work, "kinaesthetic judgments" (Solomon, 2003), something that sound very similar to somatosensory perceptions. So, the propositional or linguistic character of his theory, proper of the intentional stance, it is not applied as the model he uses for explaining emotions. The judicative role that language plays in the theory it is not placed in the defense of a propositional, linguistic judgment as the model of emotions. The model is perceptual. The role of language is that of delimiting the place of the public dimension of emotions, "the politics of emotions" (Solomon, 1998). This public dimension refers to what can be shared, what would allow emotions participate, as cognitions, in public activities, as for example those of giving reasons. Using the famous statement of Pascal, it is not that "heart *has* reasons that reason [<therefore>, I would include in this context] cannot know", like in Prinz's view, where emotions are affects and cognitions are relegated as non emotional. Cognition is just input content, "calibration files" that causes emotions but are not part of them (like a folder has files but they are not the same). It is the idea of heart *is* a reason. Emotions are constituted under the background of language, which opens them to public

and shared aspects of the world. This public character of intentionality has to be explained. The role that language plays in this intentional theory is that of point to the public dimension of emotions, which is illustrated in the defense of the thesis that emotions, they themselves, may participate in our linguistic activities of giving reasons.

To elucidate this point a little bit more, we have to remember Solomon´s influences. In particular, the phenomenological-existential tradition: Sartre, Merleau-Ponty, Heidegger. In this tradition it is assumed, as Solomon endorses, the crucial influence of language in experience. It is not just in a labeling sense, a language that just put some labels to some linguistic-independent phenomena; but language as constitutive of the emotional phenomena, the phenomenological experience of which has a linguistic character, flavor. In other words, and echoing the words of another author also cited by Solomon, give or take some obvious differences: "*the limits of my language* mean the limits of my world" (Wittgenstein, 1922). Linguistic human beings have experiences linguistically structured, and for this reason they have experiences of a different kind from those of animals (and therefore we can *do* things with emotions that animals cannot). In Solomon's view of emotional experience, feeling theory, so long it sees the core of emotions as the same as in non-linguistic creatures (as somatosensory markers), cannot account this linguistically public dimension of emotional experience that tinges emotions in creatures engaged in linguistic practices. The core of this idea of intentionality is expressed by Solomon with the term "organic molecule" (Solomon, 1980): for example, "being-proud-of-my reparation in my car's wheel". We cannot separate in our emotional experiences, as atomism in emotions does, this conjunction: the evaluation and the particular object, concrete aspect of the world to which this evaluation is directed towards, cannot be analyze in two separate ways, like two combined but independent elements, when considering the whole emotional experience. For Solomon, the substantial difference that linguistically intentional emotions represent respect to animals is due to the fact that being directed towards x, in linguistic creatures, most times it is only possible thanks to a language, to be engaged in linguistic practices. This represents a difference for emotions, not only in their causal relation to some eliciting conditions (as Prinz's idea of calibration files that presupposes this kind of atomistic analysis), but a difference in emotion itself, in his experience. So, the concept of intentionality in Prinz's theory of emotion is not the one that the intentional stand remembers us as fundamental. Prinz's view of intentionality is his idea of real content, a representation as a sensory marker, that lacks the linguistic form of the particular intentionality Solomon is interested in and which explains the public o political conception of emotions.

I do not want to suggest that Solomon's theory of emotion is the solution for closing the gap between animal feelings and linguistic emotions. Even though the new role Solomon gives to feelings and the body in his last writings (vehemently neglected as parts of emotions in his firsts texts) through the concept of "judgments of the body" (Solomon, 2003), I think he cannot reach the bridge, he just points at it. It is still a problem for the intentional stand explaining emotional experiences in animals, mostly the continuity in feelings with us that they seem to express. For responding these questions one has to put so many matters up for discussion, and this is not the place here. I just want to highlight, as a point finger, that Solomon saw that the key of the answer to these conflicts rests on the category of action. I think Solomon thought about action as the sustenance of those things called emotions. References to Merleau-Ponty, elephant's example of Dewey, etc. show this. In my opinion, this is why Solomon always thought of emotions as an ethical matter, and insists so much on that concept of action in his last writings. For example, in the summary of his thoughts of emotions that his last book represents, he locates the concept of "engagement" as the starting point of his theory (Solomon, 2007). The way I interpret these ideas of Solomon it is not just to see how emotions influence our ethical decisions, but how emotional experiences themselves are publically modeled by actions, by interactive practice between subjects among them and with the world. So, I restricted my thesis to show that Prinz's theory of emotions does not satisfy this public criterion of intentionality that, through a language, allow emotions to be directed towards particulars aspects of the world only linguistically structured. He probably does not want to do so. This matter affects the most

primitive intuitions we have about emotions. However, I do not want to discuss these matters. I just wanted to show the unsatisfactory response of Prinz to the problem of emotion through the eyes of one of the most influential exponent of the intentional stance. For defending such idea, I have showed that Prinz is not really using the concept of intentionality that Solomon defends as the starting point of emotions' analysis. For concluding, I would like to show these different conceptions of intentionality in the analysis of other affective phenomena also very related to emotions: moods. In the explanation of moods it is crucial how it is conceived intentionality. If my thesis is right, and there is a difference in the concept of intentionality they are using, then they have to explain moods differently.

Moods are the best examples I found for illustrating this thesis. Other good examples are the so known Penfield cases. Moods are special for this matter because they are felt like emotions, but they seem to lack intentionality, the relevant sense of intentionality for Solomon: they are not directed towards *particular* objects or *concrete* aspects of events of the world linguistically structured. If my thesis is right, this should lead to a different conception of moods. And it is so. For Prinz, moods are intentional in the same way emotions are: they represent core relational themes as emotions, but they are caused by a calibration file with a wider scope, determined by a reliable causation of some more general eliciting conditions. Despite some doubt, Prinz affirms "I am inclined to conclude that moods are just a special case of emotions. They are not an independent category" (Prinz, 188: 2004). And it has to be so. Since he has affirmed that these calibration files (cognitive elements) are not part of emotions, this cannot make a substantial difference for excluding them as emotions. On the opposite, Solomon's view maintains that moods lack intentionality at all. They are not directed to particular aspects of the world, so they can be describe as being directed towards "all" aspects or towards "none", in the sense that they are indifferent, out of the reach of intentional considerations. If they describe so differently moods, this means they are deploying different concepts of intentionality. What is, from Prinz's perspective, a cause non-constitutive of emotions, and the intentionality they involve, it is rescued as the fundamental part of the core of emotions from Solomon's view of intentionality.

REFERENCES

Descartes, R. (1649): *Las pasiones del alma*, Editorial Tecnos, Madrid, 2006.

Prinz, J. (2004): *Gut Reactions: a perceptual theory of emotion*, Oxford University Press, New York.

Prinz, J. (2004): 'Embodied Emotions', in Solomon, Robert C. (2004): *Thinking About Feelings: contemporary philosophers on emotions*, Oxford University Press, New York.

Solomon, Robert C. (1976): *The passions: emotions and the meaning of life*, Hackett Publishing Company, Indianapolis, 1993.

Spinoza, B. (1677): *Ética demostrada según el orden geométrico*, Alianza Editorial, Madrid, 1988.

Solomon, Robert C. (1980): 'Nothing to be Proud of', in Solomon, Robert C. (2003).

Solomon, Robert C. (1984): 'Emotions' Mysterious Objects', in Solomon, Robert C. (2003).

Solomon, Robert C. (1998): 'The Politics of Emotions', in Solomon, Robert C. (2003).

Solomon, Robert C. (2002): 'Emotions, Thoughts and Feelings: What is a "Cognitive Theory" of the Emotions and Does it Neglect Affectivity?', in Hatzimoysis, A. (ed.) (2002): *Philosophy and the Emotions*, Cambridge University Press, Cambridge.

Solomon, Robert C. (2003): *Not passion's slave: emotions and choice*, Oxford University Press, New York.

Solomon, Robert C. (2007): *True to Our Feelings: What our emotions are really telling us*, Oxford University Press, New York.

Wittgenstein, L. (1922): *Tractatus Logico-Philosophicus*, Alianza editorial, Madrid, 1999.

What Feels Right, Objectively:
A Sentimentalist Rebuttal to Prinz's Sentimentalist Moral Relativism

Jake Davis (CUNY Graduate Center)

1 INTRODUCTION

On the sentimentalist line that Prinz (2007) borrows from Hume, moral concepts are response dependent properties. In particular, the values that are expressed in moral judgment are constituted by emotional dispositions – sentiments. Prinz takes this sentimentalist line of thought to a natural conclusion, though interestingly not one it seems Hume shared. Different cultures and subcultures inculcate varied and opposing moral sentiments. If sentiments are the truth-makers for moral facts, then there are varied and opposing moral facts. Indeed, Prinz suggests that there are no moral truths that hold for all human beings. Moral relativism may be a painful truth, since it does not allow us the comfort of privileging any particular system of morality. Nonetheless, hiding from this truth is hardly a solution; moral relativism is a fact we ought to learn to live with.

I develop in this paper a sentimentalist approach to morality that does not lead to moral relativism. While recent theorists of meta-ethics extensively employ (and some cases generate) research within moral psychology, they have not in general made use of a complementary set of research programs within cognitive psychology and affective neuroscience, investigating mechanisms of attention and emotional awareness. Prinz, for one, has developed elsewhere empirically oriented accounts of attention and consciousness (Prinz 2005a; Prinz 2010), emotion (Prinz 2004), and emotional consciousness (Prinz 2005b). Nonetheless, prominent sentimentalist theorists, Prinz included, have not made full use of recent work on the role of attention and emotional consciousness to refine their accounts of moral epistemology. Drawing on empirical research on attention training as a means of developing emotional awareness, I argue that our ability to converge on a thorough and unbiased awareness of the relative painfulness of various types of emotion can ground a circumscribed set of universal truths about how we ought to live, while leaving many other aspects of how we ought to live open to cultural determination.

2 PAINFUL EMOTIONS

I start from the intuitive claim that all human beings have strong preferences against being subject to unpleasant internal states such as painful emotions, and that we express this attitude in evaluative claims about which states are good and bad. Put another way, we share at least one value, that against painful internal states, and thereby a relative value in favor of pleasurable emotions over painful ones.

2.1 Valence and Preference

The notion of affective valence, as it is used in recent empirical literature, often conflates a number of separable aspects (Colombetti 2005). Emotions such as joy are often associated with approach (towards a pleasurable object), and emotions such as sadness with withdrawal. But the dimension of approach and withdrawal needs to be

separated from the hedonic tone of an emotion. Both craving and anger motivate approach behavior, for instance, but it does not follow that these are pleasurable; on my view they are both unpleasant. Indeed, it is the unpleasantness of craving that motivates us so powerfully to obtain whatever will appease it. Conversely, I suggest that the feeling of goodwill has a positive hedonic tone, and that we can be motivated to act in benevolent ways simply because it feels so good.

For this account to be viable, it is crucial that certain physiological reactions, for instance those involved in ill-will, have a negative hedonic tone, for all human beings. Importantly, I do not deny that ill-will is pleasurable for some of us, in addition to being unpleasant. We can like pain, and more generally we can have a preference for things that are negatively hedonically valenced. One way to make sense of this conflict is to suggest that certain physiological reactions have an intrinsic negative hedonic tone, independent of whether we have a preference for or against these reactions. There empirical as well as phenomenological reasons to be skeptical of such an account of hedonic tone as intrinsic; Prinz (2004) suggests that the (un)pleasantness of a perceptual objects consists in nothing more than that we (dis)like it. That is, our own preferences determine our pain and pleasure. I am agnostic on this question. However, if pleasantness is determined just by our preferences, then I hold that some preferences must be hard-wired and universal, for instance a preference against tissue damage in virtue of which it is negatively hedonically valenced. Thus I suggest that the reason the masochist gets pleasure from tissue damage is that she has a preference for something that is actually painful. We can make sense of this either by saying that she has a preference for something that is intrinsically painful, or else by saying that she has conflicting preferences. For my purposes here, either will do.

The pleasure of feeling goodwill can on some occasions have a kind of purity, I suggest, in virtue of not being mixed with painful feelings. In contrast, the pleasure that one might take in feeling ill-will towards an enemy will be always mixed with the pain of the physiological reaction involved in ill-will. It is not the case, however, that we are aware of the negative hedonic valence of emotional reactions such as ill-will on every occasion we have such an emotion. Indeed, it is crucial to my account that we often are not accurately aware of the pain and pleasure or our own emotions, but that with the appropriate training of attention, we can come to feel and to know the relative painfulness of various emotion types.

2.2 Some Improper Parts of Emotional Reactions

For the specific practical purpose of giving an account of episodes of joy, resentment, and so on in philosophical or moral psychology, I suggest, we can proceed by identifying the psychological and physiological changes present during these episodes, as well as the activity in the brain or elsewhere in the body that sustains these effects throughout the emotional episode. We need not establish which of these aspects, if any, corresponds to the folk-psychological notion of emotion. Such an approach can address the commonalities and differences between feelings of ill-will and feelings of benevolence, say, while remaining agnostic about whether emotions are a natural kind. This allows us to avoid debates about whether emotions are essentially cognitive or instead body-based, and whether emotions are essentially conscious. This does not mean that substantive aspects of research into emotional reactions are left out; on the contrary, the nature and causal relations of somatic and cognitive aspects may well come into more precise focus when not lumped together, for instance. This approach

also allows us to ask whether psychological processes that are especially associated with emotional reactions, such as affect valence, might nonetheless be present in cases where we would not be tempted to attribute an emotional reaction. And indeed, recent empirical work suggests that affect valence is pervasive in human psychology, being implicated in evaluative decision-making about everything from consumer choices to moral judgments (Loewenstein and Lerner 2003; Haidt 2007).

On this an approach we might think of emotional episodes as often involving a cycle of initial perception, triggering associated affective and somatic responses. These in turn can trigger thoughts, which may in turn trigger further affective and somatic reactions, and so on. The central point of this model is that while we cannot change the fact that being a conscious being interacting with the world involves both pleasure and pain, we can take responsibility for the pain and pleasure we cause ourselves in reacting to the world. There horrible things that happen in the world, and so on many occasions to perceive things as they are is to perceive things as painful. Put in terms of the model sketched above, an initial perception may often be associated with negative affect. My account makes an empirical prediction that these initial appraisals need not lead to further proliferation in cycles of emotional reaction.

This distinction between initial appraisal and subsequent cycles of emotional elaboration is of central importance for my purposes here, because it allows us to separate two questions about emotion that are apt to be conflated. The first is a question of warrant; we can ask whether an initial perception and its associated affective valence get the world right. Empathetic pain in response to seeing another's pain is unpleasant, but it does not follow on my account that we ought not to feel empathy. The project I undertake here is to provide a means to evaluate the various possible ways of further reacting to an initial painful or pleasurable perception of things in the world. Any way of reacting strengthens habits of reacting in that same way, and the ethics of emotion that I seek to develop suggests that some ways of reacting to pain and pleasure ought to be cultivated, and others ought to be attenuated. The means I suggest for discerning between these two is pragmatic, even hedonist. Some ways of responding to pain and pleasure feel much better than others.

3 Feeling and Knowing

The changes in neural activation and peripheral physiology underlying emotions can be fruitfully investigated (LeDoux 2000). In order to bracket the controversy over whether the term "emotion" should refer only to consciously experienced states, I will refer to affective reactions in general as "emotional". It is then a further question, but also an empirical one, under what conditions various aspects of these emotional reactions come to be consciously experienced, in the sense that there is "something it is like" for one to be undergoing these processes (Nagel 1974; Lambie and Marcel 2002).On the view I endorse below, only a subset of those stimuli that are consciously experienced become encoded in working memory and available for report and other explicit cognitive processing. To mark this difference, I will refer to subjects as aware of or knowing of their emotional reactions only in cases where they have the ability to explicitly report, recall, or deliberate on these emotions.

-

3.1 Attention as Unmasking

It is intuitively plausible that there is some relation between, on the one hand, attention to the somatic, affective, cognitive, and volitional aspects of our emotional reactions, and on the other our conscious experience of these aspects. Dehaene and colleagues suggest that attention is "prerequisite" for consciousness (Dehaene and Naccache 2001, 7–8). Their global-workspace theory of consciousness is motivated in part by appeal to studies on backward and simultaneous masking, especially in vision. When a stimulus is salient, it can mask others such that subjects do not report being conscious of weaker concurrent stimuli. When attention is cued, subjects are able to report on stimuli that were previously masked and unconscious.

However, global-workspace accounts illicitly collapse conscious experience with the ability to report a stimulus, in advance of the empirical data, as Ned Block (1995; 2007) has argued persuasively. On Block's view, we need to make a tripartite distinction between perceptions that are unconscious, those that are phenomenologically conscious, and those that are available to cognitive functions such as report, deliberation, and storage in long-term memory. Agreeing with this tripartite division, Prinz (2005a; 2010) makes the ambitious claim that attention is both necessary and sufficient for conscious experience, functioning to make perceptual information available for encoding in working memory. He thus holds that the somatic perceptions involved in our emotional "gut reactions" are consciously experienced only when they are modulated by attention (Prinz 2005b).

For Prinz, attention is necessary for phenomenal consciousness in a constitutive sense. If so, we will only be aware of the painfulness of emotional reactions to which we pay attention. It is worth noting in passing, nonetheless, that a more easily defensible account of the relation between attention and consciousness can deliver this same conclusion. Prinz's account neglects two crucial distinctions. First, cognitive scientists distinguish a number of different types of attention. Jin Fan, Michael Posner, and colleagues for instance, have distinguished top-down selective attention from the alerting mechanisms necessary for sustained vigilance, using behavioral as well as neurophysiological measures (Fan et al. 2002; Fan et al. 2005; Fan et al. 2007). Given these distinctions, it is not clear what Prinz's general claim that attentional modulation is necessary and sufficient for phenomenal consciousness amounts to.

Secondly, Prinz follows Block in conceiving of phenomenal consciousness as a state of experiencing in a rich and vivid way certain objects or properties, for instance a state of seeing red. Without such a notion of phenomenally conscious states as essentially including modality-specific content, it would make little sense to suggest, as Block does, that visual phenomenal consciousness might be realized by certain patterns of recurrent neural activity in visual areas of the brain (Block 2005). Prinz likewise argues that particular perceptual representations become phenomenally consciousness only through the modulation of attentional mechanisms. In contrast, Parvizi and Damasio suggest that there is a basic, core level of consciousness, dependent on the thalamus and brainstem, that occurs independently of selective attentional processes in higher cortical areas (Parvizi and Damasio 2001). This core or ground floor level of consciousness depends on a basic kind of alerting function distinct from the higher-level mechanisms of selective attention that come into play in determining what one is conscious of. On this view, the fact *that* there is a phenomenal feel—the fact that there is something it is like for a subject—depends on the basic alerting func-

tion. In contrast, the *content* of phenomenal consciousness—what it is like for a subject—depends also on how this consciousness is directed to particular objects and properties through selective attention. Put another way, the particular contents of phenomenal consciousness can be seen as modifications or modulations of a basal level of awareness dependent on the alerting function.

This distinction between the content and the occurrence of consciousness allows for a more easily defensible account of the relation between selective attention and consciousness experience. On the "biased-competition" model of attention developed by Desimone and Duncan (1995), representations in early sensory areas compete with one another for access to downstream resources, such as the mechanisms involved in conscious experience as well as those involved in cognitive access. The early visual system is tuned to pick up particular types of stimuli: motion, sharp edges, bright colors, and so on; for this same sort of reason, in the somatosensory modality, more intense stimuli tend to win out over weaker ones. But, crucially, top-down modulation by short-term task-goals and other representations in working memory also serves to bias these competitions in early sensory areas in favor of certain representations.

This approach allows that under normal conditions, where subjects are presented with numerous stimuli competing for processing resources, selective attention functions to make certain of these conscious. On this view, however, selective attention is not even partly constitutive of phenomenal consciousness. In the absence of competing stimuli, no modulation by the cortical areas involved in selective attention is necessary for a subject to be conscious of a particular stimuli. Understanding selective attention as unmasking selected stimuli provides a defensible account of how we consciously feel certain somatic and affective aspects of emotional reactions, and how these reactions can further come to be available for report, recall, and deliberation.

3.2 Developing Unbiased Emotional Awareness

In discussing the relations of attention, consciousness, and cognitive access, I draw in particular on recent empirical research on one kind of attention training, "mindfulness" meditation. Mindfulness practice can be broadly characterized by the aim to cultivate a clear awareness of one's own bodily, affective, mental, and perceptual processes, as they are occurring. The practice is derived from Buddhist sources, especially Theravada Buddhist teachers from countries such Burma, Thailand, and Sri Lanka. Nonetheless, it is the secular form of mindfulness practice pioneered in hospital settings by Jon Kabat-Zinn, and now widespread in clinical settings around the world, that has been the subject of the majority scientific investigations in this area. Drawing behavioral and neurophysiological evidence of changes correlated with mindfulness practice in emotional awareness and emotional biases, I emphasize the potential of this type of attention training to help us to correct the mistakes we make about the relative value of various emotion types.

Initial results do indicate that increases in bodily awareness due to mindfulness practice correlate with increases in emotional awareness. Comparing the effects of different types of bodily awareness training on subjective awareness of emotional response, Sze et al. (2010) found that meditators showed significantly more coherence between physiological changes and subjective awareness of emotional response than dancers and controls, and dancers showed an intermediate level of coherence. In reporting similar evidence of increased interoceptive awareness in a sample of female under-

-

graduates engaged in mindfulness training, Silverstein et al. (2011) suggest that women who were distracted by emotionally-driven self-evaluative thoughts were much slower in registering their bodily reactions, as measured by reaction time in rating physiological response to sexual stimuli, and that meditation training increased awareness of bodily reactions by deceasing self-evaluative thoughts. This explanation draws support from evidence that training in mindfulness meditation is associated with decreases in a network of brain regions associated with mind-wandering (Christoff et al. 2009; Brewer et al. 2011; Berkovich-Ohana, Glicksohn, and Goldstein 2012), and corresponding increased activation in visceral and somatic areas associated with body sensation (Farb et al. 2007; Farb et al. 2010). We are unaware of many of our own emotional reactions, but it is possible to train attention so as to develop emotional awareness.

Understanding mindfulness as a strategy of decreasing elaborative thought and enhancing phenomenal awareness helps to distinguish it from more cognitive strategies, such as changing how one thinks about the challenging or distressing situations one encounters in daily life. Garland and colleagues toward emphasizing the ability of mindfulness to facilitate specifically positive reappraisal, suggesting that a mindful attitude might allow individuals to reappraise a serious heart condition as "an opportunity to change their lifestyle and health behaviors rather than as a catastrophe portending imminent doom" (Garland, Gaylord, and Fredrickson 2011, 60). Traditional Buddhist presentations do not support a conception of mindfulness as biasing subjects specifically towards positive appraisal of life situations. Instead, the claim is that developed mindfulness allows subjects to 'see and know things as they are'. Affective bias underlies emotional distortions of attention and memory (Elliott et al. 2010). Judson Brewer, Hani Elwafi, and I have suggested that the role of mindfulness meditation in dispelling emotional distortions rests on its ability to attenuate positive as well as negative affective biases (Brewer, Elwafi, and Davis, forthcoming). This is a testable hypothesis; as opposed to putative biases in ethics, objective criteria in attention and memory tasks can be used to measure these more basic types of affective bias – and their attenuation in mindfulness.

If affective biases distort attention and memory, they will have impacts on the accuracy of our normative evaluations. In the case of internal states, even when we are aware of an emotional reaction, such affective biases might distort our awareness of its hedonic tone. Suppose that ill-will is actually painful, but also that in our culture ill-will towards certain groups is encouraged, in particular by the use of negative evaluative judgments about these people. Then, even if we are aware of the reaction of contempt, habituated affective biases may prevent us from attending to or accurately identifying the painful aspects of this emotional reaction. Conversely, if mindfulness can attenuate such affective bias, this kind of present-centered attention can give subjects more accurate knowledge of the relative pain and pleasure of various types of emotional reactions.

4 The Ethics of Emotion

In Sections 2 and 3, I have argued that increasing awareness of our emotional reactions in general and decreasing distorted awareness of their hedonic valence in particular can lead to convergence on the relative hedonic weight of various emotional types. In this section, I explore some implications of this account for issues in norma-

tive ethics. Although different cultures express and inculcate diverse attitudes toward the state of ill-will, for instance, if increased clear awareness causes subjects to realize that the physiology of such internal states is strongly unpleasant, this provides a defeasible but powerful reason for agreeing that it ought not to be cultivated. Moreover, this universal ethics of emotion has substantive implications for the ethics of action and character. Take an action such as expressing ill-will towards a group of individuals because of their ethnicity or sexual orientation. If it is the case that such an action can only be performed when one is motivated by ill-will, and if it is also the case that we ought not to be motivated by ill-will, then this provides a defeasible but powerful reason for agreeing that no one ought to act in such a way.

Yet, this cannot be the whole story. Considered as an internal state, the ill-will a Holocaust survivor feels towards her persecutors may be indistinguishable from that of the homophobe, but the surivivor's ill-will has a much better justification. In this final section I address the roles of reasoned justification and emotional feelings in deciding how we ought to live.

4.1 Reason, Rationalization, and the Currency of Decisions

Work by Jonathan Haidt and colleagues suggests that when pushed back far enough, people sometimes confabulate reasons to justify their moral judgments, reasons that cannot provide justification for the specific judgments in question (Haidt, Bjorklund, and Murphy 2000). Nonetheless, even if subjects themselves don't have access to good reasons for holding the values they do, as theorists we can still ask whether there are good reasons to hold these values. The problem for rationalist accounts of ethics is not that reasons can't be given to justify normative claims, or even that the available justifications can't be assessed. Rather, the problem is that the criteria different people and different peoples use for assessing those reasons, and the criteria for assessing those criteria, and so on, are themselves varied and variable. Standards of what is just, for instance, vary widely. But more importantly, the weight given to issues of desert and justice relative to other issues such as respect for authority or purity of heart, for example, vary radically between cultures, subcultures, and even between Utilitarians and Kantians inhabiting similar a similar intellectual culture. This is what Prinz and John Doris term the problem of "outer pluralism" in ethics (Doris and Prinz 2009).

This challenge to universal ethical claims is paired with the problem that Doris and Prinz term "inner pluralism." As a matter of descriptive fact, individuals hold various sorts of ethical values. Notions of duty, responsibility, and respect for others as well as expectations of pain or pleasure can be involved in determining our choices about how to live. Indeed, in many cases these various considerations compete to determine our actual choices. In order to compete in this way, importantly, there must be some common psychological currency our values share in. I follow Prinz and Jonathan Haidt, as well as Hume and James, in holding that affect provides this currency. On this approach, considering the various issues and entailments involved in a decision serves to trigger emotional reactions. Faced with a choice between smothering one's crying baby and causing the death of the whole group, the function of recruiting cognitive resources is to trigger affective responses that may compete with one's initial intuitive reaction. As Prinz (2007, 25) puts it, "we deliberate about moral dilemmas by pitting emotions against emotions." Haidt and Björklund take subjects' response to this crying baby dilemma as a paradigm case of the sort of affective reasoning described by Antonio Damasio, "there is indeed a conflict between potential responses,

and additional areas of the brain become active to help resolve this conflict, but ultimately the person decides based on a feeling of rightness, rather than a deduction of some kind" (Haidt and Björklund 2008, 195). My endorsement of this claim is qualified. First, it should be clear by now that on my account we need not consciously feel a certain emotion for it to play a role in ethical decision-making. Secondly, drawing on the account of emotional reactions sketched earlier, I suggest more specifically that it is various affectively-backed preferences that are in competition when values on holding perpetrators of injustice responsible come into conflict with values against having the pain of ill-will.

4.2 Which Emotions are Worth What?

What reasoning cannot do, Prinz and I agree, is to secure the priority of certain evaluative considerations over others. Difficult decisions about how to live always involve a contest between different values. My suggestion is first that because one value we share is a preference for not being subject to pain, for pragmatic reasons, we all ought to be aware in a thorough and unbiased way of the actual sources of our pain. And I maintain, secondly, that once we pay careful attention so as to feel and know the nature and weight of our own emotional reactions, cultivating painful emotions generally won't feel worth it. Thus even if one has a sentiment in favor of ill-will towards those who commit atrocities, or especially towards those who commit atrocities towards oneself, the pleasure one takes in maintaining this emotional reaction will pale in comparison to the pain it causes. Knowing in abstract terms that ill-will harms the person who has it much more than the one to whom it is directed may not motivate change, but fully and accurately feeling the pain of ill-will, I suggest, is a powerful motivation not to cultivate it in oneself.

Returning to the case of the victim, then, on my account it is better to be compassionate towards the perpetrator than to be vengeful, because it feels better. In many cases, fierce and forceful action may be required, out of compassion for the suffering perpetrators cause themselves by acting out of painful emotional states, as well as out of compassion for the suffering of their victims. Nonetheless, there is at least a conceptual possibility that an initial, painful, perception of an agent as causing intentional harm can lead to forceful action without requiring elaborative cycles of internally agitating emotional reaction, and therefore an empirical question as to whether it does. If it is possible to cultivate ways of being that achieve what is good more effectively than ill-will does, there is strong pragmatic sense in which that is how we ought to live.

We are also impacted by the emotions of those around us. Seeing another in pain activates areas in observers' brains associated with negative affect, for instance (Singer et al. 2004). Those fully and accurately aware of the pain of ill-will in themselves will be pained also by seeing other people consumed by ill-will, and will want people to be free from the suffering of such painful states. This gives rise to a number of crucial points.

We may express liking or dislike of certain emotional states in others and in ourselves by calling such states good or bad. However, one might like certain types of actions, traits, or states in an aesthetic sense, while being indifferent as to whether other people share these preferences. What makes ethical judgments interesting is that in such cases we are not indifferent about the preferences that others have. It is not just that

we think rape or genocide are blameworthy, we also think that liking such things is blameworthy. Thus Simon Blackburn suggests that it is only when such second-order preferences dispose us to praise or critize the preferences held by others that it becomes "a public matter, something like a moral issue" (Blackburn 1998, 9). In this sense we might not only prefer that others not have certain painful emotional states, we might also prefer that they share our preference against such states, and be disposed to blame those of who lack such preferences.

Justin D'Arms and Daniel Jacobson suggest that evaluative language functions as an interpersonal means of emotion regulation. "We use terms like 'disgusting' to do such things as criticize, persuade, or simply express disdain for others, and most generally to guide feelings—our own and other people's" (D'Arms and Jacobson 2000, 727). If this is right, evaluative language may be used to encourage others not to cultivate painful emotional states. Moreover, if it is preferences about the preferences that people have that dispose us to employ ethical language, the function of calling certain emotional states not just good but right may be to encourage people to like such states. Similarly, using the evaluative language of ethics to blame others for liking states that are objectively painful may function as an interpersonal means of encouraging preferences against such states.

This understanding of ethical evaluation, in turn, opens a way for understanding how expressions such as right and wrong can be used in a manifestation of compassion. One who feels and knows for herself the pain of objectively painful emotional states will be motivated to encourage others not to cultivate such states. One might accomplish this by explicitly telling others not to cultivate ill-will, for instance. Less explicit means may in fact be more effective, however; motivating people to change the preferences they have so as to dislike objectively painful states will not only change their relation to present emotional states but also increase their willingness to work to change their emotional dispositions. In calling ill-will wrong, one may succeed in alleviating not only other's present suffering, and not only their tendency to cause themselves harms in the future, but also their disposition to work so as not to be disposed to cause themselves the harm of painful emotional states. For one who feels and knows the pain of painful emotions, this will seem a good outcome, even the right one.

Building an account of ethical evaluation of this basis of empathy, however, also opens me up to a number of objections. For one, those endowed with of empathetic dispositions to feel the pain of others might nonetheless develop or even cultivate a disposition to feel great about causing harm, including feeling great about the bad feelings we have when seeing others pain. If my account in section 2 is cogent, however, such states inherently involve a conflict between preferences, a conflict that is painful in itself. Moreover, it if proves possible to have more purely pleasurable emotional states, that would be better.

Nonetheless, an ethics focused primarily on emotion rather than action may have nothing to say about why Stalin ought not to have had millions of his citizens murdered, if he was not in fact motivated by painful emotions such as fear or ill-will. Perhaps some psychopaths feel purely and simply great in carrying out atrocities. I don't think this is a weakness in the theory. We don't play the morality game with dangerous reptiles; the thing to do with a loose Tyrannosaurus Rex is not to evaluative his actions as atrocious, but simply to contain the threat. Similarly, towards those other-

wise human but utterly lacking in the basic emotional building blocks of morality, the appropriate response is containment, rather than punishment or other sorts of blame.

On the more positive end, because we are pained by others' pain, to the degree we allow ourselves to feel this pain we will be motivated to do what we can to create alleviate this suffering. Thus we will be motivated to act so as to stop people from causing pain to others, but also to themselves. If I am right, one primary way people cause pain to themselves is by cultivating painful types of emotional reactions. If so, we ought to be motivated to encourage others not to engage in cultivating painful emotions, or in acting out of them.

I have suggested that for those who are fully and accurately aware of their own emotions, ill-will won't feel worth it. Even if it does turn out that in some rare and particular cases, cultivating painful emotions has instrumental value, still, there will be vanishingly few cases in which a community of such ideal-observers of emotions will be motivated to encourage ill-will across the board, by maintaining a general norm in favor of such emotional reactions. Indeed, on my account, in some cases rationalizing norms in favor of the cultivation of painful emotions may itself be blameworthy.

5 Conclusion

Perhaps victims of rape or genocide ought to feel anger and ill-will towards their aggressors, at least for a while. Anger may be important for the psychological resolution of such traumas, and on my account there can be cases where painful emotions are instrumentally valuable. Nonetheless, for anger to be instrumentally valuable is for it to be a means to an end that is good and right. Some cultures may not value resolving the trauma of rape, or may not value it sufficiently to allow victims to feel or express anger. Indeed, this is likely the case in places where women are stoned to death for the putative crime of being raped. The challenge of moral relativism is the suggestion that there is no universal fact that causing that sort of trauma is the wrong thing to do, or that resolving the pain of trauma is the right thing to do. Prinz's moral relativism offers us no way independently of the idiosyncratic preferences and values held by the particular culture a rape victim finds herself in to answer the question of whether she ought to allow herself to be stoned to death or instead ought to be angry. My aim in this paper has been to push back against this sort of moral relativism from a plausible metaphysics of morals, establishing that there are some universal ethical ends to which anger, in particular cases, might be a means.

References

Berkovich-Ohana, Aviva, Joseph Glicksohn, and Abraham Goldstein. 2012. "Mindfulness-induced Changes in Gamma Band Activity – Implications for the Default Mode Network, Self-reference and Attention." *Clinical Neurophysiology* 123 (4) (April): 700–710. doi:10.1016/j.clinph.2011.07.048.

Blackburn, S. 1998. *Ruling Passions: A Theory of Practical Reasoning*. New York: Oxford University Press, USA.

Block, N. 1995. "On a Confusion About a Function of Consciousness." *Behavioral and Brain Sciences* (18): 227–47.

———. 2005. "Two Neural Correlates of Consciousness." *Trends in Cognitive Sciences* 9 (2) (February): 46–52. doi:10.1016/j.tics.2004.12.006.

———. 2007. “Consciousness, Accessibility, and the Mesh Between Psychology and Neuroscience.” *Behavioral and Brain Sciences* 30 (5-6): 481– 548. doi:10.1017/S0140525X07002786.

Brewer, Judson A, Hani M. Elwafi, and Jake H. Davis. “Craving to Quit: Psychological Models and Neurobiological Mechanisms of Mindfulness Training as Treatment for Addictions.” *Psychology of Additive Behaviors*.

Brewer, Judson A, Patrick D Worhunsky, Jeremy R Gray, Yi-Yuan Tang, Jochen Weber, and Hedy Kober. 2011. “Meditation Experience Is Associated with Differences in Default Mode Network Activity and Connectivity.” *Proceedings of the National Academy of Sciences* 108 (50) (December 13): 20254–20259. doi:10.1073/pnas.1112029108.

Christoff, K., A. M Gordon, J. Smallwood, R. Smith, and J. W Schooler. 2009. “Experience Sampling During fMRI Reveals Default Network and Executive System Contributions to Mind Wandering.” *Proceedings of the National Academy of Sciences* 106 (21): 8719.

Colombetti, G. 2005. “Appraising Valence.” *Journal of Consciousness Studies, 12* 8 (10): 103–126.

D’Arms, J., and D. Jacobson. 2000. “Sentiment and Value.” *Ethics* 110 (4): 722–748.

Dehaene, Stanislas, and Lionel Naccache. 2001. “Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework.” *Cognition* 79 (1-2): 1–37.

Desimone, R., and J. Duncan. 1995. “Neural Mechanisms of Selective Visual Attention.” *Annual Review of Neuroscience* 18 (1): 193–222.

Doris, J.M., and J.J. Prinz. 2009. “Experiments in Ethics.” *Notre Dame Philosophical Reviews* (October 3).

Elliott, Rebecca, Roland Zahn, J F William Deakin, and Ian M Anderson. 2010. “Affective Cognition and Its Disruption in Mood Disorders.” *Neuropsychopharmacology* 36 (1) (June 23): 153. doi:10.1038/npp.2010.77.

Fan, Jin, Jennie Byrne, Michael S Worden, Kevin G Guise, Bruce D McCandliss, John Fossella, and Michael I Posner. 2007. “The Relation of Brain Oscillations to Attentional Networks.” *The Journal of Neuroscience* 27 (23) (June 6): 6197–6206. doi:10.1523/JNEUROSCI.1833-07.2007.

Fan, Jin, Bruce D. McCandliss, John Fossella, Jonathan I. Flombaum, and Michael I. Posner. 2005. “The Activation of Attentional Networks.” *NeuroImage* 26 (2) (June): 471–479. doi:10.1016/j.neuroimage.2005.02.004.

Fan, Jin, Bruce D. McCandliss, Tobias Sommer, Amir Raz, and Michael I. Posner. 2002. “Testing the Efficiency and Independence of Attentional Networks.” *Journal of Cognitive Neuroscience* 14 (3): 340–347. doi:10.1162/089892902317361886.

Farb, N. A. S., A. K. Anderson, H. Mayberg, J. Bean, D. McKeon, and Z. V. Segal. 2010. “Minding One’s Emotions: Mindfulness Training Alters the Neural Expression of Sadness.” *Emotion* 10 (1): 25–33.

Farb, N. A. S., Z. V. Segal, H. Mayberg, J. Bean, D. McKeon, Z. Fatima, and A. K. Anderson. 2007. “Attending to the Present: Mindfulness Meditation Reveals Distinct Neural Modes of Self-reference.” *Social Cognitive and Affective Neuroscience* 2 (4): 313.

Haidt, J. 2007. “The New Synthesis in Moral Psychology.” *Science* 316 (5827): 998.

Haidt, J., and F. Björklund. 2008. “Social Intuitionists Reason, in Conversation.” *Moral Psychology* 2: 241–254.

Haidt, J., F. Bjorklund, and S. Murphy. 2000. “Moral Dumbfounding: When Intuition Finds No Reason.” *Unpublished Manuscript, University of Virginia*.

Lambie, J. A, and A. J. Marcel. 2002. “Consciousness and the Varieties of Emotion Experience: A Theoretical Framework.” *Psychological Review* 109 (2): 219–259.

LeDoux, J. E. 2000. “Emotion Circuits in the Brain.” *Annual Review of Neuroscience* 23: 155–184.

-

Loewenstein, G., and J. S. Lerner. 2003. “The Role of Affect in Decision Making.” In *Handbook of Affective Sciences*, ed. R.J. Davidson, K.R. Scherer, and H.H. Goldsmith. Oxford University Press, USA.

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

Parvizi, J., and A. Damasio. 2001. “Consciousness and the Brainstem.” *Cognition* 79 (1-2): 135–160.

Prinz, Jesse J. 2004. *Gut Reactions: A Perceptual Theory of Emotion*. New York: Oxford University Press.

———. 2005a. “A Neurofunctional Theory of Consciousness.” *Cognition and the Brain: The Philosophy and Neuroscience Movement*: 381–396.

———. 2005b. “Are Emotions Feelings?” *Journal of Consciousness Studies, 12* 8 (10): 9–25.

———. 2007. *The Emotional Construction of Morals*. New York: Oxford University Press.

———. 2010. “When Is Perception Conscious?” In *Perceiving the World: New Essays on Perception*, ed. B. Nanay, 310–332. New York: Oxford University Press.

Silverstein, R. Gina, Anne-Catharine H. Brown, Harold D. Roth, and Willoughby B. Britton. 2011. “Effects of Mindfulness Training on Body Awareness to Sexual Stimuli: Implications for Female Sexual Dysfunction.” *Psychosomatic Medicine* 73 (9) (December): 817 –825. doi:10.1097/PSY.0b013e318234e628.

Singer, Tania, Ben Seymour, John O’Doherty, Holger Kaube, Raymond J Dolan, and Chris D Frith. 2004. “Empathy for Pain Involves the Affective but Not Sensory Components of Pain.” *Science* 303 (5661) (February 20): 1157–1162. doi:10.1126/science.1093535.

Sze, Jocelyn A., Anett Gyurak, Joyce W. Yuan, and Robert W. Levenson. 2010. “Coherence Between Emotional Experience and Physiology: Does Body Awareness Training Have an Impact?” *Emotion* 10: 803–814. doi:10.1037/a0020146.

Morality and the pro-social emotions: a nativist view.

Alejandro Rosas (Universidad Nacional de Colombia)

1. Introduction: Morality differs from pro-social emotions

In one of his central arguments against nativism, Prinz uses a very important distinction between dispositions towards morally praiseworthy behavior (that we probably share with other pro-social animals) and dispositions to behavior guided by moral evaluations or morality proper. Prinz develops two sides to this distinction and I agree that he captures distinctive features of morality. But I am less convinced that they speak against nativism. This passage summarizes the two relevant features:

"...biologically based behaviors...are culturally malleable and insufficient to guide our behavior without cultural elaboration. I think culture makes two contributions to the biological inputs. First, it converts these behaviors into moral norms, by grounding them in moral emotions. Second, it takes the biologically based norms that have highly stereotyped...behavioral effects in our primate cousins and alters them into culturally specific instructions."(Prinz 2007, 277, my italics)

What he calls "biologically based norms" are in fact pro-social emotions like altruism and concern for others. They are inputs for moral emotions, but differ from them for two reasons. First, they differ along the axis generality/specificity. The evolved (innate) pro-social emotions are "too vague to translate into action" (276). They are best viewed as general constraints that can only acquire specific content through culture; and with specificity comes cultural variability:

"...strictly speaking, there is no such thing as an evolutionary ethics…evolved norms [and their adjoined emotions, AR] do not constitute an innate morality. They are, instead, flexible constraints within which morality emerges." 259. We can understand many human moral norms as culturally specific variations on the same biological themes." (Prinz 2007, 274)

Secondly, moral emotions are a distinct type of motivation when compared to mere pro-social emotions. For example, behaviors against sexual fidelity or rank naturally elicit aversion, even in some non-human animals. But moralization of rank and sex is not equivalent to aversive emotion. Moralization means that the thought "it is wrong" is added on top of these emotions. The distinctive character of moral wrongness is important to moral philosophers (see Joyce (2006) and comments in De Waal (2006)). In itself, however, it does not preclude a nativist view, as we shall see.

Both cultural variability and distinctive motivation are important for morality, but Prinz lacks a proper view of how both belong together. This shortcoming emerges in a passage where he tries to explain the need for a distinctive moral motivation. Prinz says that in small-scale

societies “natural niceness”, without morality, was probably enough to produce praiseworthy behavior:

“Very small-scale human societies may not require moral rules, because members of those societies are close enough to be naturally inclined to treat each other well. As population size grows, however, we find ourselves in contact with people who are not close friends or family… Expansion places pressure on cultures to devise ways of extending our natural niceness to strangers. Moralization offers a solution.” (Prinz 2007, 273)

But if un-moralized pro-social emotions suffice to guide praiseworthy behavior in some human societies, it cannot be right that, without culturally constructed rules, evolved emotional constraints are “too vague to translate into action”. Prinz says elsewhere, commenting on Joyce 2006, that all (human) societies need rules (Prinz 2008). In his replies, Joyce puts his finger on the contradiction between this assertion and the one about “natural niceness” (without moral rules) in small-scale societies (see Joyce, 2008). One way out of the contradiction is to deny that rules are dispensable in small-scale societies. But we still need an explanation for why rules are always needed, and I think “vagueness” of evolved emotions is the wrong turn to take. I think two biological facts can provide the explanation. In the following I link these two facts to the distinctive features of morality; and I shall argue that this connection makes the case for an innate morality more attractive than Prinz is willing to acknowledge.

2. Motivational distinctiveness

It is possible, though not yet completely clear, that apes feel pro-social emotions like altruism or concern for others. However, these emotions alone do not make for a moral creature. Morality requires a second-order evaluation of behaviors and first-order emotions. These evaluations are, for example, present in self-directed moral emotions like guilt and shame. These are necessary psychological prerequisites for having ought-thoughts or “oughtitudes” (Prinz 2007, 262). Prinz says:

“We often say that genuine altruism is a form of moral behavior. But the phrase ‘‘moral behavior’’ is ambiguous. It can mean either behavior that we find morally praiseworthy or behavior that is driven by moral evaluations. Suppose apes help each other out of genuine concern. ...This tells us something about the evolution of moral decency, but it tells us nothing, I submit, about the evolution of morality... A creature could behave in noble ways without any capacity to judge that actions are good…These are different...I do think there are important psychological prerequisites on having ought-thoughts, or ‘‘oughtitudes.’’ For an ape to think that he ought to share…he must feel guilty if he doesn’t share. He must also feel angry at those who do not share with him. This kind of motivation differs from what evolutionary ethicists call altruistic motivation..” (Prinz 2007, 261-262, my italics).

However, though evolved pro-social emotions like altruism and concern are not equivalent to moral emotions, I do think that they are necessary, though not sufficient, for having moral feelings. This seems likely, particularly in creatures that also have strong evolved drives towards selfishness. When inclined to follow a course of action that makes me better off at the expense of making others worse off, I can feel the contrary pull of other-regarding feelings of concern. In such a conflict, a third feeling, a moral one, would stop me from too readily yielding to selfish desires. Moral emotions rule that I should act on the requirements of concern, and that I should feel bad about the prospect of ignoring them. This capacity for norms about how I ought to feel and which of two conflicting desires I ought to follow, is essential for morality. Here lies precisely the element that distinguishes a praiseworthy emotion and/or behavior from a strictly moral emotion or behavior.

Prosocial emotions are necessary if a selfish creature is to be able to evolve moral feelings. More precisely, it is the biological conflict between selfishness and pro-social emotions that makes possible the emergence of a meta-emotion telling me that concern should override the selfish temptation. Prinz admirably recognizes the link between meta-emotions and morality. But he is more concerned about explaining their derivative character: He says: “Indignation is not a basic emotion; it derives from anger. Indignation is anger calibrated to injustice.” Or “Guilt is sadness that has been calibrated to acts that harm people about whom we care.” (Prinz 2007, 77). Instead, I would here emphasize the importance of the conflict between prosocial emotions and selfishness, and the fact that the meta-emotions are calibrated to avoid neglect of the prosocial emotions. Norms of justice or norms against harming others only originate in the first place when this third moral feeling emerges to resist the pull of selfishness and avoid the neglect of prosocial emotions.

3. Cultural variability and the public negotiation of rules

The conflict between selfish tendencies and pro-social emotions is one of the necessary biological conditions for the emergence of moral norms or meta-emotions, but Prinz’s makes no use of it in his theory. Moral emotions condemn the selfish neglect of pro-social emotions, which partly explains concepts like duty or moral wrongness. However, these concepts involve one further and very important factor. The organism that experiences the inner conflict and evolves meta-emotions is also engaged in a process of public negotiation. The reason is that the internal conflict affects other group members if it is resolved in favor of selfishness. Others have an interest in the selfish outcome being avoided as often as possible. The solution must be in the range of what is usually understood as fairness, taking both self and others equally into account. Solution points are publicly negotiated and then fixed through explicit and public rules. This drive to publicity is a further biological pre-condition for the emergence of the typically moral emotions and norms at a higher level, and it completes the concept of duty or moral wrongness. The management of internal conflicts with a view to public rules explains both the distinctive character of moral motiva-

tion and the fact that moral rules are specific and culturally variable, for public negotiation happens in a particular time and place. The cultural variability of specific moral rules comes from the conflict between pro-social and selfish emotions and from its public negotiation. We do not understand meta-emotions like guilt or shame unless we factor in their connection to the public negotiation of rules.

This type of relationship between the moral and the basic pro-social emotions can be traced in any of the domains for which we have moral rules, including the domains of sexuality and rank or hierarchy. A necessary condition for the emergence of morality is the existence of a conflict between selfish and pro-social drives. For example, natural selection will program feelings motivating fidelity (a pro-social emotion) between mates in species where bi-parental care is obligate. But feelings driving to infidelity persist, because they are also adaptive. Animals follow one or the other according to circumstance. They may form private rules for when to follow one or the other; or they may not, and wantonly follow their moods. But animals lack public rules to solve their inner conflicts, whether they form private rules or not. As Prinz says:

> "Infidelity does occur. …even bird species known for their long-term monogamous relationships often sneak in some romance on the side. There is no evidence that non-human animals regard such behaviors as immoral, rather than merely risky." (280)

Only humans (for all we know) have developed the moral emotions and meta-norms to manage those conflicts. Only humans publicly negotiate the conditions under which following their drives to infidelity constitute, or not, a breach of fidelity. This negotiation produces different norms depending on contingent cultural or environmental factors. Prinz refers to cultures that allow women extramarital sex, not considered infidelity. The source of this variation is not that pro-fidelity emotions are intrinsically vague, but that conflicts between selfish and pro-social feelings in humans are negotiated in a public way. When all parties have to be publicly committed to a solution that represents a regular and predictable behavior, specific rules are negotiated. Since the negotiation happens in a particular context, variable rules come into being.

4. Innate morality

In Prinz's view, morality is a bio-cultural phenomenon, where culture has incomparably greater weight than biology. The "vagueness" of the biologically evolved dispositions serves to emphasize this greater weight in his theory. I have rejected this idea above. Cultural variability and specificity comes from a biologically given inner conflict and its public negotiation.

Another strategy to emphasize the weight of culture is to present morality as a by-product. "The best way to defeat nativism" Prinz says, "is to present an alternative account. ... If

morality can be explained as a byproduct of other capacities, there is little pressure to say that it is innate." (Prinz 2007, 269) The capacities he has in mind include emotions, memory, rule-formation, imitation, and mind-reading (272).The two distinctive features of morality – cultural specificity and motivational distinctiveness – are closely connected to mind-reading and rule-formation. This connection can be illuminated by viewing moral emotions as built upon two biological facts: a conflict between selfishness and pro-social emotions, and a pressure to manage this conflict through public rules. There is little doubt that the existence of a conflict between selfish and pro-social emotions is an inherited fact in many organisms. On the other hand, the pressure to manage it in a public way was probably ubiquitous and ancestral in human social evolution. Meta-representational abilities, the attribution of mental states to others (mind-reading) and the ability to follow rules were required, as humans had to be aware of the conflict and of its public management.

Prinz's by-product theory assumes that meta-representation, mind-reading and the ability to follow rules were already there and were "exapted" for morality. But what other function could they have served, besides the one we are here concerned with? What else could have driven their evolution? One possible answer is manipulation of others (Humphrey 1976; Alexander 1987). Agreed, but arguably, morality and manipulation are related phenomena and evolved under similar pressures. In any case, it seems likely that the pressure to deal with an inner conflict in a public manner drove and further shaped the evolution of those abilities. At the least, they co-evolved with the pubic management of inner conflict. Their genetic basis must have been influenced by that fact. Given this, it is not preposterous to assume that a tendency to public justification is biologically inherited in humans (see Carruthers and James 2008). As for why our ancestors felt a social pressure for the public negotiation of conflicts between selfishness and pro-social emotions, I already mentioned that the conflict harms other group members if it is resolved in favor of selfishness. I close with a quote from Joyce (2006, p. 117) that connects moral judgment to group cohesion through shared or public justification: "...moral judgments can act as a kind of "common currency" for collective negotiation and decision making. Moral judgment thus can function as a kind of social glue, bonding individuals together in a shared justificatory structure and providing a tool for solving many group coordination problems."

References

Alexander Richard (1987) The Biology of Moral Systems. New York: Walter de Gruyter

Carruthers, P and James, Scott (2008) Evolution and the possibility of moral realism. Philosophy and Phenomenological Research 77(1): 237-244

De Waal, Frans (2006) Primates and Philosophers. Princeton and Oxford: Princeton University Press.

Humphrey, Nicholas (1976) "The Social Function of Intellect," in Bateson, P. P. G. and Hinde, R.A., (eds.) Growing Points in Ethology, p. 303-317. Cambridge: Cambridge University Press.

Joyce, R. (2006) The Evolution of Morality. Cambridge MA, The MIT Press.

Joyce R. (2008) Replies. Philosophy and Phenomenological Research 77(1): 245-267.

Prinz, Jesse (2007). The Emotional Construction of Morals. Oxford: Oxford University Press.

Prinz Jesse (2008). Acquired moral truths. Philosophy and Phenomenological Research 77(1): 219-227