Volume 8, Number 2
2015

ISSN 1807-9792

abstracta

Linguagem, Mente & Ação

http://abstracta.oa.hhu.de

d|u|p

abstracta
Linguagem, Mente & Ação

ISSN 1807-9792

Volume 8, Number 2
2015

Contents

abstracta Linguagem, Mente & Ação ISSN 1807-9792 Volume 8, Number 2, pp. 3–16, 2015

On the Metaphysics of Mental Causation

Dwayne Moore[1] and Neil Campbell[2]

[1]Philosophy Dept. Trent University
1755 West Bank Drive, Peterborough, Ontario K9L 1Z6, Canada
dwaynemoore@trentu.ca

[2]Philosophy Dept. Wilfrid Laurier University
75 University Avenue W., Waterloo, ON N2L 3C5, Canada
necampbe@wlu.ca

Abstract

In a series of recent papers, Cynthia MacDonald and Graham MacDonald offer a resolution to the twin problems of mental causation and mental causal relevance. They argue that the problem of mental causation is soluble via token monism – mental events are causally efficacious physical events. At the same time, the problem of mental causal relevance is solved by combining this causally efficacious mental property instance with the systematic co-variation between distinct mental properties of the cause and the action-theoretic properties of the effect in question. In this paper we argue that the solution offered by MacDonald and MacDonald faces significant difficulties in resolving both of the twin problems of mental causation and mental causal relevance.

In a series of recent papers (1986; 1989; 2006; 2007; 2007; 2010), Cynthia MacDonald and Graham MacDonald offer a resolution to the twin problems of mental causation and mental causal relevance. They argue that the problem of mental causation is soluble via token monism – mental events are causally efficacious physical events. At the same time, the problem of mental causal relevance is solved by combining this causally efficacious mental property instance with the systematic co-variation between distinct mental properties of the cause and the action-theoretic properties of the effect in question. In other words, their model is an instance of the familiar strategy of yoking token monism with property dualism. MacDonald and MacDonald, however, endorse this nonreductive monism from within a property exemplification account of events. In this paper we argue that nonreductive monism, when yoked with the property exemplification account, faces significant difficulties in resolving the twin problems of mental causation and mental causal relevance.

This paper is divided into four sections. First, we outline the position of MacDonald and MacDonald in some detail (§1). We then point to a number of difficulties that their model of mental causation faces, all of which resolve around their attempt to combine the property exemplification account of events with a co-instantiation thesis of property exemplification (§2). Next, we show that MacDonald and MacDonald also have problems securing mental causal relevance of mental property instances (§3), each of which arise from difficulties associated with the co-instantiation thesis implying either the simplicity or complexity of events. More specifically, if events are complex, as many take them to be, there is too little causal relevance, and mental causal relevance fails. But, if events are simple, which is difficult to establish, there is too much causal relevance, and every property instantiated as the event is causally relevant. Finally, we show the difficulties that MacDonald and MacDonald have in establishing the causal

relevance of mental properties (§4). Namely, MacDonald and MacDonald preserve the causal relevance of mental properties by employing a dubious formulation of the exclusion principle, while a more appropriate exclusion principle falsifies their solution.

1.

The problem of mental causation states that every mental event lacks causal efficacy in generating its effects if a physical event is causally sufficient for that effect. MacDonald and MacDonald solve this problem by endorsing token monism – the view that mental events are causally potent physical events: "We can take it that the putatively two events are really the same event" (G. MacDonald, 2007, 242). As Graham MacDonald notes, this is "essentially the Davidsonian solution" (G. MacDonald, 2007, 242), whereby mental causation is secured via token monism.

There are, however, important differences between Davidson's approach and the one proposed by MacDonald and MacDonald, which have their origins in their differing views of events. Whereas Davidson endorses token monism within a coarse-grained model of events, MacDonald and MacDonald advocate for token monism within the fine-grained property exemplification model of events. Davidson takes events to be entities with numerous properties, or capable of supporting multiple true descriptions. However, given his nominalism about properties he did not think of events as literally constituted by clusters of properties (or their instances). Property exemplification accounts, by contrast, involve a far more robust ontology that treats properties as more than ways of describing events; they are metaphysically constitutive of events. Such views typically construe an event as the instantiation of a property in an object at a time (Kim, 1993, 33-52; Lombard, 1986). Since these three parts constitute or make up the event, they are called the "constitutive object," "constitutive property" and "constitutive time" of the given event, respectively.

According to the property exemplification theory, although an event is the exemplification by an *object* of a property at a time, *events* themselves can also exemplify properties. MacDonald and MacDonald call these further properties "characterizing" properties (MacDonald and MacDonald, 2006, 560), while Jaegwon Kim, another proponent of the property exemplification account, prefers to say that events have intrinsic descriptions that highlight the constitutive property exemplified by the constitutive *object*, and extrinsic descriptions, that pick out the event by properties the *event* exemplifies (Kim, 1993, 42-43). On the Davidsonian view there is no meaningful contrast between so-called "characterizing" and "constitutive properties" since all properties that can be truly ascribed to an event are a matter of description rather than the metaphysics of the event itself.

MacDonald and MacDonald's property exemplification model delivers token monism as follows. According to the property exemplification account, properties are universals which are instantiated as the thing (i. e., object/event) that has it:

> On the universalist conception presumed by the PEA [property exemplification account], things exemplify properties, and a thing just is (i. e. is identical with) an instance of each property that it has. Thus, an event exemplifies its properties, and it is (= is identical with) an instance of each property it has. (MacDonald and MacDonald, 2007, 14; see also MacDonald and MacDonald, 2006, 562; MacDonald, 2005, 197)

In the same way that a four-legged, yellow, ferocious, … , male, furry lion is one particular object which is the compresent instantiation of all of these properties, so the determined, barefooted, lengthy, … , winding, six mile per hour running is a single event, where this event just is an instance of all of these properties. This suggests that one instance, or event, can

be an instantiation of many distinct properties. MacDonald and MacDonald call this the co-instantiation thesis:

> *Co-Instantiation Thesis*: Two or more properties of an event can be co-instantiated in a single instance, that is, there can be just one instance of distinct properties (MacDonald and MacDonald, 2006, 562).

As the property of being a running and the property of being a movement are both instantiated as one event, which is John's running at noon, so a mental property and a distinct physical property can be co-instantiated as one causally efficacious event, thereby securing mental causation while also preserving property dualism.

This may solve the problem of mental causation, but a number of authors suggest that token monistic solutions of this sort gives rise to the related problem of mental causal relevance (Honderich, 1984, 86; Honderich, 1988, 15; Horgan, 1989; Kim, 1993b, 21; McLaughlin, 1993; Sosa, 1984, 277-278). MacDonald and MacDonald articulate the problem of mental causal relevance by first of all introducing the following four theses:

1. *The Principle of the Causal Relevance of Physical Properties*: Physical properties of physical events are causally relevant to the physical effects those events bring about (MacDonald and Macdonald, 2006, 544).
2. *The Principle of the Causal Relevance of Mental Properties*: Mental properties of physical events are causally relevant to some of the mental and physical effects those events bring about (MacDonald and Macdonald, 2006, 544).
3. *Exclusion*: If a property, *P*, of a cause, *c*, is causally sufficient for an effect, *e*, then no other property, *Q*, distinct from and independent of *P*, is causally relevant for *e* (MacDonald and Macdonald, 2006, 544).
4. *Closure*: If a physical event or phenomenon has any cause, it has a sufficient physical cause, whose physical properties are causally sufficient for its effect (MacDonald and Macdonald, 2006, 546).

Assuming that mental properties are distinct from physical properties, it seems that theses (1), (3) and (4) imply the falsity of (2). Suppose my being in pain causes me to utter a colourful metaphor. It seems, in accordance with the causal relevance of mental properties, as though the mental property (pain) is causally relevant to my utterance. However, since my utterance is a physical event, according to the causal relevance of physical properties and closure, there is a physical property that is sufficient for its occurrence. Exclusion tells us that no other distinct and independent property can be causally relevant to my utterance, in which case it seems as though the mental property is excluded from being causally relevant to my utterance.

MacDonald and MacDonald's solution to the problem of mental causal relevance has two components. It has a causal component pertaining to property instances, and a nomological/explanatory component pertaining to the abstract, universal properties themselves. With respect to the causal component, MacDonald and MacDonald argue the mental property instance of the event is identical to the physical property instance of the event, in virtue of the fact that the mental property instance is the event that is the physical property instance. Thus, mental property instances are causally efficacious, as they are causally efficacious events:

> " [...] exemplifications of mental properties of mental events are identical with exemplifications of physical properties of physical events (since each mental event is identical with a physical event). So, to say that a mental property of a physical event is causally relevant (that is, that a mental event is causally efficacious *qua* mental) is to say *at least* that an exemplification of that

> property, that is, that event, is causally efficacious in bringing about an effect of that event." (MacDonald and MacDonald, 2006, 562, see also MacDonald and MacDonald, 2006, 566; and MacDonald and MacDonald, 2006, 541)

In other words, since an instance of a mental property simply is the mental/physical event that has it (in accordance with their universalist understanding documented above), and an instance of the physical property simply is that same mental/physical event which also has it, the instance of the mental property *is* the instance of the physical property. Thus, the mental property instance is (i. e., the mental event) is causally efficacious, and this efficacy of the mental property instance is a necessary but insufficient condition for mental causal relevance.

While the causal efficacy of the mental instance is necessary, mental causal relevance also requires, "systematical property dependence or co-variation" (MacDonald and MacDonald, 2006, 574) between the mental properties of the cause and the properties of the effect as well. If mental properties systematically co-vary with the occurrence of the action-theoretic properties of the effect, then it is reasonable to suppose that these mental properties are relevant to the occurrence of the action-theoretic properties of the effect. This systematic co-variation arises because mental properties strongly supervene on physical properties (MacDonald and MacDonald, 2006, 565), so the mental properties of the cause necessarily precede the action-theoretic properties of the effect. The systematic covariance between the mental properties of the cause and the action-theoretic properties of the effect, combined with the causal efficacy of the mental property instance (i. e., the event), generates mental causal relevance.

It may be objected that this model of mental causal relevance fails on account of the fact that the physical property is causally sufficient for the effect, so the distinct mental property cannot be causally relevant for the effect. MacDonald and MacDonald solve this problem by pointing out that exclusion pressures only arise if the mental properties are, "distinct from and independent of" (MacDonald and MacDonald, 2006, 544) physical properties. Given strong supervenience, mental properties are, "distinct from but not independent of physical ones" (MacDonald and MacDonald, 2006, 566), so there is no exclusionary tension between the causally relevant physical properties and the causally relevant mental properties.

In these ways, MacDonald and MacDonald claim to secure (1) the causal efficacy of mental events via event identity; (2) the causal relevance of mental properties of mental events via the efficacy of the mental instance (which is implied by the event identity) combined with the systematic co-variation of mental properties of the cause with the action-theoretic properties of the effect. Unfortunately, there are problems with these solutions, which we will now discuss in turn.

2.

As outlined above, Cynthia MacDonald and Graham MacDonald secure the causal efficacy of mental *events* by advocating a form of the token-identity thesis. While this kind of approach is generally regarded as successful, and few have objected to it – even in its original Davidsonian form – we suspect that there is a tension between adopting this approach while simultaneously advocating the property exemplification account of events. This is because the criterion for event identity on the property exemplification model suggests one event cannot be the instantiation of two different properties. Here is the condition governing event identity on the MacDonaldian (and Kimian) model:

> *Identity Condition*: Event [x, P, t] is identical with event [y, Q, t'] if and only if the object x is identical with the object y, the property P is identical with

> the property *Q*, and the time *t* is identical with the time *t'* (MacDonald and MacDonald, 2006, 557; see also Kim, 1993, 9).

According to this condition, mental events cannot be physical events ($m \neq p$) if, among other things, mental properties are not physical properties ($M \neq P$). Kim uses this identity condition to argue that if mental properties are not physical properties, mental instances cannot be physical instances (Kim, 2005, 42; see also, Marras and Yli-Vakkuri, 2008, 117). Call this "the single-instantiation thesis" (in contrast to the MacDonaldian co-instantiation thesis). According to the single-instantiation thesis the instantiation relation is exceedingly tight, so two different properties, when instantiated, must yield two different property instances. Thus, where events are indicated by lower-case variables and properties by upper-case variables, if *p* is an instance of *P*, *m* is an instance of *M*, and *h* is an instance of *H*, it will not be the case that *p* is an instance of *P* and *M* and *H*, as the co-instantiation thesis allows unless the properties *P*, *M*, and *H* are really the same property. If Kim's single-instantiation thesis is true, then the MacDonaldian co-instantiation thesis fails, and the mental/physical event identity cannot go through (at least, not without the accompanying property-identity). For their own part, a number of authors agree with Kim's single-instantiation thesis (Ehring, 1996, 462-463; Gibb, 2004, 469; Lowe, 1989, 113; Menzies and List, 2010, 110; Whittle, 2007, 64-65). The single-instantiation thesis that is plausibly derived from the property exemplification account undermines MacDonald and MacDonald's solution to the mental causation problem.

Graham MacDonald, however, argues that Kim's single-instantiation thesis is "blocked by compelling reasons" (G. MacDonald, 2007, 243). According to him, the most plausible interpretation of the familiar determinable/determinate relation is to posit a co-instantiation thesis:

> The colour-property [red] is a different property from the property [light red], given that it can be present when [light red] is not. But when [light red] is instanced, it is clear that [red] is instanced as well, and it is natural to assume that these are not two separate instances (G. MacDonald, 2007, 243).

To switch the illustration somewhat, a dark red token may be a red token as well. The alternative to this natural assumption is to countenance the, "multiplication of many instances whenever a determinate property is instanced" (G. MacDonald, 2007, 243). In other words, if a dark red token is distinct from a red token, then both a dark red token and a red token are instantiated in an object at a time, which involves an excessive proliferation of events.

Unfortunately, it is not natural to assume the MacDonaldian co-instantiation thesis provides the best interpretation of this situation. Consider, for example, some of the other properties that are co-instantiated with the dark red token. This dark red token is also an instance of being coloured, being visible, being the colour of Mars, being the colour of Mars and ketchup, being the colour of Mars and ketchup or grass, *et cetera*. This token will be the instance of an endless number of properties that stand in some sort of dependency relation. Imagine that a woman walks into a store and buys a dark red shirt. In addition, imagine that this woman hates the look of ketchup and has never seen the planet Mars. According to the co-instantiation thesis, the instance of the property of 'being the colour of Mars and Ketchup' will be as causally efficacious in producing her purchase as the instance of the property of 'being dark red'. In contrast, the single-instantiation thesis, simply asserts that the dark red token is an instance of dark red, period. Therefore, the dark red shirt is purchased because of the dark redness of the shirt alone.

The MacDonaldian co-instantiation thesis suffers from a further complication, which is similar to an objection Sophie Gibb levels against tropist versions of the co-instantiation thesis (Gibb, 2004, 470ff). Namely, every instance *m*, or exemplification *m* of a property *M*, if it genuinely is an example of property *M*, will bear a strong affinity to property *M*. We can syntactically represent this affinity between an instance and the property through the use of the lower case and upper case versions of the same letter. Thus, instance *m*1 will strictly resemble *M*, and *m*2 will strictly resemble *M*, thereby ensuring that *m*1 strictly resembles *m*2. Coupling the co-instantiation thesis with multiple realizability, however, it is clear that *m*1 can be a physical token *px*1 of *Px*, while *m*2 can be a physical token *pz*1 of type *Pz*. The property *Px* is not identical to the property *Pz*, and therefore they do not share a close essential affinity. This being the case, the token *px*1 of *Px* and *pz*1 of *Pz* need not resemble one another. The result is that the token *m*1/*px*1 both must and must not resemble the token *m*2/*pz*1. Something is amiss. This problem arises in large part because of the co-instantiation thesis. After all, if the *m*1 token is not the *px*1 token, then the simultaneous resemblance and distinction among instances that is so troublesome does not arise. For example, on the Kimian single-instantiation thesis, the instance *m*1 of *M* and *m*2 of *M* resemble, while the distinct instances *px*1 and *pz*1 do not resemble.

Not only does the MacDonaldian co-instantiation thesis lead to these counterintuitive results, but the Kimian single-instantiation thesis provides a more natural interpretation of causal efficacy. To use a different example from Graham MacDonald, Sally, who weighs 115 kilograms, steps on the scale and this causes the arrow to point at the 115 kg mark (G. MacDonald, 2007, 245-246). On the co-instantiation thesis, the 115 kg stepping is also a 'greater than 100kg' stepping, a 'less than 116kg' stepping, a nervous stepping, an afternoon stepping, *et cetera*. On the co-instantiation thesis all of these properties are instanced as one event, so they all have the same amount of causal efficacy. This includes the seemingly irrelevant properties instanced in the event, such as the nervousness of the stepping, and it also includes the instantiations in the event that stand in dependency relations but seem to be unlikely causes of the effect, such as the stepping 'being less that 424 kilograms' (or, 'having weight' in general), being the cause of this scale reading 115kg.

In contrast, the single-instantiation thesis suggests that the constitutive property of an event reveals the fundamental essence of it, while characterizing properties are accidental, extrinsic and/or derivative properties of the event. For example, the constitutive property of the causal event is a 115 kilogram stepping, which causes the 115kg reading on the scale. This event has various characterizing properties as well, such as the event's being a nervous stepping, the event's being a 'greater than 100kg' stepping, and the event's 'being of the same weight as Frank's' stepping. The 115kg stepping is constitutive in the sense that it bears a nomological and explanatory relation to the 115kg scale reading. No such law exists between 'greater than100kg' steps and 115kg readings, so the event is not as fundamentally a 'greater than100kg stepping' as it is a 115kg stepping. The 'greater than 100kg' stepping is also an extrinsic property of the event since, if the event is taken in isolation it is not related to other steppings or weights, so it may not have the properties of being 'greater than 100kg' or of 'being the same weight as Frank's'. These characterizing properties are also derivative properties of the cause in the sense that the event is only a 'greater than100kg' stepping because it is actually a 115kg stepping. The converse does not hold; the event is not only a 115kg stepping because it is actually a 'greater than100kg' stepping. This asymmetry suggests the fact that this event is more fundamentally a 115kg stepping, as the single-instantiation thesis suggests.

This discussion hints at a resolution to the second objection lodged against the single-instantiation thesis. Recall that Graham MacDonald argued that a rejection of the co-instantiation thesis results in a proliferation of events. It is possible to reject the co-instantiation thesis without a resulting proliferation of events in a number of ways. First, one could simply argue that determinable properties such as those mentioned above do not exist, resulting in no proliferation of instances (Gillett and Rives, 2005). Or, less radically, according to the single-instantiation version of the property exemplification model discussed above, an event is the instantiation of one constitutive property, while this event has a number of other characterizing properties. This does not imply that there is a distinct event for every instantiation; rather, it implies that there is one event, and this event has a number of characterizing properties. For example, Brutus' stabbing is one event that has the characterizing property of being a killing. On this model, there is no proliferation of events, for there remains only one event. There are, however, numerous characterizing properties of the one event. MacDonald and MacDonald, however, agree that events have a number of characterizing properties, so it is not clear how this single-instantiation thesis is excessive. The single instantiation thesis, therefore, stands clear of objection, and, for a number of reasons, is the preferable interpretation of the property exemplification model of events. But, if the single-instantiation thesis is true, it is not possible for the distinct mental property to be instantiated as the same event as the instantiation of the causally efficacious physical property, which results in the loss of mental causation.

3.

In the preceding section we argued that the problem of mental causation cannot be solved by appealing to MacDonald and MacDonald's version of the co-instantiation thesis. In the next two sections we argue that the further problem of mental causal relevance cannot be solved using the tactics suggested by MacDonald and MacDonald either. Recall that the problem of mental causal relevance states that despite the mental/physical event identity, it is plausible that the event causes in virtue of its physical properties, thereby excluding the causal relevance of the mental properties. MacDonald and MacDonald solve this problem by suggesting both that (1) the mental property instantiated as the event is causally efficacious, and that (2) the mental property systematically co-varies with the action-theoretic properties of the behavioural effect, so the mental property is causally relevant to the effect.

In this section, we establish a number of problems with the first criterion of this solution to the problem of mental causal relevance. Notice, first of all, that this criterion can be interpreted as saying either that events are ontologically simple or that events are ontologically complex. Serious difficulties arise on both interpretations, and we will begin with the former. According to this interpretation, MacDonald and MacDonald can be read as stating that an event is (identical with) an instance of many properties, which means that there is only one instance (i. e., the event) and this event cannot be broken into components. Thus, it makes no sense to suggest that there is a mental property instantiated in the event and a distinct physical property instantiated as another component or constituent of the event. Rather, there is only a simple event that cannot be divided in these ways.

There are a number of difficulties with this model. First of all, many, if not all, of the critics mentioned above who delineate the problem of mental causal relevance reject the simplicity of events. Therefore, the solution offered by MacDonald and MacDonald fails to solve the problem formulated by these critics. To see this problem in detail, consider Ted Honderich's charge that mental properties of causes are excluded by the physical properties of causes. When Honderich argues that mental properties of events lack causal relevance, he is not considering

events as ontologically simple. Rather, Honderich argues that events are ontologically complex entities in the sense that they have a number of components or aspects that can be thought of as particularized properties, concretized properties, or property instances. Examples of particularized properties include 'this hardness of this brick', 'this height of this building', and 'this redness of this brick'. The last example indicates that the same object or event (i. e., this brick), can instantiate two different particularized properties (i. e., this hardness of this brick and this brick redness of this brick). Although these two particularized properties are instantiated in the same brick, this hardness of this brick is a distinct component, aspect, or particularized property, of this brick from this redness of this brick. As Honderich explains:

> It is not the age or the sheen of the teapot that is flattening the napkin, but its weight. It is not the weight of the door but its colour that makes it reflect the light. The most natural answer to the question of what caused something, then is a property of an ordinary thing. What needs to be resisted immediately, however, is that what is in question is a general property, a universal. It is not the general property of weighing a pound, which is other or more than this teapot weighing a pound, which is flattening the napkin. That general property will exist if the weight of the teapot is changed and the napkin isn't flattened . . . . We come to the idea, then, that what is flattening the napkin is *this teapot's weight,* an individual property of this teapot ... . It is not *all* of the teapot, or any individual property of it other than its weighing a pound, that is an instance of the general property of weighing a pound ... causes strictly speaking are individual properties. (Honderich, 1982, 292; see also Honderich, 1984, 86 and Honderich, 1988, 15; Horgan, 1989; McLaughlin, 1993)

With this metaphysical backdrop in mind, it is clear that when Honderich suggests that a brick has the property of being heavy and a distinct property of being red, he is suggesting that this redness of this brick is a distinct component of the brick from this hardness of this brick. And so, it is perfectly coherent to wonder whether this brick broke this glass in virtue of this redness of this brick or in virtue of this hardness of this brick. With regards to the case of mental causation, this model indicates that the mental/physical event has a particular mental property instance and a distinct physical property instance. Thus, it is reasonable to ask which aspect of the complex event caused the effect. Since the physical component of the complex event is sufficient to cause the effect, the mental component of this complex event is irrelevant and hence excludable.

MacDonald and MacDonald reject the formulation of the causal relevance problem that Honderich suggests (Macdonald and Macdonald, 1991, 25-29). Although events are the compresent instantiation of many properties for MacDonald and MacDonald, events are still simple in the sense that events lack distinct components by which to raise questions about which particular aspect of the event is causally responsible for the effect. But, this move evades rather than addresses the problems raised by Honderich and others. That is, MacDonald and MacDonald suggest that events are simple, but the critics continue to argue that events cause in virtue of one aspect of the event.

Not only do a number of critics argue that the problem of causal relevance is a problem pertaining to which aspect of a complex event is causally relevant, but there is reason to prefer the view that events are complex. First of all, it is intuitively plausible that objects/events are complex in the sense that they have differing components as ontological constituents. While

it is true that the earth is the compresent instantiation of many properties, it also appears to be the case the earth has a variety of components to it. In other words, this wetness of this earth is not identical to this weight of this earth.[1] While walking around the earth, it appears clear that the earth has these distinct components to it. Or, to borrow an example from Honderich, it is perfectly reasonable to suspect that this weight of this pear is a distinct component of this pear from this greenness of this pear. If objects/events can be analyzed in this manner, then the question about which component of the complex event is causally relevant remains poignant.

Secondly, there are a number of reasons to think that simple events cannot be accommodated within the property exemplification model that MacDonald and MacDonald deploy. According to the property exemplification model, events are complex in the sense that they are constituted by, at least, a constitutive object, constitutive property and constitutive time (cp. Ehring, 1996; Robb, 1997). It may be possible to object to the view that these three constituents imply that events are complex on the grounds that an event is the specific structure of an object's having a property at a time, it is not an object, a property and a time (MacDonald and MacDonald, 2006, 559). Even if this response is viable, the property exemplification account implies that events have components in another way as well. Namely, every event consists of a constitutive property and many characterizing properties. For example, John's run is a run, and it is winding, determined, long and occurred outside of Boston. It is natural to suppose that this run has a number of components, whereby this run's windingness is not identical to this run's length, which is not identical to this run's location. This is especially so if constitutive properties are instantiated in objects while characterizing properties are instantiated in events. After all, clearly instances can only be identical if, among other things, they are co-located in the same entity. The constitutive property is instantiated in an object, which is distinct from the event, so these instances cannot be identical. Since these instances are not identical, it is not viable to suppose that they are co-instantiated as a simple event.

Beyond these general concerns related to the property exemplification account, the MacDonaldian model of simple events also renders every property instanced as the event causal. To borrow another example from Ted Honderich, John's slipper just is the instantiation of many properties, such that John's fleecy, mauvish, comfortable ... light, stinky, slipper is on his foot. On MacDonald and MacDonald's model, we attain causation in virtue of a certain property due to the fact that the property just is instantiated as the slipper. Thus, the slipper warms in virtue of its fleecyness because the fleecyness just is instantiated as this object. At the same time, however, the slipper can be said to warm in virtue of the mauvishness because the mauvishness just is instantiated as this object. Similarly, the slipper can be said to warm in virtue of the foot-like odor of the slippers, for the foot-like odor just is instantiated as this object. One hopes that causation in virtue of the mental instance is not equivalent to the causation in virtue of the mauvish instance found in this example; surely mental instance causation is not secured because every property instantiated as the event is efficacious! A number of critics have leveled this 'too much efficacy' charge against MacDonald and MacDonald (Wyss, 2010, 174).[2]

[1] MacDonald and MacDonald object to this model which, in their words, invokes "property instances that mediate between particular and ... universal" (MacDonald and MacDonald, 2006, 562; see also MacDonald, 2005, 212), as problematically tropist. If, however, it is intuitively plausible that objects have components (i. e., this wetness of this earth is distinct from this rockiness of this earth, which is distinct from this weight of this earth), then it is difficult to see why this model is problematic.

[2] MacDonald and MacDonald accept this result as an "inevitable consequence" of their theory (MacDonald and MacDonald, 2006, 563) since, "on our account all properties 'sharing' an instance that is causally efficacious are causally efficacious properties" (G. MacDonald, 2007, 245).

MacDonald and MacDonald face a final troublesome consequence. As hinted at with the slipper, when applying the MacDonaldian model to objects we see that an instance of the property, red, "just is the red bird" (MacDonald and MacDonald, 2006, 562), so an instance of the property of being alive just is the alive bird, and an instance of the property of being winged just is the winged bird. And so the bird that is red just is the bird that is alive. It is natural to wonder whether this also means that this life of this bird, is identical to this redness of this bird. Or in other words, is this example of red that this bird has identical to this example of life that this bird has? According to the model of simple events (or, in this example, simple objects), the answer seems to be that yes, this life of this bird is identical to this redness of this bird. After all, this bird is simple, it lacks distinct instances of properties. Since the property of being alive is instantiated as the bird and the property of being red is instantiated as the bird, so this life of this bird is this redness of this bird, at least in the sense that the live bird is the red bird. The problem here is that identity is transitive. Thus, if the bird is red (r is b) and the bird is winged (w is b) and the bird is alive (a is b), the implication is that the bird's redness is the bird's wingedness, which in turn is the bird's life (r is w is a) – which is absurd. Or, with respect to the crucial case of events, Susan's run is six kilometers long ($6l$ is Sr) and Susan's run is barefooted (b is Sr), so this example of six kilometer length that Susan's run exemplifies is this example of barefootedness that Susan's run exemplifies – which is absurd. This consideration, combined with the others listed above, indicates that events ought to be construed as having distinct aspects, or, in other words, that events ought to be construed as being complex.

Given these difficulties, it is worth considering whether MacDonald and MacDonald endorse the view that events are ontologically complex. MacDonald and MacDonald appear to appeal to a constitution relation rather than an identity relation at times. For example, they state, "This redness, this shape, this size, and this position, related to one another by compresence relations, together 'constitute' or comprise the cardinal sitting on the branch of the tree" (MacDonald and MacDonald, 2006, 548).[3] And, at least one critic has interpreted MacDonald and MacDonald as endorsing the view that although many properties are instantiated as the same event, this event continues to have components (Crane, 1995, 222). While complex events would avoid the aforementioned problems with simple events, it gives rise to different problems. In a recent paper, Graham MacDonald considers the possibility that a physical property P and a mental property M can have two distinct instances m_{it} and p_{it} within the same event. He rejects this possibility because it re-introduces the problem of mental causal exclusion at the level of instances:

> If we accept the move from [mental property $\neq$ physical property] to [mental property instance $\neq$ physical property instance], that will commit us to saying that a single event can exemplify two different properties by possessing two instances, one for each property. If we grant this, then we will have saved physicalism, and avoided overdetermination in the form of event-overdetermination. But that should not satisfy anybody, because what we will be left with is instance-overdetermination (G. MacDonald, 2007, 243).

[3] To be fair, they state this in the context of the trope theory, which they later reject. On other occasions, however, when discussing the relationship between substances and their properties, Cynthia MacDonald appears to endorse a constitution relation between substances and their properties as well (MacDonald, 2005, 121). She argues that a cat has the constitutive property of being a cat, and this cat is constituted by a variety of characterizing properties as well, such as the cat's blackness, and the cat's weight. She does not, however, argue that events exemplify a constitutive property in an object, while this event is constituted by its characterizing properties.

In other words, the mental/physical event will be causally efficacious, so mental causation will be preserved. However, the mental instance that is instantiated in the mental/physical event is distinct from the physical instance that is instantiated in the mental/physical event. Thus, since closure and physical causal relevance suggests that the event will cause in virtue of the physical property instance, the event will not cause in virtue of the mental property instance, and we will fail to secure this criterion for mental causal relevance. To avoid this consequence MacDonald and MacDonald need to appeal to the identity relation, stating that this mental property instance of this event is this physical property instance of this event. But, this identity relation returns them to the difficulties listed above.

4.

Even if MacDonald and MacDonald can somehow circumvent these difficulties associated with criterion (1) of causal relevance, there are still certain problems that criterion (2) of their model of mental causal relevance faces as well. MacDonald and MacDonald suggest that mental properties supervene on physical properties, and this dependency relation ensures that mental properties consistently co-vary with the action- theoretic properties of their effects. It is possible to object along exclusionary lines: the physical properties are sufficient for the effect, so the mental properties are not causally relevant after all. MacDonald and MacDonald reply that mental properties are not independent of physical properties, so it is possible for the physical properties to be sufficient for the effect, while also including the dependent mental properties. There are, however, several problems with this line of reasoning.

First, as previously noted, MacDonald and MacDonald recognize that the concern that too many properties seem to be causally relevant on their model. Seemingly, the greenness of the pear is causally relevant to the scale's pointing at the two-pound mark since the green pear causes the scale to point at the two-pound mark. Or again, the loudness of the shot is causally relevant to the victim's death since the loud shot causes the death, *et cetera*. MacDonald and MacDonald respond by pointing out that causal relevance also requires properties to systematically co-vary with the properties of the effect. Instances of greenness do not systematically co-vary with instances of two-pound readings, so there is no reason to think this instance of greenness would be causally relevant to the two-pound reading. Instances of two-poundedness, however, do co-vary with instances of two-pound readings, so this property is causally relevant. While this test may exclude a number of obviously irrelevant properties, there still remains an excess of dependent, systematically co-varying, and hence relevant, properties that withstand this test. The scale breaks when someone weighing more than 115kg steps on it. Johnson, being 170kg breaks the scale, so the property of being 170kg is relevant, but so is the property of being more than 169kg, the property of being more than 168kg, the property of being a weight more than 168kg or less than 5kg, *et cetera*. These properties are dependent upon the property of weighing 115kg, so they are causally relevant on the MacDonaldian model. Therefore, there still appears to be an excess of causally relevant properties.

Not only is this excess of causally relevant properties unwelcome, but notice as well that the strategy that MacDonald and MacDonald deploy only works if exclusion allows distinct but dependent properties to be included. While MacDonald and MacDonald contend that exclusion allows properties that are distinct from but dependent upon sufficient physical properties to be included, there is reason to think that distinct but dependent properties should be excluded once there is already one sufficient physical property relevant to the effect. As it turns out, MacDonald and MacDonald are quite forthright about the origins of their principle of Exclusion. They claim it is a variation of Kim's well-known principle of explanatory exclusion,

which states, "there can be no more than a single *complete* and *independent* explanation of any one event" (Kim, 1988, 233). Obviously, MacDonald and MacDonald have followed Kim in assuming that there is no problem of exclusion if there is a relation of dependence between the items competing for relevance. Notice, however, that Kim's exclusion principle involves relations between explanations rather than properties. MacDonald and MacDonald are concerned with properties, and it is not clear that Kim's principle can be applied to properties as well. Moreover, MacDonald and MacDonald are concerned with issues pertaining to causal relevance. Not only is this a slightly different issue from Kim's concern about explanation, but it indicates that Kim's principle of causal exclusion may be more appropriately used on this occasion. Kim's principle of causal exclusion, however, intentionally excludes distinct but dependent events/properties: "No single event can have more than one sufficient cause occurring at any given time" (Kim, 2005, 42).[4] If the problem concerns causal relevance, and causal exclusion suggests that distinct but dependent causes are excluded, then MacDonald and MacDonald's suggestion that mental causal relevance is preserved due to the fact that mental properties are dependent upon sufficient physical properties is false. It seems, therefore, that MacDonald and MacDonald employ a dubious formulation of the exclusion principle that stacks the deck in their favour while a more appropriate exclusion principle would falsify their solution.

Fortunately, this is not merely a matter of personal preference over one's preferred articulation of the exclusion principle. Rather, there is reason to think that distinct but dependent properties ought to be excluded. It is worth noting that Jaegwon Kim introduced the principle of explanatory exclusion at a time when he also argued that distinct but dependent causes of the same event were not to be excluded either (Kim, 1993, 106-107). At this time, Kim argued that even though there is a sufficient physical cause for fear, pain can still be included as a cause for the fear by virtue of the fact that pain strongly supervenes upon, and is dependent upon, the physical cause of the fear. More recently, however, Kim argues that these distinct but dependent mental causes of the effect ought to be excluded. His reasoning is that the physical cause is sufficient on its own, so these mental causes are like shadows which, though dependent, still only come along for the ride (Kim, 1998, 37; Kim 2005, 62; Kim). According to Kim's intellectual progression, distinct but dependent entities may seem includable, but on closer analysis they ought to be excluded as they are not, ultimately, necessary. Thus, if Johnson breaks the scale in virtue of his 170kg stepping, and this is the sufficient cause of the scale breaking, the additional dependent property of 'being more than 169kg stepping' is in principle unnecessary, so it can be excluded. Or, more to the point, the physical property is sufficient, so the mental property, though dependent, is not necessary or relevant.

In summary, MacDonald and MacDonald's attempt at combining token monism, property dualism and the property exemplification model of events is fraught with difficulties. The lesson seems to be that we can retain token monism and property dualism while rejecting the property exemplification account, as Davidson suggests. Or, we can retain the property exemplification account while rejecting either property dualism or token monism, as Kim suggests. The combination of all three simply cannot succeed.

[4]The causal exclusion principle states that no event can have more than a sufficient cause, where this cause is taken to be an event. As discussed above, Kim endorses the condition on event identity outlined in the property exemplification account. Thus, if event *a* has a different constitutive property than event *b*, event *a* $\neq$ event *b*. As a consequence, if only one causal event is allowed, only one property instance is allowed. Thus, although the principle of causal exclusion is framed in terms of events, it is also true that mental properties are excluded if the physical property instance is sufficient.

References

Crane, T. (1995) 'The Mental Causation Debate', *Proceedings of the Aristotelian Society,* **69**, 211-253.

Davidson, D. (1993), Thinking Causes, *in* J. Heil and A. Mele, eds., 'Mental Causation', Clarendon Press, Oxford, pp. 3-17.

Davidson, D. (1995), 'Laws and Cause', *Dialectica* **49**, 263-279.

Ehring, D. (1996), 'Mental Causation, Determinables, and Property Instances', *Nous* **30**, 461-480.

Gibb, S. (2004), 'The Problem of Mental Causation and the Nature of Properties', *Australasian Journal of Philosophy* **82**, 464-475.

Gillett, C. and Rives, B. (2005), 'The Nonexistence of Determinables', *Nous* **39**, 483–504.

Honderich, T. (1982), 'Causes and If p, Even If x, Still q', *Philosophy* **57**, 291-317.

Honderich, T. (1984), 'Smith and the Champion of Mauve', *Analysis* **44**, 86-89.

Honderich, T. (1988), *A Theory of Determinism*, Clarendon Press, Oxford.

Horgan, T. (1989), 'Mental Causation', *Philosophical Perspectives* **3**, 47-76.

Kim, J. (1988), 'Explanatory Realism, Causal Realism, and Explanatory Exclusion', *Midwest Studies in Philosophy* **12**, 225-239.

Kim, J. (1993), *Supervenience and Mind,* Cambridge University Press, Cambridge.

Kim, J. (1993b), Can Supervenience and "Non-Strict Laws" Save Anomalous Monism, *in* J. Heil and A. Mele, eds., 'Mental Causation', Clarendon Press, Oxford, pp. 18-26.

Kim, J. (2005), *Physicalism, or Something Near Enough,* Princeton University Press, Princeton.

Lombard, L. (1986), *Events: A Metaphysical Study*, Routledge & Kegan Paul, New York.

Lowe, E. (1989), *Kinds of Being*, Basil Blackwell, Oxford.

MacDonald, C., and MacDonald, G. (1986), 'Mental Causes and Explanation of Action', *The Philosophical Quarterly* **36**, 145-158.

MacDonald, C. (1989), *Mind-Body Identity Theories,* Routledge, London.

MacDonald, C. and MacDonald, G. (1991), 'Mental Causation and Non-Reductive Monism', *Analysis* **91**, 23-32.

MacDonald, C. (2005), *Varieties of Things*, Blackwell, Cambridge.

MacDonald, C. and MacDonald, G. (2006), 'The Metaphysics of Mental Causation', *Journal of Philosophy* **103**, 539-576.

MacDonald, C. (2007), 'Physicalism or Something Near Enough Review', *Philosophical Books* **48 (2)**, 155-161.

MacDonald, G. (2007), 'Emergence and Causal Powers', *Erkenntnis* **67**, 239-253.

MacDonald, C. and MacDonald, G. (2007), Beyond Program Explanation, *in* G. Brennan, R. Goodin, M. Smith, eds., 'Common Minds: Essays in Honour of Philip Pettit', Oxford University Press, Oxford, pp. 1-27.

MacDonald, C. and MacDonald, G. (2010), Emergence and Downward Causation, *in* C. MacDonald and G. MacDonald, eds., 'Emergence in Mind', Oxford University Press, Oxford, pp. 139-169.

Marras, A. and Yli-Vakkuri, J. (2008), The Supervenience Argument, *in* S. Gozzano and F. Orilia, eds., 'Tropes, Universals and the Philosophy of Mind', Ontos Verlag, Frankfurt.

McLaughlin, B. (1993), On Davidson's Response to the Charge of Epiphenomenalism, *in* J. Heil and A. Mele, eds., 'Mental Causation', Oxford University Press, Oxford, pp. 27-40.

Menzies, P. and List, C. (2010), The Causal Autonomy of the Special Sciences, *in* C. MacDonald and G. MacDonald, eds., 'Emergence in Mind', Oxford University Press, Oxford.

Robb, D. (1997), 'The Properties of Mental Causation', *Philosophical Quarterly* **47**, 178-195.

Sosa, E. (1984), 'Mind-Body Interaction and Supervenient Causation', *Midwest Studies in Philosophy* **9**, 271-281.

Whittle, A. (2007), 'The Co-Instantiation Thesis', *Australasian Journal of Philosophy* **85**, 61-79.

Wyss, P. (2010), Identity With a Difference, *in* C. MacDonald and G. MacDonald, eds., 'Emergence in Mind', Oxford University Press, Oxford, pp. 169-179.

abstracta Linguagem, Mente & Ação ISSN 1807-9792 Volume 8, Number 2, pp. 17–30, 2015

Relativizing the Opposition between Content and State Nonconceptualism

Roberto Horácio de Sá Pereira

University of Rio de Janeiro/UFR
Department of Philosophy
robertohsp@gmail.com.br

Abstract

Content nonconceptualism and State conceptualism are motivated by different readings of what I want to call here Bermúdez's conditions on content-attribution (2007). In one reading, what is required is a neo-Fregean content to solve problems of cognitive significance at the nonconceptual level (Toribio, 2008; Duhau, 2011). In the other reading, what is required is a neo-Russellian or possible-world content to account for how conspecifics join attention and cooperate, contemplating the same things from different perspectives in the same perceptual field. The solution to this apparent contradiction is the rejection of the real content view and the adherence to what I call here Content-pragmatism: there is no such thing as *the* content of experience. According to content-pragmatism, "proposition" is not as real as a mental state, but rather it is a term of art that semanticists use, as a matter of theoretical convenience, to classify mental states. What follows from Content-pragmatism is Content-pluralism: there are so many contents that are required to meet Bermúdez's condition on content-attribution. Because both criticism and the defense of State conceptualism overlook the real scope of Bermúdez's condition on content-attribution, they are ineffective. In this paper, I will argue that the opposition between State and Content nonconceptualism is a real one, but only a relative one, that is, relative to the opposite constraints to be met. If we want to solve the problems of cognitive significance, the best we can do is to let the content of experience be modeled as neo-Fregean content, namely, a compound of nonconceptual modes of presentation of objects and properties. In this case, Content-nonconceptualism prevails. In contrast, if we want to account for how conspecifics join attention to the same entities in the same perceptual field from different perspectives, the best we can do is let the content of experience be modeled either as a structured Russellian proposition or as a function from possible worlds to truth-values. In this case, State-nonconceptualism prevails.

Introduction

Nonconceptualism can be traced back to British Empiricism. Hume famously holds that impressions were *prior* to concepts in order of perceptual processing and in order of acquisition. In the same vein, Kant claims that sensible intuition is *prior to* and independent of concepts and thoughts both in order of perceptual processing and in order of acquisition. According to Kant's famous dictum, without concepts, sensible intuitions are blind, and conversely, without sensible intuitions, concepts are empty. In contemporary terms, the key idea behind nonconceptualism is that *some mental states* can represent the world even though the bearer of those mental states need not possess the concepts required to specify correctly what those states represent, their so-called representational content.

This basic idea has been developed in different ways and applied to different kinds of mental states according to many contemporary philosophers. However, not all of these developments

and applications are consistent with each other (Byrne, 2005; Crowther, 2006; Heck, 2000, 2007; Speaks, 2005). The main discrepancy I want to focus on in this paper is between what Heck has called the *state view* and the *content view* (2000). According to the *state view*, nonconceptualism is characterized in terms of *kinds of states*: nonconceptualism is a property of mental states, that is, a view about the relation between the subject undergoing a mental state and the content of that state. A mental state is state-nonconceptual when it is a concept-independent state. Conversely, a mental state is state-conceptual when the subject cannot be in the mental state in question without possessing the concepts involved in the correct specification of its contents:

> State-Nonconcetualism: For any perceptual state *PS* with representational content *C*, *PS* is nonconceptual if any subject *S* need not possess the concepts required for the correct characterization of *C*.

One of the key features of State-nonconceptualism is that quite different mental states, such as experiences and propositional attitudes, might share the same content. Now, when we look back to the philosophical tradition and even to the recent contemporary debate, it seems undeniable that the main purpose of introducing the notion of nonconceptual content in the literature is just to identify a form of representation or mental state (rather than a form of representational content) that is prior to and independent of concept. For example, when Hume and Kant speak about impressions or sensible intuitions that are prior to and independent of concepts, what they have in mind are perceptual states rather than the contents of experiences.

Even so, according to Content-nonconceptualism, nonconceptualism is better characterized in terms of *the kind of content* that experiences possess, as opposed to the content of beliefs and other propositional attitudes. A mental state is content-nonconceptual when the representational content of the state is of a particular type, namely, it is not composed of concepts. Conversely, a mental state is content-conceptual when it is a structured complex compounded of concepts. Now, while according to State-nonconceptualism, experiences and propositional attitudes might share the same content even when the subject is in quite different kinds of mental states, according to Content-nonconceptualism they could not possibly share the same content. Content-nonconceptualism can be couched as follows:

> Content-Nonconseptualism: For any perceptual state *PS* with representational content *C*, *PS* is nonconceptual *if C* is not a structured complex compounded of concepts (Fregean senses).

According to Speak (2005), however, most arguments in favor of Content-nonconceptualism only support State-nonconceptualism. Take, for instance, the so-called fineness of grain argument, based on the well-known idea that our experiences outstrip any conceptual abilities. The problem with this type of argument, Speak claims, is that even if we grant that the content of experience is far finer-grained than the content of corresponding beliefs, that argument itself does not support Content-nonconceptualism. The only conclusion that we can draw from this is that we do not possess the concepts required for the correct specification of everything we can and do experience.

Still, State-nonconceptualism recently has come under attack. Bermúdez (2007: 67) disregards State-nonconceptualism by arguing that it is "unmotivated and fails to address the issues that the theory of nonconceptual content is intended to address". In the same vein, Toribio (2008: 351) goes further by arguing that without assuming that State-nonconceptualism entails Content-nonconceptualism. State-nonconceptualism is untenable, since it leaves content-

attribution unsupported. According to the criticisms of Bermúdez and Toribio, either State-nonconceptualism entails Content-nonconceptualism or is inconsistent.

In both cases, criticisms of State-nonconceptualism are motivated by the general requirement that content must capture how things appear to the subject. Let us call it Bermúdez's condition of content-attribution. According to Toribio (2008) and Duhau (2011), Bermúdez's condition is required to solve problems of cognitive significance at the nonconceptual level. That is the reason why in her recent defense of State-nonconceptualism against Bermúdez's and Toribio's criticisms, Duhau (2011) holds that the *form* of mental states can solve problems of cognitive significance without appealing to any Fregean content.

In this paper, I will first argue against Duhau that the mere syntactic form cannot meet the cognitive significance requirement but only a neo-Fregean conception of content. However, I will also argue on behalf of State-nonconceptualism that Bermúdez's constraint has a further dimension that has been overlooked by both Bermúdez and Toribio. Beyond considering the way things appear to the creature, a reasonable constraint on content-attribution must also take into account how the creature interacts with their conspecifics. This requires us to assume that on several occasions, creatures from different perspectives are representing the same content, individuated either as a structured sequence of particulars, properties, and relations (Russellian proposition) or as a set-theoretical proposition, that is, a function from possible worlds to truth-values.

The claim I support in this paper is the following. The solution to this apparent contradiction is the rejection of what I call the Real-content view and the adherence to what I would like to call here Content-pragmatism: there is no such thing as *the* content of experience. In this view, "proposition" is not as real as a mental state, but rather is a term of art that semanticists use, as a matter of theoretical convenience, to classify mental states. What follows from Content-pragmatism is Content-pluralism: there are so many contents that are required to meet Bermúdez's requirement, that any attribution of content must be justified.

Because both criticisms and the defense of State conceptualism overlook the real scope of Bermúdez's condition on content-attribution, they are ineffective. In this paper, I will argue that the opposition between State and Content nonconceptualism is a real one, but only a relative one, that is, relative to the opposite constraints to be met. If we want to solve the problems of cognitive significance, the best we can do is let the content of experience be modeled as neo-Fregean content, namely, a compound of nonconceptual modes of presentation of objects and properties. In this case, Content-nonconceptualism prevails. In contrast, if we want to account for how conspecifics join attention to the same entities in the same perceptual field from different perspectives, the best we can do is let the content of experience be modeled either as a structured Russellian proposition or as an unstructured set-theoretical proposition (a function from possible worlds to truth-values). In this case, State-nonconceptualism prevails.

Nonconceptual Content and the Way the Subject grasps the World

According to Bermúdez (2007), State-nonconceptualism is unmotivated. He gives us at least three reasons for not taking it seriously. His first source of dissatisfaction emerges from the alleged difficulty of Sate-nonconceptualism to meet what I want to call here Bermúdez's condition of content-attribution under the key assumption that propositional attitudes and experiences might share the same content. Let me put Bermúdez's condition of content-attribution in the simplest way:

> Bermúdez's condition: content-attribution has to capture how things appear to the subject or how the subject represents the world as being.

Bermúdez invite us, first, to assume that both propositional attitudes and experiences possess Fregean content, that is, a structured compound of concepts (senses). By such a view, Bermúdez's condition is easily met in the case of propositional attitudes. It is perfectly comprehensible how the subject of any propositional attitude might represent a structured compound of concepts, since she possesses the corresponding conceptual abilities that reflect how she represents the world. In contrast, under the assumption that the content is Fregean, it is hard to see how Bermúdez's condition could be met in the case of experiences. If the perceptual content has to capture how things perceptually appear to the perceiving subject, the subject could not possibly represent a structured compound of concepts, if by definition (of nonconceptual content) she does not possess any of the conceptual abilities required to specify correctly the content.

According to Duhau (2011), Bermúdez's endorsement of Bermúdez's condition relies on a neo-Fregean view of the content of both propositional attitudes and experiences. For one thing, only a fine-grained notion of content seems to be able to capture the way the subject captures the world as being. Yet, she argues, State-nonconceptualism relies on a different coarse-grained view of content, understood either as a structured Russellian proposition (consisting of structured compounds of objects and properties) or as unstructured propositions (functions from possible worlds to truth-values).

However, this does not seem completely right. First, Bermúdez's condition puts reasonable general constraint on content-attribution that everyone should accept, regardless of how one understands the representational content of propositional attitudes and experiences; indeed, regardless of one's adherence to content-externalism. Thus, Bermúdez's endorsement of his own condition on content-attribution cannot be seen as a direct consequence of his adherence to a neo-Fregean view of content. Nonetheless, under this assumption, it is hard to see how Bermúdez's condition could be met. For one thing, it is hard to see how a coarse-grained notion of content could possibly capture the peculiar way things perceptually appear to the perceiving subject.

Bermúdez's second reason for not taking State-nonconceptualism seriously is this. According to him, one of the main reasons for introducing the very notion of nonconceptual content is to account for discriminative abilities of objects, properties, and relations in the distal world in a way that both constitutes a precondition for the acquisition of conceptual abilities of observational concepts. As before, the accusation is that this requirement relies on his Fregean view on the content of experience (as picture-like, or as a *positioned scenario* in the way suggested by Peacocke, 1992). Yet, Bermúdez's point is that no matter how we understand the content of experience, we cannot account for discriminative abilities in a way that is independent of conceptual abilities under the key assumption of State-nonconceptualism, according to which the subject merely stands in relation to the same content of propositional attitudes.

Bermúdez's last reason for rejecting State-nonconceptualism is the claim that without assuming Content-conceptualism, we cannot account for the individuation of nonconceptual states. The idea here is that proponents of State-nonconceptualism owe us an explanation of what constitutes nonconceptual states. The natural suggestion is to assume that nonconceptual states are those individuated by appealing to a *distinctive type of content*, namely, the content that is not composed of concepts (Content-nonconceptualism). This answer is obviously not available to State-nonconceptualism since, under this assumption, nonconceptual

states are not individuated, nor entail nonconceptual contents. One alternative is to appeal to the functional role of the nonconceptual state in question. However, since the functional role of a propositional attitude is concept-dependent while the functional role of experiences is concept-independent, any appeal to functional roles restates the problem, rather than providing an explanation for it.

Nonconceptual Content and Cognitive Significance

Like Bermúdez, Toribio explicitly endorses Bermúdez's condition on content-attribution. In her own formulation, content must reflect the way the subject grasps the world as being. According to her, the only way of endorsing the logical independence of Content-nonconceptualism and State-nonconceptualism is by adopting a coarse-grained notion of content, modeled either as a Russellian proposition (consisting of a structured compound of objects and properties) or as an unstructured set-theoretical proposition (a function from possible worlds to truth-values). In other words, proponents of the logical independence of State-nonconceptualism and Content-nonconceptualism have to endorse a coarse-grained notion of content.

This is not completely true. Crane explicitly endorses the independence of State-nonconceptualism and Content-nonconceptualism and assumes a neo-Fregean view of the content of experience (Crane, 2009). In one way or the other, according to Toribio, those coarse-grained views of content are unsuitable for accounting for the subject's intentional behavior in a way that reflects how the subject grasps the world as being. It is hard to see how *n*-tuples of objects and properties (Russellian propositions) or a set of worlds could capture the different ways the subject believes something.

Likewise, if we let the content of experience be modeled either as a Russellian proposition or as an unstructured set-theoretical proposition, it can hardly meet Bermúdez's condition. As before, it is difficult to see how compounds of objects and properties or any set of possible worlds could possibly capture the different ways the subject experiences the world. Moreover, the assumption that experiences and propositional attitudes might share the same coarse-grained content blocks the natural way of understanding the process of conceptualization as consisting in subsuming entities under concepts (by picking out referents by Fregean senses).

Now, while Bermúdez raises the question of whether the representational content of experience is Fregean or a function from possible worlds to truth-values, Toribio (2008) clearly endorses a neo-Fregean fine-grained notion of content. She does so by going a step further and connecting the satisfaction of Bermúdez's condition to the solution of problems of cognitive significance:

> The content of propositional attitudes must account for (i) the fact that a rational subject can believe *Fa* while disbelieving *Fb* even when *a* is *b*, and (ii) the fact that a rational subject can believe *Fa* even when *a* lacks reference.

A classic example is the famous case of Tybalt in Shakespeare's Romeo and Juliet. Tybalt believes that Romeo loves Juliet but disbelieves that the son of Montague loves Juliet, for the simple reason that he (Tybalt) ignores the fact that Romeo and the son of Montague are the same person. Tybalt's problem is easily solved when we assume that he has two ways (two concepts or modes of presentation) of representing the same person without realizing it. This is what makes reasonable his belief and disbelief that the same person loves Juliet. Thus, according to Toribio, Bermúdez's condition is entailed as a solution to problems of cognitive significance; additionally, the only way to address to those problems is to adhere to a neo-Fregean view of

content, according to which the content of beliefs and other propositional attitudes must be seen as structured compounds of concepts or modes of presentations of the referents.

Even though it is an open question as to how we can or must understand the nonconceptual analogue of Fregean senses, we can easily formulate the same requirement for the content of experience in the following terms:

> The representational content of experience must account for the fact that, (i) under normal conditions, a subject may perceive *a* as *F* but not *b* as *F* even when *a* is *b*, and for the fact that, (ii) under abnormal conditions, a subject can experience *a* as *F* even when *a* lacks reference.

In the case of experience, we might say in an equally loose sense that a Fregean case is one in which a subject has two experiences of the same object without realizing it. Thus, there will of course be countless cases in this sense: seeing the same object from two sides in a mirror, and so on. Now, since State-nonconceptualism assumes a coarse-grained notion of content, it certainly cannot address the problems of cognitive significance. Thus, the only remaining alternative is Content-nonconceptualism: nonconceptual mental states are those whose contents are not complex compounds of concepts. The moral is that, under the requirement of solving problems of cognitive significance, the only notion of content relevant in this debate is neo-Fregean one. Thus, State-nonconceptualism is not logically independent of Content-nonconceptualism, but rather entails it.

Nonconceptual Content and the Constraint of Same-Representing

There is no doubt that both Bermúdez and Toribio are right in endorsing Bermúdez's condition and claiming that Bermúdez's condition is entailed as a solution to the problems of cognitive significance. Moreover, as I will argue in sequence, they are also right when they claim that only a neo-Fregean fine-grained notion of content can address those problems. However, I believe that Bermúdez's condition on content-attribution is not required only to solve problems of cognitive significance (which is what leads them directly to claim that State-nonconceptualism either presupposes Content-nonconceptualism or is an inconsistent view). In other words, I reject the shared view that there is one way (a single proposition) of expressing how the subject grasps the world as being or how the world perceptually appears to the perceiving subject.

First, let me defend the claim that only a neo-Fregean fine-grained notion of content can address the problems of cognitive significance. In a recent paper, Duhau (2011) has argued in support of State-nonconceptualism by holding that solutions to the problems of cognitive significance can also be found at the level of mental representation (rather than at the level of content). According to her, what State-nonconceptualism does is account for the way the subject grasps the world as being by holding that the same content might be represented in different ways. Although she does not, she could appeal here to the linguistic meaning of sentences and terms to make her point. Indeed, according to Kaplan (1989) and the older view of Perry (1979), mental states are individuated by their "character" or their "role" (linguistic meaning) rather than by their content ("what is said"). According to Kaplan's famous example, what accounts for the difference in behavior of a person seeing her pants on fire in the mirror without realizing she is the one whose pants are on fire, and the same person thinking to herself in the first person that her pants are on fire, is not the content (what is said) of respective utterances of thoughts, but rather the "character" or linguistic meaning of the sentences and terms involved. Thus, even though the subject's utterances might represent the

very same propositional content, the subject's being in different states of mind (accounted for by the different linguistic meaning of her two utterances) is what accounts for the difference in cognitive significance of her behavior. In this way, we find an answer to Bermúdez's challenge (the last of his arguments) that State-nonconceptualism cannot account for the individuation of nonconceptual mental states.

However, following Fodor (1998), Duhau endorses a further alternative: mental states are not individuated by meaning but rather by form or syntactic structure. Thus, problems of cognitive significance could be addressed at the level of mental representation in the following way. A rational person might simultaneously believe *Fa* and disbelieve *Fb*, even when *a* is *b*, provided she fails to realize that she is representing the same object but under the different syntactic representations of *a* and *b* (Duhau, 2011: 9).

The difficulty here is in understanding how the subject could fail to realize that she is referring to the same object under different modes of presentation *if the difference of those modes of presentation lies in the form of her representation*. The syntactic form of a mental representation (or even its linguistic meaning, if you will) is certainly not processed at the level of consciousness or at the so-called personal level. In contrast, problems of cognitive significance are raised at the personal or conscious level. To address those problems in a way that accounts for how a rational subject could fail to realize that the she is undergoing different experiences of the same object, we have to assume that certain identifying properties of the object in those experiences come to the foreground and are represented as part of the representational content of experience; otherwise, the explanation is empty. It is only in assuming that the subject is somehow consciously representing identifying features of the object that sense can be made of the fact that the subject fails to realize that she is experiencing the very same object under two different modes of presentation.

However, as I anticipated, I also reject Bermúdez and Toribio's view that Bermúdez's condition on content-attribution is required only to solve problems of cognitive significance. I reject the view that there is one way (a single proposition) of expressing how the subject grasps the world as being. To capture how things perceptually appear to the perceiving subject, we also need a coarse-grained notion of the representational content of experience.

To begin with, we need coarse-grained content to account for the fact that the subject experiences the same compound of objects and properties under quite different perceptual conditions such as luminosity, distance, angle, and the perceiver's perspective in general. This is what is known in psychological literature under the label of "perceptual constancy." For example, regardless of whether the sun is shining or it is a cloudy day, intuitively, what I see in both cases is the same greenness of the grass rather than some mode of presentation of it (roughly the color that is causing the color experience in me here and now). Thus, the natural suggestion here is to assume that the greenness of the London gardens is part of the coarse-grained content rather than any mode of presentation of it.

The main reason that militates in favor of my claim is this. Any conscious perception has a *cognitive impact* on a system of thoughts and beliefs. Let us suppose that I see the same greenness of the London gardens under quite different weather conditions: I see it first on a sunny day and then on a cloudy day. Let us assume further, for the sake of argument, that by means of my perceptual experiences I acquire different *de dicto* thoughts or beliefs about the greenness of the London garden (roughly the color that caused in me the experience of light green on a sunny day and the color that caused in me the experience of dark green on a cloudy day). My point is that it is certainly much harder to account for how I can come to the conclusion, for example, that the color of the London gardens are dark green if I am not

thinking *de re of that* greenness but rather *de dicto* under the different modes of presentation: the color that caused in me the experience of light green on a sunny day and the color that caused in me the experience of dark green on a cloudy day. By all accounts, I am thinking *of that* greenness rather than any mode of presentation of it. Thus, the natural suggestion is that my thought is *de re* of *that same* greenness. Moreover, the further natural assumption here is that that same greenness belongs to some representational content of our perceptual experiences (coarse-grained individuated) that is best modeled either as a structured Russellian proposition or as an unstructured set-theoretical proposition.

Thus, by perceiving the greenness of the London gardens on a sunny day or on a cloudy day, what I naturally do is acquire a *de re* thought *of that* greenness rather than the *de dicto* belief or thought. All I need now to support my claim is a further natural assumption, namely, that the content of *de re* thought is inherited from the content of the original perception. Under this assumption, the conclusion is that the content of perception is also coarse-grained individuated, modeled either as a Russellian proposition or as a set-theoretical proposition.

Bermúdez's condition on content-attribution also requires a coarse-grained notion of the content of experience to account for the way different subjects capture the world from different vantage points when they have joint attention. Joint attention is the ability to share a common focus on something with someone else (objects, properties, people, etc.). As such, it can be seen as the most primitive form of nonverbal communication. Joint attention skills can be a predictor of future language development. Joint attention starts in infancy between a child and a caregiver. Early skills can include
reaching to be picked up by a caregiver, pointing to a stuffed animal, or looking at the same page in a book, etc.

Now, as the children and adults involved have different perspectives on the same object or property, we can only meet Bermúdez's condition by letting the content of experience be modeled as either a Russellian proposition or a set-theoretical proposition (consisting of the very same entities rather than modes of presentations of them). For example, when a child and its caregiver apply joint attention to the same color on a book's page, what their visual experiences represent from their own different perspectives is the color itself painted on the book, rather than any mode of presentations of it. If each of them were representing an identifying property of that color (roughly the color that is causing the color experience in each of them) rather than the color itself, it would be difficult to understand how they could communicate.

Now, the argument behind my claim is similar to the previous one. Any conscious perception has a *cognitive impact* on a system of thoughts and beliefs. Let us suppose that two children are attending to the same object or property (for example, the same toy) from their own perceptual perspectives. Let us suppose now that by means of their own perspectives they both acquire different *de dicto* thoughts about the same toy, roughly the toy that caused in one of them the toy experience from her perspective, and the toy that caused in the other the toy experience from her different perspective. My point, as before, is that it is much harder to account for the possibility of a cogent agreement and disagreement between them when, for example, one of them asks the other about their favorite toy. By all accounts, we are talking *of that* toy rather than any mode of presentation of it. Indeed, it is hard to see how there could be any cooperation or dispute between them when, for example, one of them asks the other to give her *that* toy. In cases like this, the natural assumption is that agreement and disagreement, cooperation and dispute, require *de re* thought of the *same* toy. And, as before, the natural assumption here is that *that same toy* belongs to some coarse-grained content of their perceptual

experiences of a toy that is best modeled either as a Russellian proposition or as a set-theoretical proposition.

Thus, by perceiving the same toy, both children naturally acquire a *de re* thought *of* that toy rather than a *de dicto* belief or thought consisting of the object that is causing in one of them her toy experience (mode of presentation). Now, assuming also that the content of such a *de re* thought is best modeled as a coarse-grained proposition (consisting of the very property of greenness rather than the mode of presentation of it) and, further, that this content of a *de re* thought is naturally inherited from the content of the original perception, the natural conclusion here is to assume that the content of perception is also a coarse-grained proposition, modeled either as a Russellian proposition or as a set-theoretical proposition.

Moreover, the assumption that the content of experience is best modeled as either a Russellian proposition or a set-theoretical proposition is the best available account for the fact that joint attention is a condition for future language development. In a wide variety of cases, we can only make sense of *what is said* from different utterances if we assume that the truth of falsity turns on the same objects having the same properties. For example, if I say to you "that food is poisoned" and you disagree, what really matters in our communicative exchange, and what is really said, is that that object possesses the property of being poisoned.

Disclaimer: my complaint here is not that there is no possibility of agreement and disagreement over Fregean content or any other kind of content; what I mean is that in a wide variety of cases, identifying properties that you and I might have used to identify the object in question, or that you and I might have associated with the phrase "that food," do not really matter.

To be sure, Bermúdez and Toribio are right when they claim that Bermúdez's condition is required to solve problems of cognitive significance. Still, they seem to overlook that the same condition on content-attribution also requires what I would like to call here same-representing:

> The content must account for the fact that the same individual might represent the same entities under changing conditions of experience, as well as account for the fact that different individuals might apply joint attention to the same entities from their own viewpoints.

Nonconceptual Content: Content-Pragmatism versus Content-Realism

Now, if Bermúdez (2007) is right by proposing a condition on content attribution and Toribio is also right by claiming that such a condition is required to solve problems of cognitive significance, the intriguing question is how they both overlook that Bermúdez's conditions also requires same-representing. My explanatory hypothesis is that they both rely on what I want to call here a Content-realist view on the content of mental states:

> For every experience or propositional attitude, there is one and only one real content or proposition by means of which the mental state of the subject is individuated or constituted.

Bermúdez tacitly relies on the Content-realist view when he claims that: "appealing to a content is not a *matter of theoretical convenience*. We must also make sense of the idea of the subject being related to the relevant abstract object" (2007: 67, my emphasis). Toribio is even more explicit when she promises to prove that State-nonconceptualism does entail Content-nonconceptualism "on the only notion of content that is relevant for this debate"

(Toribio, 2008: 355). Now, the Realist view directly opposes what I would like to call Content-pragmatism.

Propositions are just tools we use, as a matter of theoretical convenience, to classify mental states.

I would like now to defend Content-pragmatism against Content-realism, showing that Content-realism relativizes the debate between State and Content-nonconceptualism. Let me begin with Bermúdez's claim:

> Fregean contents are *abstract objects*, but appealing to contents *is not a matter solely of theoretical convenience. We must be able to make sense of the idea of the subject being related to the relevant abstract object.* And, of course, we want the subject's being so related to explain how she represents the world. In the case of beliefs and other propositional attitudes with conceptual content, this is perfectly comprehensible, because we assume that the subject possesses all the relevant concepts. In fact, talk of conceptual contents is in some respects just a complicated way of talking about which subset of the subject's conceptual abilities is currently being deployed. So the connection between the content of belief and the way the subject represents the world is immediate. But this obvious way of bridging the gap between subject and content is closed to us if there is no requirement that the subject possesses the relevant concepts. It is hard to know what it means to say that being in a perceptual state with a particular content is a matter of standing in relation to a complex of concepts, none of which is possessed by the perceiver. This makes the idea that perception is related to a complex of concepts completely mysterious (2007: 67, Emphases are mine).

In this passage, Bermúdez characterizes the core of Content-realism by making two related claims. The first is the claim that contents are not a matter of theoretical convenience. The second is the traditional view that propositional attitudes are a *sui generis* relation between the subject, as a concrete entity in space and time, and a proposition, as an abstract entity. Therefore, by entertaining a propositional attitude, we are actually related to some abstract entity.

I begin by remarking that there is something quite amiss with the traditional idea of propositional attitudes as subjects standing in relation to propositions as abstract entities, regardless of whether the subject possesses the concepts required to specify that proposition or not. Nevertheless, even more amiss is the suggestion that conscious experience is constituted by a "sensing relation to a proposition" as held by Papineau: "There seems something quite amiss with the suggestion that my here-and-now conscious feelings are constituted by my bearing any kind of relation to abstract entities" (2014: 6).

As we have seen, there are different ways of understanding propositions. We may understand them as unstructured set-theoretic propositions (functions from possible worlds to truth-values), or as Russellian structured sequences of objects and properties, or finally as Fregean structured complexes of concepts. Even though there are interesting differences between these views, the point I want to make here is that propositional attitudes can hardly be seen as relations between a concrete subject in space and time and propositions as abstract entities.

However, if propositional attitudes are not *real* relations between subjects, as concrete entities in space and time, and propositions as abstract entities, the natural question that emerges is about the role of a proposition in the specification of the content of experience. My suggestion

here is to give up Content-realism and embrace Content-pragmatism. My complaint is not that perceptual states or even propositions are illusory. It must be clear that I am not endorsing either an eliminativism about consciousness or an eliminativism about propositions. The idea is rather that propositions are different ways of classifying states as a matter of theoretical convenience (to put it in Bermudez's words).

As I understand it, Content-pragmatism is not a new view, but rather a view shared by different philosophers of mind and language, such as Perry, Chalmers, and Crane, among others. The key idea is that neither propositional attitudes nor experiences are relations to propositions. "Proposition" is just a term of art, created by semanticists to classify subjects' states. I subscribe here to Crane's analogy between the role of propositions and the role of "models" in science: we let some aspect of a subject's mental state be *modeled* on a proposition in the same way that we let a cognitive process be *modeled* on a computer (2011: 34). Likewise, I also subscribe to Perry's analogy between propositions with lengths and weights (2009: 21): we classify mental states using propositions in the same way we classify someone's weight using pounds or someone's length using meters.

The natural consequence is Content-pluralism:

> There is no such thing as *the* representational content of experience: we are free to use different kinds of propositional and non-propositional contents (say, picture-like contents) to model the subject's perceptual states, taking into account the different constraints on content-attribution.

Thus, whenever we want to satisfy Bermúdez's condition by solving problems of cognitive significance, the best we can do is let the subject's perceptual state be modeled as neo-Fregean content consisting of a structured compound either of concepts (in the case of propositional attitudes) or of nonconceptual modes of presentation (in the case of experiences). In contrast, whenever we want to satisfy Bermúdez's condition by accounting for same-representing, the best we can do is let the subject's perceptual state be modeled either as a structured sequence of objects and properties or as an unstructured set of possible worlds.

Now, if Bermúdez (2007) is right by proposing a condition on content attribution and Toribio is also right by claiming that such a condition is required to solve problems of cognitive significance, the intriguing question is how they both overlook that Bermúdez's conditions also requires same-representing. My explanatory hypothesis is that they both rely on what I want to call here a Content-realist view on the content of mental states:

> For every experience or propositional attitude, there is one and only one real content or proposition by means of which the mental state of the subject is individuated or constituted.

Bermúdez tacitly relies on the Content-realist view when he claims that: "appealing to a content is not a *matter of theoretical convenience*. We must also make sense of the idea of the subject being related to the relevant abstract object" (2007: 67, my emphasis). Toribio is even more explicit when she promises to prove that State-nonconceptualism does entail Content-nonconceptualism "on the only notion of content that is relevant for this debate" (Toribio, 2008: 355). Now, the Realist view directly opposes what I would like to call Content-pragmatism.

Propositions are just tools we use, as a matter of theoretical convenience, to classify mental states.

I would like now to defend Content-pragmatism against Content-realism, showing that Content-realism relativizes the debate between State and Content-nonconceptualism. Let me begin with Bermúdez's claim:

> Fregean contents are *abstract objects*, but appealing to contents *is not a matter solely of theoretical convenience. We must be able to make sense of the idea of the subject being related to the relevant abstract object.* And, of course, we want the subject's being so related to explain how she represents the world. In the case of beliefs and other propositional attitudes with conceptual content, this is perfectly comprehensible, because we assume that the subject possesses all the relevant concepts. In fact, talk of conceptual contents is in some respects just a complicated way of talking about which subset of the subject's conceptual abilities is currently being deployed. So the connection between the content of belief and the way the subject represents the world is immediate. But this obvious way of bridging the gap between subject and content is closed to us if there is no requirement that the subject possesses the relevant concepts. It is hard to know what it means to say that being in a perceptual state with a particular content is a matter of standing in relation to a complex of concepts, none of which is possessed by the perceiver. This makes the idea that perception is related to a complex of concepts completely mysterious (2007: 67, Emphases are mine).

In this passage, Bermúdez characterizes the core of Content-realism by making two related claims. The first is the claim that contents are not a matter of theoretical convenience. The second is the traditional view that propositional attitudes are a *sui generis* relation between the subject, as a concrete entity in space and time, and a proposition, as an abstract entity. Therefore, by entertaining a propositional attitude, we are actually related to some abstract entity.

I begin by remarking that there is something quite amiss with the traditional idea of propositional attitudes as subjects standing in relation to propositions as abstract entities, regardless of whether the subject possesses the concepts required to specify that proposition or not. Nevertheless, even more amiss is the suggestion that conscious experience is constituted by a "sensing relation to a proposition" as held by Papineau: "There seems something quite amiss with the suggestion that my here-and-now conscious feelings are constituted by my bearing any kind of relation to abstract entities" (2014: 6).

As we have seen, there are different ways of understanding propositions. We may understand them as unstructured set-theoretic propositions (functions from possible worlds to truth-values), or as Russellian structured sequences of objects and properties, or finally as Fregean structured complexes of concepts. Even though there are interesting differences between these views, the point I want to make here is that propositional attitudes can hardly be seen as relations between a concrete subject in space and time and propositions as abstract entities.

However, if propositional attitudes are not *real* relations between subjects, as concrete entities in space and time, and propositions as abstract entities, the natural question that emerges is about the role of a proposition in the specification of the content of experience. My suggestion here is to give up Content-realism and embrace Content-pragmatism. My complaint is not that perceptual states or even propositions are illusory. It must be clear that I am not endorsing either an eliminativism about consciousness or an eliminativism about propositions. The

idea is rather that propositions are different ways of classifying states as a matter of theoretical convenience (to put it in Bermudez's words).

As I understand it, Content-pragmatism is not a new view, but rather a view shared by different philosophers of mind and language, such as Perry, Chalmers, and Crane, among others. The key idea is that neither propositional attitudes nor experiences are relations to propositions. "Proposition" is just a term of art, created by semanticists to classify subjects' states. I subscribe here to Crane's analogy between the role of propositions and the role of "models" in science: we let some aspect of a subject's mental state be *modeled* on a proposition in the same way that we let a cognitive process be *modeled* on a computer (2011: 34). Likewise, I also subscribe to Perry's analogy between propositions with lengths and weights (2009: 21): we classify mental states using propositions in the same way we classify someone's weight using pounds or someone's length using meters.

The natural consequence is Content-pluralism:

> There is no such thing as *the* representational content of experience: we are free to use different kinds of propositional and non-propositional contents (say, picture-like contents) to model the subject's perceptual states, taking into account the different constraints on content-attribution.

Thus, whenever we want to satisfy Bermúdez's condition by solving problems of cognitive significance, the best we can do is let the subject's perceptual state be modeled as neo-Fregean content consisting of a structured compound either of concepts (in the case of propositional attitudes) or of nonconceptual modes of presentation (in the case of experiences). In contrast, whenever we want to satisfy Bermúdez's condition by accounting for same-representing, the best we can do is let the subject's perceptual state be modeled either as a structured sequence of objects and properties or as an unstructured set of possible worlds.

Conclusion

We have seen that the main reason for favoring Content-nonconceptualism and disregarding State-nonconceptualism is the assumption that a coarse-grained notion of content cannot account for the cognitive significance of a subject's beliefs and experiences. We have also seen that the main line of defense for State-nonconceptualism against such criticism is the mistaken assumption that the syntactic form of representations can solve problems of cognitive significance (Duhau, 2011). Against this last view, I have argued that only a neo-Fregean notion of content of attitudes and experiences can solve problems of cognitive significance. Identifying properties must come to the foreground and be represented as part of the representational content.

Still, I have argued that Bermúdez's condition on content-attribution also entails the satisfaction of an opposite constraint: same-representing. Moreover, I have argued that only a coarse-grained notion of content of experience can meet that constraint. Now, considering that solving problems of cognitive significance and meeting the requirement of same-representing clearly pull in opposite directions, my next step was to argue in favor of Content-pragmatism and against Content-realism. Once we recognize both constraints and endorse Content-pragmatism, both Bermúdez and Toribio's reasons against State-nonconceptualism, and Duhau's defense of State-nonconceptualism against Bermúdez and Toribio's criticism become ineffective: Content-nonconceptualism and State-nonconceptualism are simultaneously logically independent and consistent because they are based on opposite constraints on content-attribution.

The moral to be drawn is that the opposition between Content-nonconceptualism and State-nonconceptualism is a real one, but it is only relative, that is, relative to the opposite con-

straints on content-attribution to be met. If we want to make sense of the perceiver's behavior that perceives that *a* is *F*, while misperceiving *b* as not being *F*, even when *a* is *b*, the best we can do, *as a matter of theoretical convenience*, is let the content of both experiences be modeled as neo-Fregean content, compounded of different (non-conceptual) modes of presentation of the same object, associated with *a* and associated with *b* (whatever these are). In this sense, Content-nonconceptualism prevails: the representational content of experience is nonconceptual in the relevant sense of not being composed of concepts, but rather of nonconceptual modes of presentations.

In contrast, if we want to make sense of joint attention or non-verbal communication between perceivers experiencing the object from different vantage points, the best we can do, *as a matter of theoretical convenience*, is let the content be modeled either as a Russellian proposition or as a set-theoretical proposition composed of objects and their properties, rather than of modes of presentations of them. In this sense, State-nonconceptualism prevails: the perceptual state is nonconceptual in the relevant sense that the subject does not need to possess the concepts required to specify correctly the content of her experiences.

References

Bermúdez, J. L. (2007), 'What is at stake in the debate on nonconceptual content?', *Philosophical Perspectives* **21**, 55–72.

Byrne, A. (2005), Perception and Conceptual Content, *in*: E. Sosa and M. Steup (eds.): 'Contemporary Debates in Epistemology', Blackwell, Oxford, pp. 231–250.

Crane, T. (2009), 'Is perception a propositional attitude?', *Philosophical Quarterly* **59 (236)**, 452–469.

Crane, T. (2011), 'The Singularity of Singular Thought', *Aristotelian Society* Supplementary Volume **85 (1)**, 21–43.

Crowther, T. (2006), 'Two Conceptions of Conceptualism and Nonconceptualism', *Erkenntnis* **65**, 245–276.

Duhau, L. (2011), 'Perceptual Nonconceptualism: Disentangling the Debate between Content and State Nonconceptualism', *European Journal of Philosophy* **22 (3)**, 358–370.

Fodor, J. (1998), *Concepts: Where Cognitive Science Went Wrong*, Oxford University Press, Oxford.

Heck, R. (2000), 'Nonconceptual Content and the 'Space of Reasons', *The Philosophical Review* **109**, 483–523.

Heck, R. (2007), Are There Different Kinds of Content?, *in*: J. Cohen and B. McLaughlin (eds.), 'Contemporary Debates in the Philosophy of Mind', Blackwell, Oxford, 117–138.

Kaplan, D. (1989), Demonstratives, *in*: Almog, Perry, and Wettstein (eds.), *Themes from Kaplan*, Oxford University Press, USA, 481–563.

Papineau, D. (2014), 'The Presidential Address: Sensory Experience and Representational Properties', *Proceedings of the Aristotelian Society* **114**, 1–33.

Peacocke, C. (1992), *A Study of Concepts*, MIT Press, Cambridge MA.

Perry, J. (1979), 'The Problem of the Essential Indexical', *Noûs* **13**, 3–21.

Perry, J.(2009), *Reference and Reflexivity*. Center for the Study of Language and Information, University of Chicago Press, Chicago.

Speaks, J. (2005), 'Is There a Problem about Nonconceptual Content?', *The Philosophical Review* **114**, 359–398.

Toribio, J. (2008), 'State versus Content: The Unfair Trial of Perceptual Nonconceptualism', *Erkenntnis* **69**, 351–361.

abstracta Linguagem, Mente & Ação ISSN 1807-9792 Volume 8, Number 2, pp. 31–51, 2015

In defense of empathy: A response to Prinz[1]

Claudia Passos-Ferreira

Universidade Federal do Rio de Janeiro
Departamento de Filosofia, Post-Doc
Largo de São Francisco de Paula 1 sala 310 Centro 20051070
Rio de Janeiro, RJ - Brasil - Caixa-postal: 22246379
cpassosferreira@gmail.com

Abstract

A prevailing view in moral psychology holds that empathy and sympathy play key roles in morality and in prosocial and altruistic actions. Recently, Jesse Prinz (2011a, 2011b) has challenged this view and has argued that empathy does not play a foundational or causal role in morality. He suggests that in fact the presence of empathetic emotions is harmful to morality. Prinz rejects all theories that connect empathy and morality as a constitutional, epistemological, developmental, motivational, or normative necessity. I consider two of Prinz's theses: the thesis that empathy is not necessary for moral development, and the thesis that empathy should be avoided as a guide for morality. Based on recent research in moral psychology, I argue that empathy plays a crucial role in development of moral agency. I also argue that empathy is desirable as a moral emotion.

1 Empathy and morality

A prevailing view in moral psychology holds that the cognitive abilities of empathy and affective perspective-taking play key roles in morality and in prosocial and altruistic behaviors. According to numerous psychologists (Eisenberg & Strayer 1987; Batson et al. 1981; Batson & Shaw 1991; Zahn-Waxler & Radke-Yarrow 1990; Hoffman 2000; Vaish et al. 2009, 2011; Decety 2011) and philosophers (Hume 1739/1978; Smith 1759/2009; Slote 2010; Goldman 2006; Darwall 1998; de Vignemont & Frith 2008), empathy is fundamental to morality.

Recently, Jesse Prinz (2005, 2011a, 2011b) has challenged this trend in moral psychology. Prinz has two major theses. First, he argues that empathy is not a necessary precondition for moral approval and disapproval. Second, he argues that empathy is prone to biases that render it potentially harmful and frequently produce morally undesirable results. His counterintuitive conclusions are that empathy plays no essential role in morality and that it interferes negatively with the ends of morality; therefore, it should not be cultivated.

I will argue against both theses. First, I will argue that empathy plays a necessary role in human moral development. I argue that empathy – understood either as vicarious sharing

[1]Research for this paper was supported by a five-year grant (Programa Nacional de Pós-Doutorado/CAPES) from the Philosophy Graduate Program of Federal University of Rio de Janeiro (UFRJ) for which I am grateful. Special thanks to David Chalmers, David Copp, Norman Mandarasz, Philippe Rochat, Noel Struchiner and an anonymous reviewer for helpful comments on earlier versions of this paper. For helpful comments, I would also like to thank Caroline Marin, Wilson Mendoncá, Cinara Nahra, Ricardo Bins de Napoli, Adriano Naves, Jesse Prinz, Roberto Sá Pereira, Tom Sorell, Flavio Williges, and participants of the June 2012 Workshop on Ethics and Metaethics at UFSM, the June 2012 Brazilian Analytic Philosophy Society Meeting at Fortaleza, the August 2012 Second Latin American Analytic Philosophy Conference α2 in Buenos Aires, and the December 2012 Seminar on Bioethics and Environmental Ethics at UFRJ.

of emotion or as affective perspective-shifting through simulation and imaginative reconstruction – is fundamental to the development of moral agency. The absence or deficiency of these processes leads to the absence or deficiency of a crucial element of our morality. Second, I will argue that there is a moral benefit associated with empathic feelings. I also argue that there are certain morally demanding situations in which empathy is our best guide to moral judgment.

This paper falls into four sections. In the first section, I spell out Prinz's negative view of an empathy-based morality, and I clarify my thesis about the necessity of empathy in moral development, suggesting that the thesis should be understood as a claim about moral development in humans.

In the second section, I describe Prinz's argument against the developmental necessity thesis. Prinz's discussion of moral development focuses on three issues: moral deficits in psychopaths, moral deficits in autistic people, and theories of moral development. He argues that psychopaths' and autists' impaired empathic abilities are not responsible for their impaired moral competence, and he proposes an imitation-based account of moral development to explain how the development of moral competence involves the acquisition of emotional capacities via imitative learning. I review the empirical literature concerning psychopaths and autistic people and offer an alternative explanation for psychopaths' and autists' moral deficit which favors the empathy-based account of moral development that I propose.

In the third section, I defend a broader conceptualization of empathy that helps us to understand moral competence, and I distinguish two roots of empathy – perceptual empathy and imaginative empathy – based on distinct underlying mechanisms.

In the fourth section, I discuss Prinz's argument against the normative claim, and I argue that, despite potential biases of empathy, there are ways to resist the conclusion that empathy should not be cultivated. I argue that there are moral contexts in which empathy is a good moral guide.

2 Does empathy play a necessary role in morality?

I start with the issue of whether empathy is necessary for morality. Prinz is not the first to challenge the view that empathy is necessary here. Jeannette Kennett in "Autism, Empathy, and Moral Agency" (2002), Victoria McGeer in "Varieties of Moral Agency: Lessons from Autism (and Psychopathy)" (2008), Heidi Maibom in "Feeling for Others: Empathy, Sympathy, and Morality" (2009), and Peter Goldie in "Anti-empathy" (2011) have adopted this perspective as well. Many of those who argue against empathy as a precondition for morality adopt a Kantian rational account of morality. Kennett, for instance, argues that an examination of moral thinking in autistic people shows that moral agency can be developed in the absence of empathy, and that this evidence can support a Kantian account of moral agency.

Unlike these rationalist critique[2], Prinz's critique (2007) defends a sentimentalist account of morality. As a sentimentalist, Prinz shares the Humean intuition that emotions are essential for moral judgment and moral motivation and that moral judgments involve approval and disapproval. According to Prinz (2011a), Hume's sentimentalism can be formulated as a constitution claim: to believe that something is morally right or wrong consists of approving or disapprov-

[2]McGeer (2008) also endorses a Humean view of morality. She argues that reason-based judgments play an instrumental role in morality, that only emotions have the required motivational force that accompanies moral attitudes, and consequently that all kinds of human moral agency are rooted in affect. However, she denies that empathy and perspective-taking abilities are the basis of morality. She suggests that people with autism challenge this view. In autism, the deficit of empathy and perspective-taking abilities do not lead to a deficit in morality. She concludes that empathy should not be considered the only emotion to provide moral motivation; other kinds of affective dispositions (which are available to people with autism, such as affective concern) also play a role in morality.

ing it. Prinz endorses the Humean view of morality but dismisses empathy as the basis for moral approval or disapproval. He argues for what he calls an *antiempathic sentimentalism*.

In "Against Empathy," Prinz takes issue with the Humean thesis that empathy is a precondition for morality. He analyzes two central points of Hume's project: 1) the *definitional thesis* that empathy is feeling an emotion that we take another person to have; 2) the *precondition thesis* that empathy is a constitutive precondition for moral approbation or disapprobation. When the precondition thesis is combined with Hume's sentimentalism, it follows that empathy is a precondition for moral judgment. Prinz endorses Hume's sentimentalism but rejects his precondition thesis.

More precisely, Prinz denies that "moral approbation involves any kind of congruence between the emotions of the one who approves and those on either side of the action being approved of" (2011a: 218). There are many ways to connect empathy and moral judgment. However, Prinz rejects all theories that necessarily connect empathy and morality. He argues against six versions of the precondition thesis, corresponding to the following six types of precondition: 1) constitutive (empathy is a necessary element of moral judgment); 2) causal (empathy is necessary for causing moral judgment); 3) developmental (empathy is necessary for developing moral agency); 4) epistemic (empathy plays a necessary epistemic role in moral judgment); 5) normative (empathy is necessary for justifying moral judgment); and 6) motivational (empathy is necessary for moral motivation).

The general form of Prinz's arguments against these theses might be reconstructed like this:

P1. Moral judgments are constituted by sentiments of approbation and disapprobation.

P2. Empathy plays a contingent role in moral sentiments of approbation and disapprobation.

P3. If empathy plays a contingent role in moral judgment, empathy is not necessary for morality.

C1. Hence, empathy is not necessary for morality.

Premise P1 expresses Prinz's emotionism[3] and will not be the target of my discussion. Premise P2 is a modal thesis, which Prinz supports by counterexamples and psychological research. The conditional premise P3 covers all the types of necessary connection between empathy and morality denied by Prinz.

In effect, there are six different versions of the argument corresponding to the six different kinds of preconditions. As we have seen above, these six precondition theses make six different claims about a necessary role for empathy in morality. For example, the developmental precondition thesis says that empathy is necessary for moral development. In Prinz's argument against this thesis, premise P2 says that empathy plays a contingent role in moral development, and the conclusion says that empathy is not necessary for moral development. This is the version I will focus on.

I will deny premise P2, arguing that empathy plays a necessary role in moral development in human beings. I argue that there are aspects of our morality for which empathetic emotions are necessary. Where Prinz uses empirical results to argue against the developmental precondition thesis, I suggest different interpretations of the results. Recent experimental work shows that empathy plays a key role in the emergence of moral agency in human beings. I also endorse

[3]Emotionism – such as defined by Prinz (2007) – is any theory that claims that emotions are essential to morality. Prinz distinguishes the term 'emotionism' – which he defines as an overarching label for any view that claims that feelings are essential to morality – from the term 'emotivism' which is a specific version of emotionism. Prinz argues for a strong form of emotionist view, defending what he calls 'constructive sentimentalism,' the view that "sentiments literally create morals, and moral systems can be created in different ways" (Prinz 2007: 9).

an ontogenetic account of the development of moral agency based on empathy. I use these considerations to reject conclusion C1, which claims that empathy is not necessary for morality, at least when this claim is restricted to human beings.

The idea that empathy is not necessary for moral judgment can be understood either as a claim about all moral systems or as a claim about human beings. The former claim is that there are possible moral systems that have moral judgment without empathy. I accept that empathy is not necessary for moral judgment in this sense. We might conceive a primitive system, for instance, like a mythical savannah (Prinz 2007) populated by early humans living in a natural state governed by pre-moral values, where people do not need empathy to follow rules and respect authority. In this primitive situation, empathy might not be involved in attitudes of approval and disapproval. In a natural state, a primitive agent might use other elements to moralize. For instance, he might use reactions of disgust or anger as guides to moral action (although one question is whether this kind of behavior counts as moral thinking or as mere regulation of behavior).

For present purposes, the necessity claim should be understood as a claim about human beings. That is, because of the way human beings are psychologically constituted, empathy is necessary for human moral development. It is this claim that I will defend. It is true that even humans can morally disapprove of others without directly passing through empathy or the affective perspective-taking process. Recent psychological studies (Haidt 2012) show that moral judgment can be a result of automatic affective reactions. We can form a moral judgment on the basis of gut feelings. However, on the view I will argue for, gut feelings in humans only count as moralizing when they express attitudes of disapproval of a moral agent, and the characteristic features of a moral agent depend on the capacity to arrive at moral attitudes as a result of a process of empathic simulation and affective perspective-taking.

3 Autism, psychopathy, and moral development

Many contemporary works in moral psychology emphasize the constitutive role of empathy in early moral development (see Batson 1981, 1991, 2011; Eisenberg 2000; Hoffman 2000, 2011; Tomasello & Vaish 2013). The central thesis defended by many moral psychologists is that empathic processes are the psychological mechanisms underlying moral agency. Morality involves processes of behavioral regulation toward others. Quite early in ontogeny, infants start using empathic processes to regulate their behavior toward others. When a child shows no empathy to others, we can predict that she will fail both to acquire concern for others and to be able to appreciate how her actions affect others. Consequently, the lack of empathy will affect her capacity to evaluate her own actions and others' actions as being right or wrong and to react to those actions either expressing disapproval or approval. We can predict that people who display difficulties in empathic abilities are also impaired in morality.

The idea that empathy plays a central role in moral development is supported by several studies which investigate the correlation between empathy, prosocial behaviors, and cooperation (Batson 1981, 1991, 2011; Eisenberg & Strayer 1987; Eisenberg 2000; Hoffman 2000, 2011; Tomasello & Vaish 2013). Early in ontogeny, children start behaving prosocially and engaging in cooperative and altruistic actions. The question here is: what are the psychological mechanisms that enable children to behave prosocially? The studies have shown that emphatic processes provide children with the affective, cognitive, and motivational abilities to behave prosocially. Empathic processes motivate altruistic behaviors, such as helping, caring, and other-directed comforting behavior and relate negatively with antisocial and aggressive behavior (Batson 1981, 1991, 2011; Eisenberg 2000; Hoffman 2000). Empirical data show that the

capacity for moral reasoning and moral agency is strongly dependent on the capacity to respond empathetically to others'affective states and to take the perspective of others into account.

Prinz challenges this interpretation and argues that there is no empirical data to provide evidence for the strong conclusion that empathy is the basis of moral development. He does not deny the positive correlation between empathy and moral judgment, and empathy and prosocial behavior. However, he denies that this correlation is evidence for the developmental necessity thesis. Prinz (2011a) addresses three potential sources of evidence for the thesis: evidence concerning psychopaths, evidence concerning autistic people, and theories of moral development.

The two pathological populations – psychopaths and autistic people – have been of special interest in moral psychology. Both populations show deficiencies in social understanding, social responsiveness, and moral competence. Both psychopaths and autistic people also have impaired empathic abilities. A popular view is that their empathic impairments explain their moral impairments (Nichols 2004). Psychological research on these populations seems to support the view that lack of empathy affects moral competence (Blair 2005), suggesting that empathy plays a key role in moral development.

Prinz (2005, 2011a, 2011b) rejects this explanation. He argues that evidence from autistic people and psychopaths does not support the developmental precondition thesis. I will consider his arguments in both of these cases, focusing especially on the evidence from psychopaths.

Psychopathy. Psychopathy is a disorder associated with callous and unemotional traits (lack of fear, guilt, remorse, and shallow affect) and antisocial and aggressive behavior (Blair et al. 2005; Patrick 2005). It is widely held that psychopaths' emotional deficits explain their lack of empathy, and their impaired empathy explains the lack of moral competence. Psychopaths lack emotions that facilitate moral education and lack emotional responses that constitute moral judgments.

According to Prinz (2005), psychopaths' moral deficits can be explained without appeal to the empathy deficit. It is the lack of basic emotions, such as fear and sadness, and not the lack of empathy that explains the impairment in moral reasoning detected in people with psychopathy. The account Prinz favors is based on the dysfunctional fear hypothesis[4]. Under this hypothesis, psychopaths are impaired in the systems modulating fear behavior (Fowles 1988). Psychopaths show reduced aversive conditioning and reduced emotional responses in anticipation of punishment and in imagining threatening events. Psychopaths' fear deficit prevents them from being socialized and from developing moral competence. Moral socialization is achieved through the use of punishment. Aggressive punishment instills fear, and fear of punishment is often used during moral training. A child that is frightened by punishment will associate this fear with the action that resulted in the punishment and will develop conditioned aversive responses to anticipated threats. A child that does not display fear of punishment will not learn good conduct if threatened with punishment.

Prinz argues that psychopaths' moral impairment can be explained by a deficit in inhibitory behavior and inhibitory emotions. He claims that the same dysfunctional system that impairs fear in psychopaths may also impair sadness, other negative emotions, and negative reactions.

[4]Roughly, there are two cognitive models in empirical literature to explain moral deficit in psychopathy. The behavioral inhibitory system put forward by Fowles (1988) and others, that claims that psychopaths have a basic deficit in their rudimentary behavior system that underlies many aspects of emotions and causes impairment in aversive behavior and fear, and the early violence inhibitory model (VIM), updated to the integrated emotion systems developed by Blair (Blair et al. 2005), that explains the nature of the emotional impairment in individuals with psychopathy as a result of impairments in different systems, such as dysfunctional empathy, dysfunctional fear, and dysfunctional VIM.

Sadness, he claims, is crucial to morality because it is involved in recognition and response to the sadness of others, and it is a basic element that can be used to create moral emotions (such as shame and guilt).

Moral emotions are complex emotions that arise in contexts that involve conformation to or violation of a moral rule. Prinz (2005, 2007) holds the view that moral emotions (such as shame, guilt, regret and indignation) are generated by a blend of basic emotions (e.g. fear, sadness and anger), which are combined with a calibration process. In the calibration process, as proposed by Prinz, a basic emotion that initially had one set of eliciting conditions can be assigned a new set of eliciting conditions that have been elaborated through experience to form an independent elicitation mechanism. For example, Prinz claims that "guilt is just sadness that has been calibrated to situations in which one has caused harm to someone that one care's about" (Prinz 2005: 273). That is, the emotional blend can be associated with situations where the child "catches"another's emotional states (distress, negative reaction, disapproving, etc.) by copying another's emotional states through imitation or emotional contagion.

In "Imitation and Moral Development" (2005) Prinz gives imitative learning a fundamental place in the explanation of moral development. To develop moral competence, a child has to be able to react with negative emotions in the presence of caregivers' disapproval or in the presence of another's distress. To be able to react with negative feelings in those contexts requires not only the basic disposition of feelings of fear and sadness, but also the ability to"catch"others' emotions and to "catch"others' distress. Emotional dispositions are established by imitation. In our socialization process, we mimic perceived emotions (facial expressions and vocalizations), and eventually we copy (via imitation) the inner states of others, such as shame, guilt, and others' distress. Seeing others' distress triggers vicarious distress and, further in development, it triggers consolation responses. So, Prinz concludes, psychopaths are bad moralizers because they cannot learn the appropriate emotional reaction to their conduct in the context of their victims' distress, neither through imitative learning nor through emotional contagion.

However, in the following sections, I will argue that imitative learning cannot fully explain moral development. Imitative learning might explain recognition of basic emotions, but it cannot explain the development of moral emotions such as guilt, shame, regret, admiration and empathic concern. As I will argue later, we cannot "catch" those complex emotional reactions by "copying another's affective state" through imitative learning, even in the broad sense of imitation adopted by Prinz (2005). The intentional and motivational elements of those affective states are not available for direct perception and associative learning. A complete explanation for the development of moral emotions must involve empathy. If this is right, Prinz's hypothesis about psychopaths is at best incomplete.

This does not mean that we must abandon the dysfunctional fear hypothesis. Instead, we can use that hypothesis but add a role for empathy. It may be that impairment in fear[5] causes impairments in shared fear and in empathy, and these impairments in turn cause the moral deficit. It is clear that being able to feel emotions (such as fear, anger, sadness, joy, disgust, and surprise) is a prerequisite for sharing those emotions and for (emotional) empathy. So, impairment in feeling emotions (as stated in the dysfunctional fear hypothesis) will, necessarily, cause a deficit in shared emotions and in empathy. It is natural to suppose that this deficit is what leads to moral deficits in psychopaths (at least in a sentimentalist approach of morality). In effect, the hypothesis is that impaired empathy mediates the connection between impaired emotions and

[5]I will not talk about feelings of sadness, as very little is known about how sadness is affected in psychopathy.

impaired morality. This provides an alternative to Prinz's hypothesis that impaired imitative learning mediates the connection.

It is also arguable that Prinz's hypothesis cannot work unless empathy is given a key role. It is widely believed that psychopaths' moral impairments are especially tied to impairments in recognizing and responding to their victims' distress. These impairments are naturally explained in terms of impairments in sharing victims' distress, which can be seen as a form of empathic concern. Prinz's hypothesis requires that imitative learning alone can explain the recognition of others' distress. He argues that concern for the victim's distress is a metacognitive ability that emerges late in development and derives from early vicarious distress, which is a more basic ability used to catch others' distress via emotional contagion. However, in the following sections, I will argue that vicarious distress requires empathic abilities. It cannot be explained simply through emotional contagion.

It is true that there is a simple form of vicarious distress by emotional contagion in infants that does not require empathic abilities. However, this sort of early vicarious distress by emotional contagion happens before the development of full self-other differentiation. In early vicarious distress, the infant is not experiencing or recognizing others' distress; the infant is experiencing her own distress, which leads to personal distress and not to empathic concern. This rudimentary phenomenon cannot explain those elements of moral development that involve the recognition of others' distress. To explain that, one needs a more complex form of vicarious distress involving empathy. On my account, early vicarious distress in emotional contagion evolves first into empathic vicarious distress, and, then, eventually, to empathic concern. Empathy helps us to get information about the manner in which an event or an action might affect a person emotionally and cause others' distress.

My explanation of the role of empathy in connecting emotions and moral development in psychopaths fits well with recent research on psychopaths by Blair and others. Psychopaths show impairment in recognition of fear expressions (face, body, and voice) (Blair 2005; Marsh, Blair 2008), reduced experience of fear (Marsh et al. 2011), impairment of response to fear in others (Marsh, Cardinale 2012), and impairment of the ability to identify behavior that causes fear and in judging the moral acceptability of those behaviors (Marsh, Cardinale 2012). Also, psychopathy affects judgments of transgressions associated with harm. Psychopaths tend to err in treating conventional violations like moral violations, and they are less likely to justify their judgments by referring to the victim's welfare (Blair 1995; 2005). Their propensity to inflict harm to others indicates a profound disturbance in their empathic response to the suffering of others (Blair 2005). The ability to recognize others' distress is crucial for the experience of empathic concern (Nichols 2001). Any impairment in the early emotional recognition ability or an innate impairment in the ability to perceive and respond to the affective expressions of others will lead to a dysfunctional emotional empathy. As Blair suggests (2005), an individual that shows impairment in emotional empathy is difficult to socialize through empathy induction, a practice that involves the socializer focusing the attention of the transgressor on the distress of the victim. All this is further evidence for a role of empathy in explaining psychopaths' moral impairment.

Autism. Prinz suggests that experimental work (Blair 1996, 2005) shows that autistic people, unlike psychopaths, seem to both acquire an understanding of moral rules and exhibit a deficit of empathy. He concludes that if this interpretation is correct, "acquisition of moral competence *may* not depend on a robust capacity for empathy" (Prinz 2011b: 222). Kennett (2002), Nichols (2004), and McGeer (2008) have also argued that in autism the deficit of empathy does not inevitably lead to a deficit in morality. People with autism show a lack of empathy,

but they still have a sense of morality. According to Nichols (2004), autists' preserved ability to make moral judgment, despite their impairment in simulating another person's perspective, reveals that perspective-taking accounts of morality must be empirically wrong. From this evidence, Prinz (and also McGeer and Kennett) concludes that empathy is not necessary for the development of moral agency; if empathy plays any role in moral development, it plays an instrumental role, hence, a contingent one.

One way to resist this conclusion has been to show that while people with autism are impaired in cognitive empathy and mind-reading abilities, they are able to experience emotions, display affective empathy, and emotionally respond to others' distress. This suggests that their moral competence might derive from their emotional empathic abilities (Blair 1996; Nichols 2004).

The view that autistic people show morality without empathy has also been challenged by de Vignemont and Frith in "Autism, Morality, and Empathy" (2008). They challenge both ideas: that autistic people show a lack of empathy and that they show a sense of morality. They argue that autistic people have some degree of automatic emotional empathy: they show emotional recognition and autonomic responses to others'facial expressions of sadness and fear. Experimental work (Blair 2005) yields evidence that autistic people may have emotional components of empathetic behaviors. They are capable of displaying empathy toward the distress of others. Accordingly, while autism may involve impairment of cognitive empathy (the ability to know what another person thinks), some emotional empathy remains intact. The lack of empathetic behavior in autism has been attributed to deficits in mentalizing processes (Batson et al. 1987; Blair 2005). Despite showing preserved emotional empathy and preserved capacity of emotion recognition, studies based on parental reports suggest people with autism show specific impairments in their affective relatedness towards other people (Hobson et al. 2006). They clearly manifest signs of happiness, distress, anger, and fear as emotional responses to the moods of others, but they present limitations in experiencing and manifesting other-person-centered feelings, such as sympathy and concern; also, they rarely express feelings for and in relation to other people (Hobson et al. 2009). According to the reports, they show jealousy towards others and are affected by others' moods, but fewer show concern, guilt, or empathetic sadness. People with autism are more likely to describe situations in terms of breaking the rules rather than in terms of causing physical or emotional harm to others (Hobson et al. 2009).

De Vignemont and Frith (2008) suggest that the presence of the emotional component in people with autism may explain why they show apparently preserved moral competence. People with autism are able to detect the transgression of normative rules and to detect distress in others. Nevertheless, they do not seem capable of detecting moral violations. This detection requires correlating two facts: a moral transgression and someone's suffering without moral justification. People with autism seem to fail to correlate these two facts. De Vignemont and Frith (2007) also suggest that the problem with autistic people in detecting moral violations may be related to the way they make the distinction between allocentric and egocentric representations. People with autism display extreme egocentrism disconnected from allocentrism, meaning their social world is self-focused, they lack social intuitions and make abstract analyses of their surroundings, and "they are more interested in normative rules than in emotions due to an abstract allocentrism disconnected from egocentric interactions with others" (de Vignemont & Frith 2008: 280). Their conclusion is that we cannot rule out the possibility that the rules followed by autistic people are merely perceived by them to be conventional rules, and that their apparent capacity for moral judgment is the result of applying those conventional rules.

Although this conclusion cannot defeat Prinz's argument that autistic people are able to make moral judgments, it can offer an alternative interpretation to this phenomenon. First, emotional empathic abilities seem to be preserved in high-functioning autistic people, and this preserved ability might explain their ability to make moral judgments, despite their limitations in experiencing and manifesting empathic concern and offer comfort in the context of other's distress. Second, there is no strong evidence that their apparent capacity to make moral judgments is the result of applying moral rules or displaying moral concern.

Theories of Moral Development. Prinz's third argument is against developmental theories that emphasize the role of empathy in moral development. Developmental moral psychology describes how we evolve to become moral agents, how we come to distinguish between right and wrong, and how we learn the distinction between conventional and moral rules. Prinz's developmental story (2005) emphasizes the central role of imitation in learning to be emotionally responsive to moral judgments. He suggests that moral learning requires a different kind of imitation; children might "copy the inner states of others," and not just "their goal-directed behaviors." His main argument is that imitation helps us to acquire forms of moral comprehension. Our moral understanding involves a range of emotional capacities that depend on imitative learning to be acquired. Prinz describes five stages of normal moral development. In the first stage, infants experience the emotions of others via facial mimicry; moral responsiveness begins with emotional contagion in newborns. This stage contributes to the emergence of concern and empathy. In the second stage, toddlers engage in prosocial behavior and early consolation. In the third stage, children become responsive to moral rules. In the next stage, children begin to engage in reparative behavior and moral condemnation. In the last stage, children distinguish different classes of norms through the attainment of moral emotions (guilt, shame, outrage). According to Prinz (2005), imitative learning makes contributions to all those stages, and none of those forms of moral learning requires empathy. He concludes that acquisition of moral competence does not depend on a robust capacity for empathy.

There are at least two obscure points in Prinz's developmental moral story. First, he explicitly neglects well-known empathetic processes that emerge in human development and are fundamental for developing sociality and moral competence (2005). It is widely accepted that those processes contribute to the acquisition of moral competence. As I argued in a previous work (Rochat & Passos-Ferreira 2008), imitation and mirroring processes are necessary but not sufficient conditions for children to develop morality. Imitation provides the basic sense of social connectedness, including mutual acknowledgment of existing with others that are 'like me' (Meltzoff 2007). However, for morality to develop, imitation and mirroring processes need to be supplemented by an open system of reciprocation and shared representations (of emotions and other mental states). Imitation and emotional contagion decreases as the individual develops other cognitive capacities. Developmental research shows that from the second month, mimicry, imitative, and other contagious emotional responses are bypassed. Imitation gives way to signs of reciprocation and emotional co-regulation. As joint attention to objects develops, shared affective representations also emerge. Eventually an explicit moral sense develops, accompanying the emergence of mind-reading and imagination by age 4. Around age 5, children show explicit understanding of the mental states that drive others in their behaviors and beliefs, allowing children to understand the motivational aspects that trigger moral attitudes (Rochat & Passos-Ferreira 2008).

Second, Prinz's imitation story is only a partial story about understanding others' minds. Imitation and emotional contagion are just the foothold for understanding others as bearers of mental states. Further developments are needed for understanding other types of mental

states, which are far from purposive actions, desires, visual perception, and basic emotions. For these states, there is a relatively close coupling between the underlying mental states and their expression in bodily action. We can infer those states through perception and imitation, as Prinz argued. Empathy is not the only way to understand others. However, it is psychologically our most pervasive method for identifying mental states in others. It enables us to infer other mental states in a faster and more accurate way. Empathy allows us to make faster and more accurate predictions about other people's needs, their emotions, and the intentions of their actions. In addition to this, empathy is the only reliable mechanism for understanding the mental states of people to whom we do not have direct perceptual access and whose thoughts are not overtly expressed in their actions. It is especially relevant to grasping false beliefs, divergent beliefs, divergent affective and cognitive perspectives, and secondary moral emotions.

To moralize – that is, to think morally – depends on sharing others' affective states and taking others' affective perspective into account. Sharing, simulating, and imagining others' emotional states is necessary for developing secondary emotions, such as feelings of empathetic concern, shame, guilt, regret, resentment, outrage, and admiration. I argue that a basic empathetic mechanism is necessary to acquire secondary moral emotions. The mere capacities to imitate basic emotions (fear, anger, disgust, happiness, or sadness) or to copy the inner states of others are not enough for understanding and internalization of secondary emotions, which are fundamental components of our moral competence. For developing moral agency, we need a complex emotional regulatory system which is more sophisticated than mere imitation and emotional contagion processes. The empathic systems play this role (Rochat & Passos-Ferreira 2008). In the next section, I will suggest another developmental story that leads from imitation to perceptual and imaginative empathy.

4 From imitation to perceptual empathy and imaginative empathy

Empathy has been defined in a number of ways (Eisenberg & Strayer 1987; Eisenberg 2000; Batson 1998, 2011; Hoffman 2000; de Vignemont & Singer 2006; Decety & Jackson 2006). The term 'empathy' ("feeling as the other feels") and the associated term 'sympathy' ("feeling concern for the other") have been used to refer to a wide family of psychological processes. To define empathy, it is important to distinguish it from a variety of other phenomena, such as emotional contagion, sympathy, mental projection, and empathic concern. Emotional contagion is a phenomenon whereby an emotion is automatically spread from one individual to another, and it is characterized by self-other non-differentiation (e.g., a baby that begins yawning when she sees another baby yawning). In contrast, empathy implies self-other differentiation. Mental projection is a mental process in which we put ourselves in the other's position in order to understand them through simulation (Goldman 2006; Decety 2004), mind-reading (Gopnik & Meltzoff 1997), or perspective-taking. Sympathy is characterized by participating in an emotion experienced by another. It involves feeling concern, sharing suffering with others, and seeking their well-being. Empathic concern, as defined by Batson (2011: 11), is "*an other-oriented emotion elicited by and congruent with the perceived welfare of someone in need*"; it includes empathic emotions, such as feelings of sympathy, compassion, sadness, distress, and concern.

I will adopt here Nancy Eisenberg's widely accepted conceptualization of empathy. According to Eisenberg and Strayer (1987), empathy involves sharing the perceived emotion of another; it is a vicarious affective reaction that "may occur as a response to overt perceptible

cues indicative of another's affective state (e.g., a person's facial expression) or as the consequence of inferring another's state on the basis of indirect cues (e.g., the nature of the other's situation)"(Eisenberg & Strayer 1987, p.5).

Traditionally, psychologists distinguish between two psychological processes involved in empathy: *emotional empathy* (vicarious sharing of emotion) and *cognitive empathy* (mental perspective taking) (Smith 2006; Davis 1983; Hoffman 1977). Cognitive empathy involves cognitive perspective taking of the thoughts and beliefs of others. Emotional empathy involves sharing affective states with another person. The different ways of conceptualizing empathy focus on one or another of those two components. Some researchers focus on the emotional aspects of empathy, while others focus on the intellectual process of inferring others' mental states. Psychologists and philosophers distinguish those processes using narrow and broad definitions of empathy. The narrow definition tries to capture empathy in its most basic form, identifying it with emotional contagion, as an automatic process of affective resonance. The broad definition describes empathy as a multidimensional phenomenon which combines both processes (affective and cognitive) involved – or a set of processes as proposed by Davis (1983) – as they emerged in early development[6].

Prinz defines empathy narrowly[7]. He defines empathy as a vicarious emotion that involves"feeling what one takes another person to be feeling"(Prinz 2011b; 215). According to his account, empathy is "a matter of feeling an emotion that we take another person to have" as a response to an automatic contagion or the result of an exercise of the imagination (Prinz 2011b: 215). In his sentimentalist account, Prinz emphasizes the perceptual and emotional aspects of empathy and downplays the rationalist and intellectualist notions of empathy that emphasize the role of imagination, simulation, and mind-reading. However, Prinz affirms that empathy is not always an automatic process in the way that emotional contagion is; "sometimes imagination is required, and sometimes we experience emotions that we think someone would be experiencing, even if we have not seen direct evidence that the emotion is, in fact, being experienced" (Prinz 2011a: 212). Nevertheless, Prinz claims that imagination is "overly intellectual"and"a mental act that requires effort on the part of the imaginer" (Prinz 2011a: 212).

As Prinz notes, his definition of 'empathy' is similar to the definition of 'sympathy' used in the tradition of moral philosophy, including David Hume and Adam Smith. However, emotional empathy alone may not play the crucial role required for a sentimentalist account of morality. Even for sentimentalists like Hume and Smith, moral approval and disapproval involve impartially placing oneself in the perspective of the person affected and sharing their emotions and reactions. To understand the role (if any) that empathy plays in morality – at least according to sentimentalists – we should adopt a broader conception of empathy that includes

[6]The multidimensional approach of empathy has been suggested by different studies. Hogan (1969) and Davis (1983) have suggested a scale of empathy (either cognitive or emotional) in which empathy is considered as a set of constructs that all concern responsivity and sensitivity to others. To some psychologists, empathy is a unitary process that includes a class of phenomena – such as emotional contagion, sympathy, personal distress, and cognitive perspective-taking – that share the same mechanism. Hoffman (1977, 2000) suggests a unitary account where the ontogenetic development of empathy starts from birth with global empathy (emotional contagion) leading to the emergence of egocentric empathic distress by 14 months, and the emergence of veridical empathy in the second half of the second year when children fully differentiate between self and other. Conversely, Blair (2005) claims that the term "empathy" subsumes a variety of different and dissociable neurocognitive processes, varying from emotional empathy and perceptual empathy to cognitive empathy.

[7]De Vignemont and Singer (2006) also suggest narrowing down the concept of empathy. However, they argue for the exclusion of the automatic component as part of its definition. They define empathy as a conscious affective state, isomorphic to another person's state, that is elicited by observation or imagination of another person's affective state.

both emotional empathy (sharing of emotions) and cognitive empathy (affective perspective-taking).

The broad definition (including both cognitive and affective empathy) specifies the content of empathy as a reaction to the observed experiences of another that is shared (sharing cognitive states and sharing emotional states). However, the distinction between cognitive empathy and emotional empathy does not capture all the processes involved in sharing emotions. I suggest an additional distinction that focuses not on the empathic reactions (cognitive or affective), but on the underlying psychological mechanisms necessary to access others' affective states. I will distinguish empathy as a response to a direct perception of others' emotions – which I call *perceptual empathy* – and empathy as a response to imaginative or projective simulation of others' affective perspective – which I call *imaginative empathy*. These mechanisms help explain how emotional empathy develops into a more sophisticated emotional state that allows us to directly perceive or imagine or simulate others' emotional states. These two processes are part of a continuum of empathetic processes that emerge in early human development of the ability to understand and identify another's emotional state. As cognitive abilities develop, there is an ontogenetic chain of processes leading from mimicry and emotional contagion to empathy, sympathy, compassion, and perspective-taking. This distinction helps us to understand the developmental basis of the connection between empathy and morality.

In early development, we distinguish different levels of empathy. This process starts from birth via neonatal imitation and emotional contagion, and leads to the capacity to mimic and resonate with other's emotional states. Later, an understanding of others' emotional states and intentions develops, along with affective perspective taking via joint attention, simulation and imagination. The ability to understand and respond to another's emotional state appears in the very beginning of an infant's development and increases to complex levels of empathy over time.

Elisabeth Pacherie (2004) suggests three degrees of empathy and of their respective psychological mechanisms, on a continuum going from imitation and emotion contagion to perceptual and imaginative empathic processes, covering different stages of child development. In each stage of ontogeny, children develop empathic abilities corresponding to the understanding of three aspects of others' mental states: 1) the type of emotion experienced by others, 2) the situation that is causing the specific emotion experienced by others, and 3) the motivational factors triggered by the emotion. The three degrees of empathy are the ability to identify an emotion, the ability to understand the intentional object of the emotion, and the ability to understand the connection between the type of emotion, its intentional object, and the motivational factors triggered by the emotion. In this respect, my developmental proposal can be seen as an elaboration of Pacherie's account.

The first level is emotion recognition, which is the ability to *identify the type of emotion* experienced by others. How does our capacity to use perceptual clues to understand the emotion experienced by others emerge? This level starts with early imitation of facial and vocal expressions in newborns. According to Meltzoff (1977), newborns can equate their own unseen behaviors with gestures and facial expressions they see others perform. Facial imitation suggests an innate mapping between observation of another's expression and execution of a motor action. In imitation, there is an automatic correspondence between the visual information of the observed facial expression and the proprioceptive information of the motor representation. When a baby imitates a facial expression, her imitation is based on a motor representation formed when she is observing another's expression. In early imitation there is a correspondence between observing an expression, adopting a facial expression or a body posture, and feeling

the corresponding emotion (Meltzoff 1977). Newborns' facial mimicry leads to emotional contagion through facial and vocal feedback. By two months, infants engage in face-to-face proto-conversations, reciprocating with others in what amounts to a process of emotional co-regulation and affective attunement (Rochat & Passos-Ferreira 2008). Imitation and emotional contagion are based on two distinct processes: a direct connection between perception and action and a direct connection between proprioceptive perception and facial expression (Pacherie 2004). Early imitation and emotional contagion always involve proprioception – an awareness of our body's movements and positions – but do not involve an explicit self-other distinction.

Unlike imitation and emotion contagion, empathy emerges when an infant becomes aware of self-other distinction. In early development, specific cognitive functions emerge that allow infants to distinguish emotional contagion – which involves no awareness of self-other distinction – and empathy – which involves awareness of self-other distinction. As Pacherie (2004) points out, the first level of empathy involves the emergence of a direct connection between the evoked motor representation and the emotional experience *without having to necessarily go through the proprioceptive stage*, i.e., without the corresponding imitation of others' expression. In the early form of perceptual empathy, infants have perceptual access to another's emotional state through facial gestures and vocal expressions without necessarily forming a motor representation through proprioception. This allows infants to distinguish between feeling their own emotions, observing the same emotions in others, and sharing others' affective states.

The second level of empathy is the ability to *understand the object of the emotion*. At this level, the subject identifies the relationship connecting another's emotion with a given situation. This ability emerges with the development of joint attention processes, social references, and intentional communication. With the emergence of the drive to co-experience events and objects in the environment with others, by nine months, babies start learning and developing shared meanings about events and objects and understanding the intentions of others' behaviors. The meaning of a perceived event (e.g., whether something is dangerous or threatening or disapproved) is now referred to through others' emotional responses; to some extent, it is evaluated in relation to others (Rochat & Passos-Ferreira 2008). This cognitive capacity allows the subject to understand others' affective states. In joint attention and social reference processes, when a child observes an object or an event that is the focus of her mother's attention, the child treats the mother's emotions and her facial and vocal expression as a commentary on the object or event. We interpret another's emotions as a commentary and an appraisal of situations and events, which gives us information about the environment. Such processes allow the child to understand the causal role of emotions and understand the motivations of others' affective reactions. As Pacherie (2004) points out, in becoming referential, toddlers develop access to agents' motivations and develop the ability to identify the immediate intentions of the agent by observing the way she behaves. Our intentions are reflected in our body movements, and the mere observation of an action performed by others allows us to detect others' intentions and motivational states.

In the first two levels of empathy, there is a direct connection between perception and action, which allows the subject to identify the type of the emotion and to understand the intentional object associated to the observed emotion (Pacherie 2004). This form of empathy as *direct perception* I call *perceptual empathy*. Perceptual empathy plays a crucial role in situations where the subject has perceptive cues that allow direct access to the type of emotion and its intentional object through perceptual mechanisms. It allows the subject to understand others' mental states, such as goals, attitudes, motivations, and affective states, and to identify the situation that is causing others' emotions.

The third level of empathy involves the ability *to understand the correlation between the type of emotion, its intentional object, and its motivational factors*. This form of empathy relies on simulation and imaginative capacities. By the age of two, children start engaging in elaborate games of imagination and symbolic pretense in which objects and actions in the actual world are taken to stand for objects and actions in a realm of make-believe. They start imagining hypothetic situations and creating imaginative characters. This capacity increases once additional cognitive abilities emerge during child development. Children progressively acquire imaginative flexibility and the ability to simulate others' cognitive and affective perspective. This more elaborate form of empathy – which I call *imaginative empathy* – is necessary when the situation provides the observer with no transparent or direct access to others' mental states. In opaque contexts, emotions are not overtly expressed, and the motivational aspects may differ from our own motivations in similar contexts. In the early stages of empathy, imitation and emotional contagion processes involve mainly basic emotions (e.g., happiness, fear, sadness, anger, surprise, and disgust), which are characterized by universal facial expressions that the subject can have direct perception and transparent access. In imaginative empathy, imagination and mental simulation are fundamental mechanisms that allow the subject to understand secondary emotions (social and moral) and to infer their motivational potential. According to Pacherie (2004), in *transparent contexts*, both forms of perceptual empathy – identification of the type of emotion and understanding the connection between emotion and its intentional object – can emerge from perceptual mechanisms that establish a direct connection between perception and action. In transparent contexts, we can overtly perceive cues that indicate another's affective state (e.g., a person's facial expressions or body gestures). However, in *opaque contexts*, in the absence of perceptible clues, we must rely on imaginative empathy to grasp the ternary connection between the type of emotion, its intentional object, and the motivational factors triggered by the emotion.

Throughout most of our lives, we are involved in opaque contexts where we need imaginative empathy and mental simulation to understand and infer others' emotional states. Empathy, defined as this capacity to understand via perception or imagination the type of emotion and the connection between emotion, motivational aspect, and intentional object, is essential for moral development. The capacity to express moral attitudes involves the capacity to understand and identify secondary emotional reactions like guilt, shame, contempt, regret, admiration, outrage, and concern. Imaginative empathy plays a central role in understanding those affective reactions and allows us to internalize those emotional reactions as we imagine or simulate them based on others. We can experience, for example, feelings of shame, guilt, regret, admiration, or outrage in certain circumstances, because we can place ourselves in the shoes of those primarily affected by the action and share their reactive attitudes. This is the way children come to understand and internalize moral rules and moral attitudes.

According to this conception, empathy involves mental simulation and imagination of others'feelings, imagination of how others perceive our actions, and imagination of whether or not they approve of us. The internalization of imagined feelings and the simulation of others' affective perspective is crucial for the development of a moral agent capable of following moral rules and behaving morally. The characteristic features of a moral agent depend on being able to arrive at moral attitudes as a result of a process of empathic simulation and affective perspective-taking. As I have argued, imitation and emotional contagion are only the first step of this process. The emergence of perceptual and imaginative empathy is required to develop the capacity to think morally.

5 Is empathy beneficial for morality?

Should empathy play a role in morality? Should we cultivate empathy in morality? According to Prinz (2011a, 2011b), empathy-based morality is harmful to society. He argues that empathetic emotions may lead to inaccuracies in our moral judgments and do not contribute to any good moral practices. An empathy-based morality has many limitations as a guide for moral motivation. Empathy can lack motivational force in driving prosocial behavior and altruistic actions; it is vulnerable to bias and tends to be highly selective. It can also lead to preferential treatment and crimes of omission. According to Prinz, if empathy produces biases in moral judgment and interferes negatively with morality, then it should be avoided as a guide for morality. Prinz concludes that empathy should be discouraged as the central motivational component of a moral system.

Prinz's prescriptive argument might be reconstructed like this:

P4. Empathy produces biases in moral judgment and interferes negatively with morality.

P5. If empathy produces biases in moral judgment and interferes negatively with morality, it should be avoided as a guide for morality.

C2. Hence, empathy should be avoided as a guide for morality.

Premise P4 is based on experimental work that suggests that empathy is harmful and produces biases in moral judgment. Those results lead Prinz to the conclusion that empathy should not be cultivated. I argue against this prescriptive conclusion by rejecting the second premise P5.

The argument derives a prescriptive conclusion (empathy should be avoided) from an epistemic premise (empathy biases moral judgment). Similar arguments have been defended by Holton and Langton (1999) and Struchiner (2011). They emphasize the limitations and distortions empathy could bring to morality. Holton and Langton (1999) worry about relying on imaginative identification as an epistemic tool for morality. One of their main concerns is that empathy leads to parochialism. Empathy is prone to parochialism because it occurs more readily vis-à-vis individuals who are salient, currently perceived, and spatially closer to us or bear resemblance to us (Goldman 2006). As noted by Hoffman (2000), though we empathize with almost anyone in distress, it is easier to empathize with those like us.

One objection proponents of empathy might raise to Prinz's argument is that empathy might be improved by combining it with additional epistemic tools and helpful devices. Goldman (2006) shows that empathy, as a process involving imagination and simulation, can be enhanced by perceptually derived information to generate more accurate representations of an anonymous and distant individual and to transcend the parochialism of a self-affective perspective. Prinz suggests that improvements of this sort might lead empathy to play an inert causal role in this process and that it could easily be replaced by other emotions such as anger and outrage.

Struchiner (2011) also argues that empathy is not necessary for moral judgments, moral development, or moral motivation. Following Prinz, he claims that empathy is a 'dangerous' emotion that leads to acts of cruelty and injustice, and it should be avoided, or even eliminated, in legal systems. Struchiner argues that an empathy-based decision-making model in legal systems can result in errors, distortions, and abuses. He argues that empathy is as potentially harmful to legal decisions as it is to morality. On that basis, he concludes that a rule-based

decision-making model captures the essence of law, and a good decision-making model for legal decisions should overrule any empathic component[8].

Struchiner proposes that legal systems should be guided by reasoning based on an autistic perspective. He defends what he calls 'the contingent morality of autistic rule-based decision-making.' The pivotal idea here is that people with autism show the right virtues for a good model – the virtues of rules. They love systematizing, they are rigid rule followers, and they take seriously the literalness in which rules are formulated. Struchiner suggests that the legal system should embrace these virtues. Autistic thinking, from which empathy is absent, would produce less bias in moral judgments and less distortion in legal decisions.

Clearly, there is a problem with this characterization of the autistic mind. As I argued before, the capacity for following rules and detecting normative transgressions that characterize autistic reasoning do not result in a capacity for detecting moral transgression. Consequently, the presence of those abilities does not result in moral competence in people with autism. Furthermore, even if we concede that we might find autistic minds as Struchiner describes them – in which moral competence relies on following normative rules – this will not vindicate the rule-based legal system. Even if such a system works better in certain circumstances, we still have to consider situations in which legal systems cannot make decisions based only on following rules. There are situations that necessarily require our imaginative power and capacity for taking the affective perspective of those affected by the action – e.g. situations involving moral conflicts are one of those circumstances. In these situations, legal systems should be able to transcend distortions of egocentric perspective views. Judges should reason as disinterested participants that can take the perspective of those affected and make decisions based on imaginative flexibility.

There are two implicit ideas in Prinz's account. The first idea is that if empathy is not necessary for all kinds of moral judgment, then empathy is not necessary for morality; it could therefore be systematically substituted by other emotions. The second idea is that morality is nevertheless based on a single kind of emotion. Prinz suggests that *outrage* might be the kind of emotion that would play a central role in morality, because it has more motivational power than empathy, and it is less susceptible to bias. According to Prinz (2011a, 2011b), we should cultivate an outrage-based morality[9].

According to the view I defend, empathy is a critical feature of moral development. Important aspects of our morality are related to empathetic feelings. Furthermore, empathy should be cultivated. The fact that empathy produces distortion and bias does not imply that it is not beneficial to morality overall. Empathetic emotions can distort our perception and are prone to self-perspective distortions, but this is true for other emotions too, such as fear, shame, regret, admiration, guilt, outrage, anger, and jealousy. Outrage, for instance, can lead to lynching, a collective violence in which a group punishes an individual that transgressed moral rules.

Our emotional reactions can distort our view and interfere negatively with our moral judgment, leading to incorrect judgments and morally wrong actions. We can be 'misguided'

[8]Struchiner's rule-based model of decision-making is based on Frederick Schauer's conception of rules, which is itself a response to Ronald Dworkin's arguments against legal positivism in "The model of rules" (1967).

[9]For example, Prinz (2011a) affirms, "We should rage against the wrong. (…) From a practical perspective, we might be best off trying to cultivate a sense of outrage for injustice wherever it occurs and a sense of joy in helping the needy wherever they may be. The assumption that empathy is essential for these ends may be mistaken, and efforts to expand our moral horizons by empathetic induction may make us more vulnerable to errors of allocation." And in a criticism of feminist ethics (2011b), he affirms "a feminist morality bent on liberation should not be an empathy-based morality if that label is meant to describe a morality that makes empathy into the primary emotional resource. An outrage-based morality might be more effective."

by our affective reactions (Goldie 2002). Emotional reactions can lead to misunderstandings in moral judgment. This shows that we must not rely on our affective reactions as the only source of our moral attitude. Moralizing involves transcending our egocentric affective perspective (Rochat & Passos-Ferreira 2008) and taking the affective perspective of those affected by the action into account.

If we exclude all emotions that could lead to distortions, limitations, or biases in moral judgment, moral sentimentalists will be left with very little to count as a positive guide for moral approval and moral disapproval. Sentimentalists need emotions to distinguish right from wrong. For non-sentimentalists, excluding these emotions will be less problematic. For example, Kantians and other moral rationalists can rely on reasoning abilities to develop moral competence (as Kennett (2002) suggests). However, sentimentalists (like Prinz and myself) have no alternative but to rely on affective dispositions, including some biased emotions.

On the sentimentalist view that I advocate, empathy is a crucial element of morality with moral motivational force. In some specific circumstances, it is our best guide for morality. The reason is that empathy allows us to transcend our egocentric-affective perspective and to simulate the perspectives of those affected by an action. Transcending our own perspective and taking the affective perspective of those affected by an action is required for moral judgment, and empathy often provides the best way to access the perspectives of others. This access may be imperfect, but, nevertheless, it is highly beneficial overall.

A similar perspective has been defended by some advocates of the ethics of care. This tradition has helped to emphasize the role of empathy-related emotions. As Virginia Held (2006) notes, the ethics of care values emotions, such as empathy and sympathy, as an epistemic tool to ascertain what morality recommends. The ethics of care also rejects the idea that you should favor abstract reasoning and impartiality to avoid bias and arbitrariness (Held 2006). Empathy plays a key role in contexts of caring and helping for individuals who cannot express their emotions, desires, and beliefs.

Furthermore, imaginative empathy is often beneficial in these cases of caring. In "Who Needs Empathy?" Coeckelbergh (2007) analyzes the decision-making process in an intensive care unit for babies, and he shows that imaginative empathy plays a central role in decisions about the lives of those who cannot express themselves in a transparent manner. Health care units for babies are opaque contexts where we cannot rely exclusively on perceptive cues and perceptual empathy to decide what is morally right or wrong. We have to use our imagination and simulate babies' internal affective perspective to make decisions. We should exercise empathy with people who suffer and are not able to express their desires and concerns. They appeal to our imaginative powers and they might want us to share their vulnerability and suffering as fellow humans. As Coeckelbergh (2007) affirms: they appeal to "an imaginative effort on the part of the helper to imagine *what it is like to be the other person* by taking the internal perspective (imagine what it would be like to be the sufferer) and by the communication of this imaginative effort."(2007: 69).

In our society, caring for those who lack autonomy, including those who never developed autonomy and those who are temporarily unable to act as autonomous moral agents, is a central value. As part of our morality, we approve of the practice of caring for those who, for various reasons, are not able to express their concerns and desires. This group includes babies, people with disabilities, and patients in terminal states who have lost consciousness or bear other mental illnesses. In these situations, empathetic abilities are essential in avoiding the distortions from taking our own perspective or applying rigid rules that might potentially produce errors and injustices. In cases where we must provide aid and care to those whose affective experience

is opaque to us, a lack of imaginative empathy will lead to distortions in moral judgment and also in legal decisions. In these cases, empathy is morally beneficial.

6 Conclusion

I have argued that empathy is a crucial element in morality and that in some specific circumstances it is our best guide for morality. I have argued against two theses of Prinz's antiempathic sentimentalism. I have argued against his developmental thesis, which says that empathy is not necessary for moral development. I have also argued against Prinz's normative thesis, which says that empathy should be avoided as a guide for morality.

To think morally, we need to transcend our egocentric affective perspective in order to correct the limitations and distortions of this perspective. We can do this by sharing affective states and imagining the reactions of those affected by our actions. In this way, empathy serves as a positive guide in moral judgment.

References

Batson, D. C. et al. (1981). 'Is empathic emotion a source of altruistic motivation?', *Journal of Personality and Social Psychology* **40**: 290—302.

Batson, D. C., Fultz, J. & Schoenrade, P. A. (1987). Adults emotional reactions to the distress of others, *in*: Eisenberg, N. & Strayer, J. (eds.), 'Empathy and its development', Cambridge University Press, Cambridge, 163—184.

Batson, D. C. & Shaw, L. L. (1991). 'Evidence for altruism: Toward a pluralism of prosocial motives', *Psychological Inquiry* **2**: 107—122.

Batson, D. C et al. (1997). 'Is empathy-induced helping due to self-other merging?', *Journal of Personality and Social Psychology* **73**: 495—509.

Batson, D. C. (2011). *Altruism in humans*, Oxford University Press, New York.

Blair, R.J.R. (1995). A cognitive developmental approach to morality: investigating the psychopath. Cognition 57 (1995) 1-29.

Blair, R. J.R. (1996). 'Brief report: Morality in the autistic child', *Journal of Autism and Developmental Disorders* **26.5**: 571—579.

Blair, R.J. R. (2005). 'Responding to the emotions of others: dissociating forms of empathy through the study of typical and psychiatric populations', *Consciousness and Cognition* **14**: 698-718.

Blair, R.J.R. et al. (2005). *The psychopath: emotion and the brain*, Blackwell Publishing, Oxford.

Coeckelbergh, M. (2007). 'Who needs empathy? A response to Goldie's arguments against empathy and suggestions for an account of mutual perspective-shifting in contexts of help and care', *Ethics and Education* **2(1)**: 61—72.

Copp, D. (2011). Jesse Prinz, The Emotional Construction of Morals (Oxford: Oxford University Press, 2007): Prinz's Subjectivist Moral Realism. *Noûs* **45(3)**, 577—594.

Darwall, S. (1998). 'Empathy, sympathy, care', *Philosophical Studies* **89.2/3**: 281—282.

Davis, M. (1983). 'Measuring individual differences in empathy: evidence for a multidimensional approach', *Journal of Personality and Social Psychology*, **44(1)**: 113—126.

Decety, J. (2004). L'empathie est-elle une simulation mentale de la subjectivité d'autrui?, *in*: Berthoz, A. & Jorland, G. (org.), 'L'Empathie', Odile Jacob, Paris, 53—88.

Decety, J. & Jackson, P. L. (2006). 'A social-neuroscience perspective on empathy', *Current Directions in Psychological Science* **15(2)**: 54—58.

De Vignemont, F. & Singer, T. (2006). 'The empathic brain: how, when and why?', *Trends in Cognitive Sciences* **10(10)**, 435-441.

De Vignemont, F. & Frith, U. (2008). Autism, morality and empathy, *in*: Sinnott-Armstrong, W. (ed.), 'The neuroscience of morality: emotion, brain disorders, and development', MIT Press, Cambridge MA, 273—280.

Dworkin, R. (1967/1977). The model of the rules I, *in*: Dworkin, R. (1977) 'Talking rights seriously', Harvard University Press, Cambridge MA.

Eisenberg, N. & Strayer, J. (1987). Critical issues in the study of empathy, *in*: Eisenberg, N. & Strayer, J. (eds.), 'Empathy and its development', Cambridge University Press, Cambridge, 3—13.

Eisenberg, N. (2000). 'Emotion, regulation, and moral development', *Annual Review of Psychology* **51**: 665—697.

Fowles, D. C. (1988). 'Psychophysiology and psychopathy: a motivational approach', *Psychophysiology* **25**: 173—391.

Goldie, P. (2002). 'Can we trust our emotions?' *Richmond Journal of Philosophy* **1(1)**: 27—35.

Goldie, P. (2006). Anti-empathy: against empathy as perspective shifting, *in*: Goldie, P. & Coplan, A. (eds.), 'Empathy: Philosophical and psychological perspectives', Oxford University Press, New York.

Goldman, A. (2006). *Simulating minds*, Oxford University Press, New York.

Gopnik, A. & Meltzoff, A. (1997). *Words, thoughts, and theories*, MIT Press, Cambridge MA.

Haidt, J. (2012). *The righteous mind*, Pantheon, New York.

Held, V. (2006). The ethics of care, *in*: Copp, D. (ed.) 'The Oxford Handbook of ethical theory', Oxford University Press, New York, 537—566.

Hobson, R.P., Chidambi, G. Lee, A., & Meyers, J. (2006). 'Foundations for self-awareness: an exploration through autism', *Monographs of the Society for Research in Child Development Serial* 71 (2 Serial No. 284).

Hobson, et al. (2009). 'Anticipatory concern: A study in autism' *Developmental Science* **12(2)**: 249—263.

Hoffman, M. (1977). Empathy: its development and prosocial implications, *in*: Keasey, C.B. (ed.), 'Nebraska Symposium on Motivation Vol. 25', University of Nebraska Press, Lincoln, 169—218.

Hoffman, M. (2000). *Empathy and moral development: the implications for caring and justice*, Cambridge University Press, Cambridge.

Hoffman, M. (2011). Empathy, Justice and the Law, *in*: Coplan, A. & Goldie, P. (eds.), 'Empathy: Philosophical and psychological perspectives', Oxford University Press, Oxford, 230—254.

Hogan, R., (1969). 'Development of an empathy scale', *Journal of Consulting and Clinical Psychology* **33**: 307—316.

Holton, R. & Langton R. (1999). Empathy and animal ethics, *in*: Jameson, D. (ed.), 'Singer and his critics', Blackwell, Oxford, 209—232.

Hume, D. (1739/1978). *A treatise of human nature*, Nidditch, P.H. (ed.). Oxford University Press, Oxford.

Kennett, J. (2002). 'Autism, empathy and moral agency', *The Philosophical Quarterly* **52**: 340—357.

Maibom, H. (2009). 'Feeling for others: Empathy, sympathy, and morality', *Inquiry* **52**: 483—499.

Marsh, A. A., & Blair, R. J. (2008). 'Deficits in facial affect recognition among antisocial populations: A meta-analysis', *Neuroscience and Biobehavioral Re- view* **32**, 454—465.

Marsh, A. A. et al. (2011). 'Adolescents with psychopathic traits report reductions in physiological responses to fear', *J. Child Psychol. Psychiatry* **52**, 834—841.

Marsh, A. A., & Cardinale, E. M. (2012). Psychopathy and fear: specific impairments in judging behaviors that frighten others, *Emotion* **12**, 892—898.

McGeer, V. (2008). Varieties of moral agency: lessons from autism (and psychopathy), *in*: Sinnott-Armstrong, W. (ed.), 'The neuroscience of morality: Emotion, brain disorders, and development', MIT Press, Cambridge MA, 227-—258.

Meltzoff, A. N. & Moore, M. K. (1977). 'Imitation of facial and manual gestures by human neonates'. *Science* **198**: 75—78.

Meltzoff, A. N. (2007). 'The 'like me' framework for recognizing and becoming an intentional agent', *Acta Psychologica* **124(1)**: 26—43.

Nichols, S. (2001). 'Mindreading and the cognitive architecture underlying altruistic motivation', *Mind and Language* **16(4)**:425-455.

Nichols, S. (2004). *Sentimental rules: On the natural foundation of moral judgment*, Oxford University Press, Oxford.

Pacherie, E. (2004). L'empathie et ses degrés, *in*: Berthoz, A. & Jorland, G. (org.), 'L'Empathie', Odile Jacob, Paris.

Patrick, C. (2005). *Handbook of psychopathy*, The Guilford Press, New York.

Prinz, J. (2005). Imitation and moral development, *in*: S. Hurley and N. Chater (eds.), 'Perspectives on imitation: From cognitive neuroscience to social science (Vol 2: Imitation, human development, and culture)', MIT Press, Cambridge MA, 267-283.

Prinz, J. (2007). *The emotional construction of morals*, Oxford University Press, New York.

Prinz, J. (2011a). Is empathy necessary for morality?, *in*: P. Goldie & A. Coplan (eds.), 'Empathy: Philosophical and psychological perspectives', Oxford University Press, New York, 211—239.

Prinz, J. (2011b). 'Against empathy', *Southern Journal of Philosophy* **49**: 214—233.

Rochat, P. & Passos-Ferreira, C. (2008). From imitation to reciprocation and mutual recognition, *in*: J. Pineda (ed.), 'Mirror neurons systems. The role of mirroring processes in social cognition', Humana Press, New York, pp. 191—212.

Slote, M. (2010). *Moral sentimentalism*, Oxford University Press, Oxford.

Smith, A. (1759/2009). *The theory of moral sentiment*, Penguin Books, New York.

Smith, A. (2006). 'Cognitive empathy and emotional empathy in human behavior and evolution', *The Psychological Record* **56**: 3—21.

Struchiner, N. (2011). No empathy towards empathy: Making the case for autistic decision-making, *in*: 'The Nature of Law: Contemporary Perspectives', based on the so-named paper presented at McMaster University Philosophy of Law Conference "The Nature of Law: Contemporary Perspectives", May 13-15, 2011, in Hamilton, Ontario (Canada).

Tomasello, M. & Vaish, A. (2013). 'Origins of human cooperation and morality', *Annual Review of Psychology* **64**: 231—255.

Vaish, A., Carpenter, M. & Tomasello, M. (2009). 'Sympathy through affective perspective taking and its relation to prosocial behavior in toddlers', *Developmental Psychology* **45(2)**: 534—543.

Vaish, A. & Warneken, F. (2011). Social-cognitive contributors to young children's empathic and prosocial behavior, *in*: J. Decety (ed.), 'Empathy: From bench to bedside', MIT Press, Cambridge.

Zahn-Waxler, C. & Radke-Yarrow, M. (1990). 'The origins of empathic concern', *Motivation and Emotion* **14(2)**: 107—130.

Zahn-Waxler, C. & Robinson, J. (1995). *Empathy and guilt: early origins of feelings of responsibility, in: Tangney, J.P. & Fischer, K. (eds.), 'Self-Conscious Emotions', Guilford, New York, 143—173.*

abstracta Linguagem, Mente & Ação ISSN 1807-9792 Volume 8, Number 2, pp. 53–66, 2015

Does Same-Level Causation Entail Downward Causation?

Neil Campbell

Philosophy Dept., Wilfrid Laurier University,
Waterloo, ON, N2L 3C5
necampbe@wlu.ca

Abstract

I argue that Jaegwon Kim's supervenience argument does not generalize to all special science properties by undermining his central intuition, employed in stage one of the argument, that there is a tension between horizontal causation and vertical determination. First, I challenge Kim's treatment of the examples he employs to support this intuition, then I appeal to Kim's own work on the metaphysics of explanation in order to dissipate the alleged tension.

A number of philosophers have discussed the concern that Jaegwon Kim's supervenience argument (1997; 1998; 2005) generalizes, thereby threatening to undermine the causal efficacy of *all* special science properties. For those who are attracted to a supervenience-based layered model of the world and who take seriously the idea that there are properties performing causal work on different levels, Kim's argument is clearly a threat. After all, Kim claims that the first stage of the supervenience argument shows that "*level-bound causal autonomy is inconsistent with supervenience or dependence between … levels*" (Kim, 2005: 40). Discussions about whether or not the supervenience argument generalizes to other special science properties have focused primarily two points. Kim's critics have tried to show either that the fact the argument generalizes shows the argument is absurd because this implies causal drainage (Block, 2003), or else that Kim's suggestions for how to prevent causal drainage (Kim, 1997; 2003; 2005) are unsuccessful (Noordhof, 1999; Gillett and Rives, 2001; Bontly, 2002). Few, however, have critically examined the first stage of the supervenience argument itself. The reason for this, I suspect, is that the supervenience argument targets nonreductive physicalism, and most nonreductive physicalists are content to grant Kim his stage one conclusion that same-level mental-to-mental causation requires downward mental-to-physical causation. However, for those who are worried about the broader implications of the supervenience argument this failure to assess the first stage of the argument is an important oversight. As we shall see, there are grounds to reject the first stage of Kim's argument in which case the proponent of the layered model can avoid the conclusion that same-level causation *entails* downward causation. This means the supervenience argument might not generalize after all and that advocates of the layered model need not worry about causal drainage.[1]

[1]Of course, this strategy is unavailable to the nonreductive physicalist, and so my argument will not constitute a defense of nonreductive physicalism. This might make my discussion seem like a side issue but to characterize it as such is a mistake. For those who endorse the layered model and who conceive of levels in terms of mereological supervenience it is surely very important to preserve level-bound causal autonomy.

My discussion is divided into four parts. In part 1, I describe the tension Kim claims is at the heart of the first stage of the supervenience argument in the form of Edwards' dictum. In part 2, I outline the first stage of the argument and focus on the role that Edwards' dictum plays within it. In part 3, I attempt to dispel the alleged tension exploited in the first stage of the argument. My strategy is twofold. First, I identify some troubling features of Edwards' original example and an important disanalogy between what Edwards describes and the metaphysical framework of the supervenience argument. These observations raise some initial doubts about the reality of the tension Kim identifies in the first stage of his argument.[2] Second, I argue that the metaphysical and explanatory principles that might be used to motivate the tension actually have the opposite effect: they either undermine the tension, or else point to a way of relieving the tension that does not require downward causation. Either way, this will cast considerable doubt on Kim's claim that same-level causation *entails* downward causation and hence, on his broader conclusion about level-bound causal autonomy as well. In part 4 I consider, and ultimately reject, two ways in which Kim might respond to the above argument.

1 Edwards' dictum

The first stage of the supervenience argument involves the claim that there is a tension between same-level or "horizontal" causation and supervenience or "vertical determination." Kim introduces the tension by drawing on an example described by Jonathan Edwards, an eighteenth century philosopher and theologian. According to Edwards, God recreates the world at each instant ex nihilo. This means, contrary to appearances, that there are no temporally persisting objects or causal relations between such objects.

> It will follow from what has been observed, that God's upholding created substance, or causing its existence in each successive moment, is altogether equivalent to an *immediate production out of nothing*, at each moment. Because its existence at this moment is not merely in part from God, but wholly from him, and not in any part or degree, from its *antecedent existence*. For the supposing that his antecedent existence *concurs* with God in *efficiency*, to produce some part of the effect, is attended with all the very same absurdities, which have been shewn to attend the supposition of its producing it *wholly*. Therefore the antecedent existence is nothing, as to any proper influence or assistance in the affair; and consequently *God* produces the effect as much from *nothing*, as if there had been nothing *before*. So that this effect differs not at all from the first creation, but only *circumstantially*... (Edwards, 1808: 331).

According to Edwards then, God's creative act, which can be thought of as a synchronic determinative relation, precludes the possibility of diachronic causal relations between objects. Kim adopts the general form of the tension that Edwards identifies, dubs it "Edwards' dictum," and defines it as follows:

> *Edwards' dictum*: There is a tension between "vertical" determination and "horizontal" causation. In fact, vertical determination excludes horizontal causation (Kim, 2005: 36).

[2]Here my argument bears some similarity to one offered recently by Jens Harbecke (2013). However, Harbecke's discussion focuses on an anti-physicalist argument by Sturgeon (1998; 1999), involves a highly technical framework, and does not appear to work within Kim's own metaphysical assumptions. My hope is that my approach will be clearer and will be more compelling by virtue of working with assumptions and principles Kim himself endorses.

To illustrate the above tension and to clarify what he means by "vertical" and "horizontal," Kim introduces a less esoteric example: the colour of a lump of bronze. If we want to understand why the bronze is yellow at a certain time, *t*, it seems we can appeal to its microstructure at that very time. That is, we can think of the colour as synchronically dependent on some subset of the microphysical properties of the bronze. Since the colour is a macroproperty in comparison to the bronze's microstructure, we are to think of the micro-macro relation as one of "upward" determination, with the micro- and macroproperties arranged in a vertical array. The upward determination of the colour of the bronze is intended by Kim to be analogous to Edwards' claim about God's creative act, which is also a synchronic relation.

There appears to be another way one might account for the colour of the bronze, however: by appealing to its causal history. According to Kim, one might think of the bronze in terms of a series of successive "time-slices" and understand its colour at *t* as the causal product of the colour of the bronze at t-Δt, i. e., its colour at the previous time-slice. Since this involves a causal relation between objects and properties at the same mereological level, Kim calls this "horizontal causation," with the causal arrow moving through time, from left to right. The diachronic causal relation between successive time-slices of the bronze should be thought of as analogous to the putative (and ultimately illusory) causal relation between temporally successive objects in Edwards' example.

We appear to have a tension on our hands when we try to combine the two metaphysical accounts of the colour of the bronze at *t*. Kim sums the difficulty up as follows:

> As long as the lump has microproperty *M* at *t*, it's going to be yellow at *t*, no matter what happened before *t*. Moreover, unless the lump has *M*, or another appropriate microproperty (with the right reflectance characteristic), at *t*, it cannot be yellow at *t*. Anything that happened before *t* seems irrelevant to the lump's being yellow at *t*; its having *M* at *t* is fully sufficient in itself to make it yellow at *t* (Ibid.: 37).

The tension appears quite nicely to illustrate Edwards' dictum. The upward determination of the bronze's colour by its microstructure appears, as Kim says, to be in tension with the diachronic causation of its colour by an earlier time-slice of the bronze. In fact, so long as the bronze has the appropriate reflectance property at *t*, this being sufficient for its colour, the diachronic causal relation seems to be completely unnecessary, and so is pre-empted by the synchronic relation. Hence, with the aid of Kim's own example, it would appear that Edwards' dictum is a plausible one and identifies a genuine tension. Now that Edwards' dictum is in reasonably clear focus, I will sketch out the first stage of Kim's supervenience argument.

2 The first stage of the supervenience argument

Although our concern here is much broader, the intended target of the supervenience argument is nonreductive or "minimal" physicalism, including those forms of emergentism that are committed to the supervenience of emergent properties on their basal conditions.[3] The argument is a reductio that forces the nonreductive physicalist either to abandon her position (and embrace the reduction of mental properties to physical properties) or else to acquiesce in type epiphenomenalism. In Kim's own words, the goal of the argument is to foist upon the nonreductive physicalist the following conditional thesis: "If mentality is to have a causal

[3] In fact, an earlier incarnation of the supervenience argument appears at the end of Kim's paper, "Making Sense of Emergence" (1999). Since, in my view, the clearest and most thorough formulation of the supervenience argument appears in (Kim, 2005), my discussion will follow this version of the argument. This improves on the version in (Kim, 1998).

influence in the physical domain—in fact, if it is to have any causal efficacy at all—it must be physically reducible" (Ibid.: 161). Since I am only concerned with the first stage of the argument I will forego a lengthy treatment of Kim's background assumptions and his alternative formulations of the second stage of the argument.

Stage one of the supervenience argument begins as follows: A putative causal relation between two mental events seems to require a causal relation between two mental property instances, such that

1. *M* causes *M** (Ibid.: 39).

   Since the nonreductive physicalist is committed to supervenience we have:

2. For some physical property instance *P**, *M** has *P** as its supervenience base (Ibid.).[4]

   This is the point at which Kim appeals to Edwards' dictum. He claims that (1) and (2) together give rise to a tension because we have two competing accounts of why *M** is instantiated: *M** is caused to instantiate by *M*, and *M** is instantiated because its physical base *P** is instantiated. Kim claims that the vertical determination of *M** by *P** appears to exclude the diachronic causation of *M** by *M*. For what we see here is a tension analogous to the one between the synchronic determination of the colour of the bronze by its microstructure and the diachronic causation of its colour, or to the tension between God's being the sustaining cause of the world and a diachronic causal relation between temporally successive objects. As with Kim's other examples, if *P** is instantiated, it doesn't seem to matter what happened *before* the occurrence of *P**. So long as the supervenience base *P** is instantiated, *M** *must* instantiate since *P** is sufficient for *M**.

   According to Kim, in the light of Edwards' dictum the only way to secure a causal role for *M* is to suppose that *M* caused *M** to instantiate by causing its physical base *P** to instantiate. The solution is elegant; since *M** will instantiate as long as *P** instantiates, the way to make *M* relevant to the instantiation of *M** is to make it relevant to the instantiation of *P**. This seems plausible since, to use Kim's analogy (Ibid.: 20), in order to change the aesthetic qualities of a work of art one would need to alter its subvenient non-aesthetic properties; one cannot directly modify a painting's beauty. Hence, we now have an appeal to what Kim calls "downward causation":

3. *M* caused *M* by* causing its supervenience base *P** (Ibid.: 40).

   At this point in the argument Kim draws the following conclusions:

   > What the argument has shown at this point is that if *Supervenience* is assumed, mental-to-mental causation entails mental-to-physical causation—or, more generally, that "same-level" causation entails "downward" causation. Given *Supervenience*, it is not possible to have causation in the mental realm without causation that crosses into the physical realm. This result is of some significance; if we accept, as most do, some doctrine of macro-micro supervenience, we can no longer isolate causal relations within levels; any causal relation at level L (higher than the bottom level) entails a cross-level, L to L - 1, causal relation. In short, *level-bound causal autonomy is inconsistent with supervenience or dependence between the levels* (Ibid. 40).

[4]Since we are concerned here with the question of whether the supervenience argument generalizes we need only to think of *M* as a special science property and of *P* as its supervenience base.

I think the supervenience argument does not establish this conclusion about level-bound causal autonomy because I do not think Kim successfully motivates the move from (1) and (2) to (3) in the first stage of the argument. While it is true that most nonreductive physicalists are already committed to downward causation I don't think this commitment is or should be driven by the above argument. This is significant for the question of whether or not the supervenience argument generalizes—our main focus here. If I am correct and there are reasons to resist the above inference, then advocates of the layered model can preserve level-bound causation without the threat of those causal powers draining into their subvenient bases.

3 Revisiting the initial tension

To help bring my concerns into focus, I would first like to make some observations about Kim's treatment of Edward's dictum. Let's begin with Kim's use of Edwards' example involving God as the sustaining cause of the world. It is important to think about the broader metaphysical picture and ask about the source of the tension between diachronic causation and God's synchronic creative acts. If we agree with Edwards that God recreates the entire world ex nihilo at every instant, this does appear to preclude the possibility of diachronic causal relations. But does the origin of the tension lie with the idea that it is problematic to have two sufficient metaphysical sources for how the world (or some part of it) came to be as it is at any particular moment? I think not. The real reason we are driven to reject diachronic causal relations in this case, I suspect, is that the metaphysical picture within which God's creative acts are embedded is one in which we lack a *necessary precondition* for diachronic causal relations. If we follow the tradition of thinking of *events* as the causal relata and embrace something like Lombard's (1986) view that events are changes in temporally persisting objects or substances, it is clear that the kind of world Edwards describes has no events since there are no temporally persisting objects. Hence, the real tension at work in Edwards' example might not originate from the idea that there are two sufficient metaphysical sources for the way the world is at any particular time-slice, but from the fact that there can be no events in the world Edwards describes, which means Edwards' dictum, as Kim articulates it, is perhaps somewhat misleading.[5]

In Kim's defense, Edwards' is a peculiar case, and it is tempting to overlook its rarefied metaphysical features and concede the point that something is amiss in the claim that it is possible to reconcile the idea that earlier events cause later ones if the latter are instantaneously brought into existence by God. Indeed, I am sympathetic with this suggestion. However, there is an important disanalogy between it and the tension Kim appeals to in the supervenience argument. In Edwards' example the competing metaphysical relations are both *causal relations*, so the source of the tension in this case (overlooking the above concerns about the overall metaphysical picture) is that we have two competing causes of the state of the world at any given time. If we take causes to be sufficient for their effects, as Kim does, this is certainly undesirable unless we are willing to endorse widespread causal overdetermination, which I assume we are not. It would be odd indeed to claim that God's omnipotent creative power is a redundant, overdetermining cause.

My reason for highlighting the fact that Edwards' example involves a tension between two competing causes is to point out that this is quite different from the alleged tension to which Kim appeals in the supervenience argument after premise (2). The upward determination of

[5]Such a world might, however, be consistent with the existence of Kimian events construed as property exemplifications, since temporal duration is arguably not an essential part of Kim's model (Kim, 1976). Of course, since Kim thinks (with Lombard) that events are the relata of causation, in a world without causation it might be difficult to motivate the existence of even Kimian events.

*M** by *P** (unlike the relationship Edwards describes between God and the world) is *not* a causal relation, it is supervenience. So why should we agree with Kim that there is a tension between the claim that *M** is superveniently determined by *P** and the claim that *M* causes *M**? If these were both causal relations there would indeed be a formal tension in need of reconciliation, and so we would have a reason to move to premise (3) of the supervenience argument. Since the situation is *not* one where we have two sufficient causes, why should we see *any* tension here?[6] It seems that the tension exists only if one conflates causal sufficiency with the sui generis sufficiency of supervenient determination. Since these are *different* metaphysical relations, why can't one claim they are compatible and simply deny the alleged tension?[7]

Kim claims in a footnote that he used to support (3) by appealing to an exclusion principle but that he now prefers "to rely on the reader's seeing the tension I spoke of in connection with the two answers to the question 'Why is *M** instantiated on this occasion?' … I don't believe invoking any 'principle' will help persuade anyone who is not with me here" (Kim, 2005: 41). In the absence of any such principle, however, the question remains why one should see the tension, especially when it involves different kinds of metaphysical relations. Since the alleged tension is what drives the idea that same-level causation entails downward causation, surely it is not enough to justify it with an unsupported intuition. In the absence of any additional premises or background principles, the first stage of the supervenience argument threatens, at worst, to collapse at premise (2) or, at best, to lead to a stalemate between Kim and those who have a differing intuition about the compatibility of diachronic causation and supervenient determination. Thus, it is important to explore what additional metaphysical principles might break the stalemate and bring opposing intuitions in line with Kim's.

In what follows, I show that the additional principles at Kim's disposal fail to support the inference from (1) and (2) to (3) because they offer ways to alleviate the alleged tension that do not require an appeal to downward causation. If I am correct, Kim's argument does not justify the claim that "level-bound causal autonomy is inconsistent with supervenience or dependence between … levels." In fact, Kim's other principles and general metaphysical framework suggest just the opposite.

In describing the alleged tension, Kim writes:

> (1) and (2) together give rise to a tension when we consider the question "Why is *M** instantiated on this occasion? What is responsible for, and explains, the fact that *M** occurs on this occasion?" For there are two seemingly exclusionary answers: (a) "Because *M* caused *M** to instantiate on this occasion," and (b) "Because *P**, a supervenience base of *M**, is instantiated on this occasion" (Ibid.: 39).

In treating the two answers as "exclusionary" it looks like Kim is appealing to one of his infamous exclusion principles. Indeed, in stage two of the supervenience argument Kim appeals explicitly to *Exclusion*, which he formulates as follows:

> *Exclusion*: No single event can have more than one sufficient cause occurring at any given time—unless it is a genuine case of causal overdetermination (Ibid.: 42).

Since *Exclusion* only generates a tension where there are multiple *causes*, yet we just observed that only one of the metaphysical sources of *M** involves a causal relation, the supervenience

[6] It is also interesting to note that Kim's use of Edwards' dictum is designed to *legitimize* diachronic causation (at the physical level)—something Edwards would clearly reject and regard as anathema to his own principle.

[7] Ausonio Marras (2007) also makes this observation but does not explore its implications since he is more concerned with the second stage of Kim's argument.

argument requires a broader principle than *Exclusion* in order to motivate the tension. Kim has two such principles at his disposal: the first is his well-known principle of *explanatory exclusion* (1988; 1989; 1995); the second is a still broader principle of *determinative/generative exclusion*, which Kim briefly mentions in his preamble to the supervenience argument (Kim, 2005). In the rest of this section I consider how these principles might operate in the first stage of the supervenience argument.

In a nearly identical version of the supervenience argument (1997), Kim appeals explicitly to the principle of explanatory exclusion in order to motivate the tension identified above. The principle is defined elsewhere[8] as follows:

> *Explanatory exclusion*: there can be no more than a single *complete* and *independent* explanation of any one event, and we may not accept two (or more) explanations of a single event unless we know, or have reason to believe, that they are appropriately related—that is, related in such a way that one of the explanations is either not complete in itself or dependent on the other (Kim, 1988: 233).

If one were to accept such a principle, then there would indeed appear to be a tension involved in having two explanations for the occurrence of M^*. Furthermore, by having M cause P^*, which then determines M^*, as Kim proposes in premise (3) of his argument, one could alleviate the tension by showing that the two explanations of M^* are not independent of one another.[9,10] The causal dependence of P^* on M would *guarantee* that the explanations are not independent, so we can (at this stage of the supervenience argument) preserve them both by appealing to downward causation in the way Kim suggests. This is too quick, however. I think it important to consider Kim's treatment of the principle of explanatory exclusion and to look more carefully at how it might figure in the tension identified in the first stage of the supervenience argument.

I won't provide much in the way of detail about how Kim develops and defends the principle of explanatory exclusion, but a few points are central and deserve some elaboration. Kim motivates the principle by means of commitments to explanatory realism, causal realism, and some general considerations about the metaphysics of explanation. According to explanatory realism, a proposition C is an explanans for proposition E in virtue of there being a determinate, objective relation R between events c and e, that is, between the events the respective propositions are about.[11] Kim takes this relation to determine the correctness of an explanation and to serve as its objective content.

According to Kim, it is most often a causal relation that plays the role of the explanatory relation, hence the assumption of causal realism: that events stand in mind-independent causal relations with one another. Kim's support for the principle of explanatory exclusion and his intended meanings of the terms "complete" and "independent" that figure within it emerge from his survey of the various ways in which two causes (c_1 and c_2) of an event e might be

[8]Kim here refers to his (1989) in which he formulates the principle as follows: "two or more complete and independent explanations of the same event or phenomenon cannot coexist" but the principle is more perspicuous in his (1988).

[9]Graham MacDonald and Cynthia MacDonald (2006) exploit this fact and a particular version of the property exemplification view of events in order to undermine the supervenience argument.

[10]Note that we have here shifted from talking about the metaphysical *relations* involving M^* to talking about the *explanations* of M^*. In the present context this is harmless since, as we shall see, the independence of explanations is a matter of the independence of the explanatory relations they track.

[11]I will here follow Kim's convention of referring to events by lower case variables and propositions to the effect that those events occurred by the corresponding upper case variables.

related to one another. Suppose, in apparent contravention to the principle of explanatory exclusion, that we have two explanations for *E:* one in terms of *C1* and another in terms of *C2.* According to Kim, there are six ways in which the underlying causes might be related to one another: (1) $c_1 = c_2$, (2) c_1 is distinct from c_2 but is reducible to or supervenient on it, (3) c_1 and c_2 are partial causes of e, (4) c_1 is a part of c_2, (5) c_1 and c_2 are different links in the same causal chain leading to e, and finally, (6) e is causally overdetermined by c_1 and c_2. In case (1) Kim treats the explanations as equivalent. In cases (2) - (5), Kim claims that an explanation of E in terms of either C_1 or C_2 alone will be incomplete or fail to be independent of an explanation appealing to the alternative event. The reason an explanation of E in terms of C_1 is *incomplete* under scenario (3), for example, is that c_1 is on its own causally insufficient for e. Since c_1 and c_2 are each partial causes, any explanation that leaves out one of the causes will be incomplete. For Kim, given his commitment to explanatory realism, the metaphysical incompleteness of c_1 in the production of e entails the incompleteness of the *explanation* of E in terms of C_1. Similarly, under scenario (4), an explanation of E that appeals to C_1 fails to be *independent* of an explanation in terms of C_2 because of the metaphysical dependence of c_2 on c_1. Only in case (6), where e is causally overdetermined, are the alternative explanations complete and independent. Since causal overdetermination is sufficiently rare, Kim is happy to treat this as an exception to the exclusion principle.

Kim's catalogue of the various ways in which two explanations might be related provides plausible support for the principle of explanatory exclusion. Indeed, the principle of explanatory exclusion appears to be a straightforward corollary to the principle of causal exclusion appealed to in the supervenience argument. Given this, it would seem Kim is correct to maintain that we should treat the two explanations of M^* as not just *in tension*, but also as *exclusionary*.

This, however, is the point at which my earlier observations about Edwards' dictum become relevant once again. In the case of Edwards' example of God being the sustaining cause of the world, we can appreciate how a causal explanation for the way the world is at t that appeals to the previous time-slice is excluded by an explanation that appeals to God's creative act. Unless we are willing to endorse rampant causal overdetermination, only one cause can be sufficient, and hence, only one explanation can be maintained. I pointed out, however, that there is an important difference between Edwards' example and what is described in the supervenience argument. In the latter, the tension that Kim claims requires reconciliation via an appeal to downward causation involves two distinct kinds of relation: a diachronic causal relation between M and M^*, and a synchronic determinative relation involving the supervenience of M^* on P^*. Given this, it is not so clear that the principle of explanatory exclusion entails that these two explanations are exclusionary.

To sort this out we need first to recognize that Kim individuates explanations by the explanatory relation and the events it relates:

> Explanatory realism yields a natural way of individuating explanations: explanations are individuated in terms of the events related by the explanatory relation (the causal relation, for explanations of events). For on realism it is the objective relationship between events that ultimately grounds explanations and constitutes their objective content (Ibid.: 233).

According to this criterion of individuation, we should see the M-M^* relation and the P^*-M^* relation as different explanatory relations that ground different explanations. Elsewhere, Kim (1994) argues that explanations track dependency relations, and he claims that there might

be many different kinds of dependency relations in the world. In fact, he treats mereological supervenience as an entirely distinct kind of explanatory relation from causation:

> Another dependence relation, orthogonal to causal dependence and equally central to our scheme of things, is *mereological dependence* (or "mereological supervenience", as it has been called): the properties of a whole, or the fact that a whole instantiates a certain property, may depend on the properties and relations had by its parts (Ibid.: 67).

Kim (1994) also claims that *mind-body* supervenience can be thought of as a different kind of dependency relation than causal dependence. Given Kim's criterion of individuation for explanations, it is clear he should think the explanation of *M** that appeals to *M* is grounded in the diachronic causal relation between *M* and *M**, and tracks one kind of dependence, whereas the explanation of *M** that appeals to *P** is grounded in the supervenience relation, and hence, tracks a different kind of dependence. Assuming we adopt Kim's principle of explanatory exclusion, how does it apply in this case? Does one explanation actually exclude the other? Since the latter is not a causal relation, as I suggested earlier, perhaps one can resist Kim's argument by pointing out that there is no *formal* tension in need of resolution. In Kim's own words, perhaps the explanation of *M** in terms of supervenience is simply "orthogonal" to the causal explanation of *M**. To the extent that the alleged tension is necessary to Kim's conclusion that same-level causation entails downward causation, perhaps it is possible to block Kim's move from (1) and (2) to (3) in the first stage of the supervenience argument. To settle this matter we need to delve a little deeper into the principle of explanatory exclusion and the potential justification for treating these two explanations as exclusionary.

The situation we face in stage one of the supervenience argument involves two different explanatory relations that converge on the same explanandum event (*M**). In order to evaluate whether or not one excludes the other, we need to determine whether or not the explanations are complete or independent of one another. In Kim's view, a causal explanation is complete when the cause it mentions is sufficient for the effect. In that case, the explanation that appeals to *M* in order to explain *M** would seem to constitute a complete explanation, given that *M* is assumed to be causally sufficient for *M**, but what of the explanation that appeals to the supervenient determination of *M** by *P**? Is that also a complete explanation? In the light of Kim's treatment of the example involving the colour of the bronze, it would seem so. Kim claimed that given the presence of the relevant base, the supervenient property *must* occur, no matter what happened at the previous time-slice. Hence, although a different kind of sufficiency is clearly involved in the case of supervenient determination, it seems the presence of *P** is (by strong supervenience) sufficient for *M**, and so the explanation of *M** in terms of *P** is also complete.

There appear, then, to be reasonable grounds to treat the two explanations as complete, but are they independent? It won't do to claim the explanations fail to be independent simply because they converge on the same explanandum event. By that reasoning it would follow that explanations appealing to overdetermining causes aren't independent—something Kim clearly denies. At first glance, it might seem unclear how we should understand the concept of independence that is relevant to the present example. After all, Kim's discussion of the ways in which two explanations might fail to be independent appears to be cashed out primarily in terms of *causal* independence. However, two of his examples suggest a broader interpretation. Kim claims that causes fail to be independent if one supervenes on the other, or if one is a proper part of the other. It would seem, then, that causes fail to be independent if there is any

kind of metaphysical dependency relation between them—one that is more general than the relation of *causal* dependence that is active in cases of joint cause or of causal chains. This is hardly contentious as it amounts to a near tautology: given explanatory realism, explanations fail to be independent if the events they appeal to are not metaphysically independent of one another.

Are *M* and *P** independent in this sense? If we limit ourselves to the first stage of the supervenience argument there is simply not enough information to answer this question. To do so we need to consider the broader metaphysical picture that emerges in the second stage of the argument. If we grant that *M* has a supervenience base *P*, and that *P* causes *P**, then there are grounds for saying that the explanation of *M** that appeals to *M* is *not* independent of the explanation that appeals to *P** since *M* supervenes on *P* and *P* causes *P**. These dependency relations would appear to suggest that we could retain both explanations since the principle of explanatory exclusion allows for more than one explanation of an event, provided the explanations are appropriately related.[12] Since we are able to trace the explanatory relations that ground both explanations to a series of connected dependency relations, it appears that the two explanations *are* appropriately related, and so the principle of explanatory exclusion is not violated. Because the causal explanation of *M** can, in this way, be reconciled with the explanation of *M** that appeals to supervenience, it seems Kim's principle of explanatory exclusion does not support the claim that there is a tension in need of resolution in the first stage of the supervenience argument. We therefore lack any convincing reasons to suppose a tension exists, or if there is any prima facie tension, it is relieved simply by identifying the above dependency relations. Consequently, there is no need to appeal to downward causation in order to eliminate the alleged tension. We therefore have no reason to suppose that mental-to-mental causation entails mental-to-physical causation, or more generally, that same-level causation entails downward causation. We can, as Kim himself at one time suggested, regard the two explanations as "orthogonal" to one another rather than in competition.

If the principle of explanatory exclusion is unsuccessful at motivating the tension in the supervenience argument, perhaps Kim's other principle of determinative/generative exclusion will be more successful. In Chapter 1 of *Physicalism, or Something Near Enough*, Kim foreshadows the supervenience argument and proposes the following "generalized version of the exclusion principle":

> *Principle of determinative/generative exclusion*: If the occurrence of an event *e*, or an instantiation of a property *P*, *is determined/generated by an event c*—causally or otherwise—then *e*'s occurrence is not determined/generated by any event wholly distinct from or independent of *c*—unless this is a genuine case of overdetermination (Kim, 2005: 17).

Kim adds that the above principle expands on *Exclusion* because it "broadens causation, or causal determination, to generation/determination simpliciter, whether causal or of another kind" (Ibid.). By broadening exclusion to include *any* determinative relation, Kim would appear to succeed at reviving the tension in the supervenience argument between horizontal causation and vertical determination. With such a principle in place, perhaps one can no longer deny the tension in the first stage of the supervenience argument.

Kim's broader principle does make the alleged tension look more robust, but there are two ways to dissipate the tension that do not involve an appeal to downward causation: (1) reject

[12]Recall that the principle states: "we may not accept two (or more) explanations of a single event unless we know, or have reason to believe, that they are appropriately related—that is, related in such a way that one of the explanations is either not complete in itself or dependent on the other" (Kim, 1988: 233).

this broader version of exclusion; (2) argue as before that the two generative relations aren't independent of one another, and so we can preserve both.

The first strategy is obviously the strongest, but I will not belabour it here since what I have already said captures the essence of this reply. The main idea would be to show that, in the light of what we have already observed about Kim's views on the metaphysics of explanation, the broader version of the exclusion principle is unmotivated.[13] If one embraces explanatory realism and thinks of supervenience and causation as different *kinds* of explanatory relations, it is entirely unclear why we should embrace determinative/generative exclusion. The principle seems entirely ad hoc: Kim proposes it simply because it renders the tension in the supervenience argument more perspicuous. So the strategy here is to dig in one's heels and demand to know what is so objectionable about an event standing in two different kinds of explanatory relations. It would appear that, against the background of Kim's work on the metaphysics of explanation, the burden of proof lies squarely with him to show why we should accept this principle. It also bears pointing out that if, like Kim, one countenances causal overdetermination as a bona fide exception to determinative/generative exclusion, there is no *principled* obstacle to the convergence of distinct determinative or generative relations on a single event, and so it is fair to question the truth and value of such a principle, given that it has clear exceptions. The principle, in effect, seems to say "no overdetermination, unless this is a case of overdetermination" where "overdetermination" has been broadened to include any generative relation that is sufficient for the outcome. I fail to see the value of such a principle.[14]

Although I think one can deny the principle of determinative/generative exclusion, it is nevertheless possible to grant Kim this principle and still avoid the tension in a way that blocks the move from (1) and (2) to (3). We have already seen there are reasons for claiming the two metaphysical sources of M^* are not independent of one another once we consider the broader metaphysical picture within which M, M^*, and P^* are embedded. The determinative/generative exclusion principle only threatens M as a cause for M^* if M and P^* are *independent*. Since, as we saw earlier, there are grounds for saying they are not independent, there is no reason to suppose that the determinative production of M^* by P^* excludes the causal generation of M^* by M. Nothing in Kim's principle suggests otherwise, and so the alleged tension evaporates once again.

4 Potential responses

I have argued that two versions of exclusion (explanatory exclusion and determinative/generative exclusion) fail to motivate a tension between vertical determination and horizontal causation, and hence, that there is as yet no convincing reason to think that same-level causation entails downward causation. Since we are working within Kim's own metaphysical framework it is worth asking if he has other resources that might provide a response to my argument. The seeds of two possible replies lie in Kim's earlier work on the metaphysics of explanation and on the concept of supervenience.

[13] It is interesting to note that Kim has his own doubts about this principle when he suggests that it might be too broad (Kim, 2005: 20).

[14] Of course, it might be implicit here that Kim thinks "genuine" cases of overdetermination are rare or that there are other reasons for thinking that we should not treat M^*'s being caused by M and being superveniently determined by P^* as a case of genuine overdetermination. While he has some plausible things to say about why we should not think that P^* is causally overdetermined by M and P (Kim, 2005: 46-52) it is not clear that those lessons will carry over to the set of relations now under consideration. I also think that Kim's claim that genuine overdetermination requires the overdetermining factors to be independent of one another is unmotivated.

The first involves an assumption about the individuation of explanations. Much of my defence of same-level causation requires the idea that it is possible for different explanatory relations to converge on a single event, and some philosophers might balk at such a suggestion. Indeed, Kim offers a brief remark on the subject that might support such a response. In footnote 23 of "Explanatory Realism, Causal Realism, and Explanatory Exclusion," where Kim discusses how to individuate explanations, he writes:

> If relations other than the causal relation can serve as [the] explanatory relation, they can also be considered as a basis for individuation; however, that probably would be redundant. It is unlikely that when the explanatory relation is different, exactly the same events would be involved (Kim, 1988: 239).

So perhaps one could argue that it would be highly unlikely that a single event can stand in more than one explanatory relation to other events. Since the denial of the tension in the supervenience argument requires something like the possibility Kim seems here to dismiss, it looks as though there might be a problem for the argumentative strategy I have proposed.

The trouble with this response is that the above remark seems to be entirely without support. Why must events stand in solitary explanatory relations with other events given that a variety of such relations are possible? Even Kim acknowledges that there are different kinds of explanatory relations in the world. More significantly, were there a principled reason to deny the possibility that one event can stand in two explanatory relations, that would rule out the possibility of causal overdetermination all by itself. Surely we are willing to allow the possibility of causal overdetermination, so why not also allow for the kind of case described above? To fail to do so seems arbitrary. Granted, in the case we are concerned with the explanatory relations are relations of different kinds, but that does not appear to be any barrier to their converging on the same event. In fact, one might claim that such a situation is *less* problematic than causal overdetermination precisely *because* the explanatory relations are relations of different kinds. Finally, it is worth noting that all Kim's remark about individuation *technically* precludes is the *very same* pair of events standing in multiple explanatory relations. I agree with Kim that it would be extremely odd if, for example, *c* were not only a sufficient cause of *e*, but if *e also* strongly supervened on *c*. This, however, is not the case in the supervenience argument since the explanans for *M** refer to two *different* events (*M* and *P**). While *M** stands in two different explanatory relations, each relation involves a distinct explanans-event.

A second possible response targets the claim, exploited above, that the two explanations of *M** in terms of *M* and *P** are not in tension because they are not independent of one another. Since *M* supervenes on *P* and *P* causes *P**, *M* and *P** are not independent of one another and so both explanations of *M** can be preserved under explanatory and determinative/generative exclusion. Since we have stipulated that *M** is determined by *P** and that *P** is caused by *P*, the only point where there might fail to be a dependency relation is between *M* and *P*. Why think there might not be a dependency relation here? The only justification I can think of involves the idea that supervenience itself fails to be a dependency relation. Though he was by no means the only one to raise this concern, Kim at one time (1995) argued that we lack an account of supervenience that is strong enough to secure dependence, yet weak enough to avoid reduction. If one shares Kim's earlier concerns about the concept of supervenience, it might be possible to resist my argument by undermining the idea that the *M-P* supervenience relation is a dependence relation, thereby breaking the chain of metaphysical dependency between *M* and *P**.

There is not much to recommend this strategy. This is so for two main reasons. First, such a claim would stop the first stage of the supervenience argument in its tracks anyway. This stage of the argument requires us to see a tension between the vertical determination of M^* by P^* and the causation of M^* by M. If we deny that supervenience is a dependency relation, we can no longer say that M^* is determined by P^*, and the tension evaporates. Hence, the argument would collapse at precisely the same point I am proposing, though for an entirely different reason. Furthermore, this would vindicate the same-level causal relation between M and M^* as the only genuine dependency relation in stage one of the argument, which is precisely the relation Kim wants to reveal as illusory.

I think this is a convincing reply, but another point against it deserves mention. That supervenience is a dependency relation is a key assumption of nonreductive or minimal physicalism—the main target of Kim's argument. To see the relation as either reductive or as one of mere property covariation is, as Ruben (1990) points out, to fail to see it as an explanatory relation at all. This, of course, is simply to deny a key claim that defines the nonreductive physicalist's position. If stage one of the supervenience argument requires such a denial as a *background assumption* in order to block my response, the argument is certainly question begging. For to assume a deflationary view of supervenience is already to deny the nonreductive physicalist's position, since she takes this dependence to define her position and to legitimize her claim to be a physicalist. Hence, this does not appear to be a plausible way to respond to the argument.

In conclusion, if one does not see a tension between the diachronic causation of M^* by M and the supervenient determination of M^* by P^*, there is no need to appeal to downward causation in order to alleviate the tension. Only if the tension is genuine should we be compelled to think that mental-to-mental causation requires mental-to-physical causation, or that same-level causation entails downward causation more generally. I hope to have shown that the tension between same-level causation and supervenient determination is unproblematic, and have done so working from within Kim's own metaphysics. If I am correct, advocates of the layered model need not worry that the supervenience argument will generalize or that the efficacy of special science properties will drain into their physical bases.[15]

References

Block, N. (2003). "Do Causal Powers Drain Away?", *Philosophy and Phenomenological Research* 67, 133–50.

Bontly, T. (2002). "The Supervenience Argument Generalizes", *Philosophical Studies* 109, 75–96.

Edwards, J. (1808). *The Doctrine of Original Sin Defended: Evidences of Its Truth Produced, and Arguments to the Contrary Answered. Containing in Particular, a Reply to the Objections and Arguings of Dr. John Taylor, in His Book Intitled, "the Scripture Doctrine of Original Sin Proposed to Free and Candid Examination"*, Boston: S. Kneeland.

Gillett, C. and Rives, B. (2001). "Does the Argument from Realization Generalize? Responses to Kim", *Southern Journal of Philosophy* 39, 79–98.

Harbecke, J. (2013). "On the Distinction between Cause-Cause Exclusion and Cause-Supervenience Exclusion" *Philosophical Papers* 42, 209–38.

Kim, J. (2003). "Blocking Causal Drainage and Other Maintenance Chores with Mental Causation" *Philosophy and Phenomenological Research* 67, 151–76.

[15]I wish to express my gratitude to the anonymous referees for this journal for their helpful comments on an earlier draft of this paper.

Kim, J. (1997). "Does the Problem of Mental Causation Generalize?", *Proceedings of the Aristotelian Society* 97, 281–97.

Kim, J. (1976). Events as Property Exemplifications, *in* Brand, M. and Walton, N., eds, 'Action Theory', Dordrecht: Reidel. Reprinted in Kim, J. 'Supervenience and Mind', Cambridge: Cambridge University Press, 1993). References are to the 1993 reprint.

Kim, J. (1995). Explanatory Exclusion and the Problem of Mental Causation, *in* MacDonald, C. and MacDonald, G., eds, 'Philosophy of Psychology: Debates on Psychological Explanation', Oxford: Blackwell, 35–56.

Kim, J. (1994). "Explanatory Knowledge and Metaphysical Dependence", *Philosophical Issues* 5, 51–69.

Kim, J. (1988). "Explanatory Realism, Causal Realism, and Explanatory Exclusion" *Midwest Studies in Philosophy* 12, 225–40.

Kim, J. (1999). "Making Sense of Emergence", *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 95, 3–36.

Kim, J. (1989). "Mechanism, Purpose, and Explanatory Exclusion", *Philosophical Perspectives* 3, 77–108.

Kim, J. (1998). *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*, Cambridge, Mass.: MIT Press.

Kim, J. (2005). *Physicalism, or Something near Enough*, New Jersey: Princeton University Press.

Lombard, L. B. (1986). *Events: A Metaphysical Study*, London; Boston: Routledge & Kegan Paul.

Macdonald, C. and Macdonald, G. (2006). "The Metaphysics of Mental Causation", *Journal of Philosophy* 103, 539–76.

Marras, A. (2007). "Kim's Supervenience Argument and Nonreductive Physicalism", *Erkenntnis* 66, 305–27.

Noordhof, P. (1999). "Micro-Based Properties and the Supervenience Argument: A Response to Kim", *Proceedings of the Aristotelian Society* 99, 109–114.

Ruben, D. (1990). *Explaining Explanation*, London: Routledge.

Sturgeon, S. (1999). "Conceptual Gaps and Odd Possibilities", *Mind* 108, 377–80.

Sturgeon, S. (1998). "Physicalism and Overdetermination", *Mind* 107, 411–32.

abstracta Linguagem, Mente & Ação ISSN 1807-9792 Volume 8, Number 2, pp. 67–77, 2015

Adverbial Account of Intransitive Self-Consciousness

Roberto Horácio de Sá Pereira

University of Rio de Janeiro/UFR
Department of Philosophy
robertohsp@gmail.com.br

Abstract
This paper has two aims. First, it aims to provide an adverbial account of the idea of an intransitive self-consciousness and, second, it aims to argue in favor of this account. These aims both require a new framework that emerges from a critical review of Perry's famous notion of the "unarticulated constituents" of propositional content (1986). First, I aim to show that the idea of an intransitive self-consciousness can be phenomenologically described in an analogy with the adverbial theory of perception. In an adverbial theory of perception, we do not see a *blue* sense-data, but we see something *blue-ly*, whereas in intransitive self-consciousness we are not conscious *of* ourselves when we undergo a conscious experience—instead, we experience something *self-consciously*. But what does this mean precisely? First, I take intransitive self-consciousness to be the first-person operator that prefixes the content of any experience that the subject undergoes, regardless of whether or not the subject is self-referred. Further, I argue that this first-person adverbial way of entertaining a content of any experience in Perry's revised framework fixes the subject as part of the circumstance of the evaluation of the content of her own experience. We can only evaluate whether the content is veridical of falsidical relative to the subject undergoing the experience. This is referred to here as "self-concernment without self-reference." When I am absorbed reading a book, I do not self-represent my own experience of reading a book, let alone see myself as a constituent of the content of this experience. Even so, I experience that reading *self-consciously* in the precise sense that *I* do belong the circumstance of the evaluation of the selfless content of my experience of reading the book. The content of the experience of reading a book is simply a propositional function, true or false of myself.

1 Introduction

The phenomenological tradition from Husserl to Heidegger and Sartre postulates a form of pre-reflexive self-consciousness in opposition to the traditional reflexive self-consciousness in which the subject knowingly self-refers. However, this pre-reflexive self-consciousness is almost always described negatively: as non-reflexive, non-cognitive, non-transitive, not object-consciousness, and so on. As Schear (2009: 14) puts it, when it is time to offer a positive description of the phenomenon, we are left with incomprehensible metaphors such as "subtle background presence" (Zahavi, 2006b: 124). Moreover, it *seems* to be meaningless to talk about a primitive and omniscient form of self-consciousness with no knowing self-reference at all.

Things become still more confusing when we are told that *all consciousness involves this pre-reflexive self-consciousness*. One might honestly suspect that what they are calling pre-reflexive self-consciousness is nothing but simple consciousness. This is the way that the Heidelberg school has attempted to solve the old dilemma between a vicious circle and an infinite regress (Henrich, 1967, 1971; Pothast, 1971; Tugendhat, 1979). Whenever I am awake, I am always

conscious, even when I am not knowingly self-referring. Is that what phenomenologists meant by phrases such as "subtle background presence?" I do not think so.

Moreover, the very idea of a *pre-reflexive* self-consciousness has been understood quite differently in old and recent literature. Fichte was certainly the first to propose an account of self-consciousness in a pre-reflexive way. He understood this pre-reflexive self-consciousness in terms of what he calls a spontaneous act of *self-posit*: a sort of intellectual intuition of the self that dispenses Reflection. Recently, Horgen and Kriegel have proposed a neo-Brentanean reading of pre-reflexivity: in an experience, it is not only the scene that is represented in the content, but also the experience itself, and the subject of experience. According to Kriegel, "a mental state is (intransitively) self-conscious when it represents its own occurrence" (2003: 103).

Bermúdez (1998) has also proposed a non-conceptual primitive form of self-consciousness that dispenses with reflection altogether. Pre-reflexivity means a kind of knowing self-reference without mastering the *token-reflexive rule* of the employment of the first-person pronoun (1998: 15). Many cognitive psychologists describe *self-awareness* in a similar sense: a self-reference given in perception and proprioception that dispenses with reflection. However pre-reflexive they are, these putative forms of self-consciousness are not *intransitive* or *non-objectual* in the original sense required by the phenomenological tradition. To avoid any further confusion, from now on I will avoid the term "pre-reflexive." Like Kriegel (2003: 103), I prefer the technical term *intransitive* self-consciousness. Still, in my adverbial account there is no self-reference, as no mental state represents itself.

When we observe the reasons given in support of the idea of intransitive self-consciousness, things do not look any better. In the recent history of continental philosophy, this idea is motivated by three reasons. The first, and perhaps most quoted one, is the alleged necessity of avoiding the infinite regress that is generated by the traditional Theory of Reflection. The phenomenological solution to this old problem (Sartre) is to postulate an intransitive form of self-consciousness in which the subject does not take herself as the object either of some intellectual intuition (Fichte) or of some nonconceptual self-awareness (Bermúdez). However, as we will appreciate, this argument is far from convincing. For one thing, only Shoemaker, in his seminal paper (1968), provides an acceptable solution.

The second reason is what Schear (2009) has recently called the "interview argument" in Zahavi's work (2006b). The idea is intuitive, but his argument is not very convincing. The idea is that intransitive self-consciousness is required to explain not only reflexive self-consciousness (what is false), but also every form of reflexive consciousness. Suppose I am reading a book when someone interrupts me to ask me what I am doing. Now, since I am able to reply immediately, without inference or observation, I must be intransitively self-conscious of my experiences the whole time that I am reading (Zahavi's favorite example, 2006a: 21).

As it stands, Zahavi's argument faces the so-called "refrigerator light problem" (Schear, 2009). Every time I open the door of my refrigerator the light is on, but of course that does not mean that the light is on the whole time that the door is closed. Likewise, every time someone interrupts me by asking me what am I doing, I can immediately reply by knowingly referring to myself as the subject that was reading. However, that does not entitle me to infer that I was intransitively self-consciousness all the time that I was reading. As before, an easier explanation is available. When I was reading the book, I was the *subject* of my experience all along. But that does not mean that I was intransitively *self-conscious* the whole time that I was reading. I first become self-conscious when the focus of my attention shifts from the book to myself as the subject undergoing the experience of reading.

However, the most promising reason in support of the idea of an intransitive form of self-consciousness is what we may call the phenomenological argument. According to Zahavi, careful attention to our conscious experience reveals what he calls "mineness" or "for-me-ness." The conscious experience is distinctively *mine* in the sense that it is *me* rather than you who is having the experience. The intuitive suggestion here is that, if there is something that is like "to read a book," than that like is *for me*. Even though I do agree with this phenomenological description, I think that it requires further explanation. What does it mean to say that, in all my experiences, there is something that is "what it is like to be in those experiences" *for me*?

This paper has two aims. First, it aims to provide an adverbial account of the idea of an intransitive form of self-consciousness. Second, it aims to argue in favor of this account. For both of these things, a new framework is required, which emerges from a critical review of Perry's famous notion of the "unarticulated constituents" of propositional content (1986). First, I aim to show that the idea of an intransitive self-consciousness can be phenomenologically described in an analogy with the adverbial theory of perception. In the adverbial account, we do not see a *blue* sense-data, but we see something *blue-ly*. However, in intransitive self-consciousness, we do not necessarily experience ourselves whenever we undergo a conscious experience. Instead, we experience something *self-consciously*. But what does that mean exactly? First, I take intransitive self-consciousness as the first-person operator that prefixes the content of any experience that the subject undergoes, regardless of whether or not she is self-referred in that content. Further, in my revised version of Perry's framework, I argue that this first-person adverbial way of entertaining a content of any experience fixes the very subject as part of the circumstance of evaluation of the selfless content of her own experience. We can only evaluate whether content is veridical of falsidical according to the subject undergoing the experience. That is what is referred to here as "self-concernment without self-reference." When I am absorbed in reading a book, I do not self-represent my own experience of reading a book, let alone myself, as a constituent of the content of this experience. However, I do experience reading *self-consciously* in the precise sense that *I* do belong to the circumstance of the evaluation of the selfless content of my experience of reading the book. The content of the experience of reading a book is simply a propositional function true of false *of myself*.

The plan of this paper is as follows. The first section reminds the reader of some truisms about self-consciousness and undertakes an analysis of the traditional Theory of Reflection. In the second section, I present the phenomenological idea of an intransitive self-consciousness as the historical solution that Sartre proposed to this dilemma. However, I reject the idea that intransitive self-consciousness is the best solution to this classical dilemma. In this section, I also manifest my misgivings about Zahavi's interpretation of intransitive self-consciousness with relation to Shoemaker's notion of self-reference without identification. I will try to persuade the reader that the old phenomenological concept of an intransitive self-consciousness is best captured by what I am calling here "self-concernment without self-reference."

The third and final section of this paper is devoted to explaining my reading. Starting with a critical review of Perry's famous notion of "unarticulated constituent," I will try to show that, when there is something that is "what it is like to be in the state representing that content" *for me*, I am entertaining the selfless content of my own experience. Therefore, the content of my own experience is meant to be evaluated in a world that is centered on me and the context of the experience.

2 Self-Refeerence an Reflexivity

The connection between self-consciousness and reflection is commonsensical. Even so, in order to establish and define the main concepts that are involved, I shall remind the reader of some well-known things here. Thus, let us suppose that Oedipus, as king of Thebes, while examining the evidence of Laius's death, is led to think the following:

(1) That murder deserves punishment.

To the extent that Oedipus happens to be Laius's murderer, by thinking (1), he refers to himself, even though he is *unaware* of that fact. Things change as the tragedy comes to its end and Oedipus realizes that he is Laius's murderer and that he married his own mother. Terrified by his discovery, he might have thought:

(2) I feel remorse!

Thus, even though Oedipus self-refers in (1), he is only self-conscious when he knowingly or reflexively self-refers in (2). In contemporary philosophy, (2) (as a vehicle of thought) is usually called an "I-thought," and its propositional content is called *de se* content. In contrast, (1) is only a thought that "happens to be" about the subject herself: that is, a thought that contains an accidental or an unbeknown self-reference and whose propositional content is either *de dicto* (such as "the murderer of") or *de re* (such as "that guy"), but never genuinely *de se*.

This uncontroversial truism clearly connects conceptual self-consciousness to knowing or reflexive self-reference. However, according to a long and living tradition (the Theory of Reflection), which extends from Locke (1959) to Rosenthal (2004), such cognitive self-reference requires *another* higher-order thought beyond the original I-thought (2) in order to make it self-conscious. The idea is as simple as this: if I becomes conscious of something by representing it, then I could only become conscious of myself by a higher-order thought representing myself as the subject that some lower-order thought is about. Thus, even when he entertains (2), Oedipus is still not self-conscious. According to Rosenthal, for example:

> "Though the thought itself does not describe that individual as thinking that thought, the individual that thinks the thought is disposed to pick that individual out in that way, by being disposed to have another thought that so identify the first thought is about." (Rosenthal, 2004: 167)

Therefore, Oedipus only becomes self-conscious when his original I-thought (2) disposes him to entertain the higher-order thought (3), which identifies him as the subject of thought (2), as the following explains:

(3) I am the individual that (2) is about.

In this Theory of Reflection, one only manages to knowingly self-refer when one's first-order thought (2) disposes one to entertain higher-order thoughts (3) that refer back to (2) and identify the subject as the individual that (2) is about. Only this meta-representation can account for the key difference between the accidental self-reference in (1) and the cognitive self-reference in (2), disposing the subject to think (3).

The main difficulty that this Theory of Reflection faces is better formulated in terms of the traditional dilemma between vicious circularity and infinite regress. This dilemma can be described as follows. On the one side of the dilemma, in order to know whether (2) is about him, Oedipus has to know the truth of (3), but that requires him to entertain the following higher-order thought:

(4) I am the individual who is having thought (2).

But, in order to know (4), Oedipus would have to entertain a further, higher-order thought, to the effect that he is having thought (4):

(5) I am the individual who is having thought (4).

The same issue will arise over and over again *ad infinitum*.

Rosenthal does not recognize the possibility of a vicious regress, because he assumes from the beginning that *unaware* higher-order thoughts could enable lower-order thoughts to become conscious (2004: 164; 167). However, as Zahavi correctly claims (2006a), it is a complete mystery how a non-conscious higher-order thought could render an equally non-conscious lower-order thought conscious just by representing it.

On the other side of the dilemma, the knowledge expressed by (3) is already rooted in Oedipus's original lower-order thought (2). But insofar as self-consciousness emerges as the result of an act of Reflection on oneself, to assume that the knowledge expressed by (3) is already rooted in Oedipus's original lower-order thought (2) is to *presuppose* knowing self-reference rather than to *provide an account of* it. That is what Fichte calls a vicious circle (1937).

Therefore, Fichte was the first philosopher who clearly saw this as a threat to the Theory of Reflection. If self-consciousness was only able to emerge as the result of thought turning back onto itself <*Sichzurückzuwenden*> (higher-order thought), in a sentence such as (2), knowing self-reference is presupposed rather than explained. He describes this dilemma in the following words:

> "We become… conscious of the consciousness of our consciousness only by making the latter a second time into an object, thereby obtaining consciousness of our consciousness, and so *ad infinitum*. In this way, however, our consciousness is not explained, or there is consequently no conscious at all, if one assumes it to be a state of mind or an object and thus always presupposes a subject, but never finds it. This sophistry lies at the heart of all systems hitherto, including the Kantian." (*Nachlass*: 356)

Henrich reformulates Fichte's paradox in the following terms:

> "(a) It is not difficult to see that the reflection theory is circular: if we assume that reflection is an activity performed by a subject—and this assumption is hard to avoid—it is clear that reflections presuppose an "I" which is capable of initiating activity spontaneously, for the "I" as a kind of quasi-act cannot become aware of its reflection only *after the* fact. It must *perform* the reflection and be conscious of what it does as the same time as it does it." (1971: 11)

However, Cramer has formulated this problem with the most clarity:

> "But how can the subject know the she in the reflection has herself as her own object? Apparently, only through the fact that the ego knows that she is identical with herself as her own object. Now, it is impossible to attribute this knowledge to reflection and to justify knowledge from it. Because for every act of reflection is presupposed that the I am already acquainted with myself, to know that the one with whom she acquainted, when it takes herself as object, is identical to the one who is making the act of reflective turning back on itself. The theory, which wants to make the origin of self-consciousness understandable, therefore ends necessarily in the circle: that knowledge already must presuppose what it wants to explain in the first place." (Cramer, 1974: 563)

Fichte's own solution to this problem is unclear, but very well known:

> "The "I" posits itself absolutely, that is, without any mediation. It is at the same time subject and object. The I only comes into being through its self-positing—it is not an preexistence substance—rather, its essence in positing is to posit itself, it is one and the same thing; consequently, it is immediately conscious of itself." (*Nl* II: 357)

However, Fichte has never explained his discussion of self-posing <*Sichsetzen*> (Henrich, 1967: 18). The formula "the 'I' posits itself" can only negatively characterize his own rejection of the Theory of Reflection. However, according to the Heidelberg School, the idea of "self-positing" seems incomprehensible. Following this traditional reading, Fichte seems to mean that the "I" comes into existence through self-positing. Still, "how could someone perform that very act of positing if it does not yet exist in the first place?" (Pothast, 1971: 71).

But there is a more charitable reading of Fichte's insight. According to Frank:

> "But there *is* consciousness; so this model must be wrong. If wrong, then consciousness must have been immediately acquainted with itself, that is, prior to any objectification by means of a succeeding consciousness. Fichte accounts for *this immediate self-acquaintance as the complete indiscernibility of subject and object in self-consciousness*. Now in Kantian terminology an immediate consciousness is an intuition. But what is intuited here, is not a spatio-temporal entity, like in sensible intuition, but rather the being of the sheer spontaneity of apperception: hence the intuition is deemed to be intellectual." (Frank, 2004: 77. The emphases are mine)

Fichte's original insight is that self-consciousness is based on self-acquaintance or a level of self-intuition. Since the self-intuition is not spatio-temporal, its form must be the sheer spontaneity of apperception: *an intellectual intuition*. I become acquainted with myself any time that I think. That is what Fichte means by self-positing. To be sure, Fichte is right when he holds on the Cartesian insight that whenever I perform an act of thinking, I knowingly refer to myself. However, in those terms we are back to circularity: how can the subject be acquainted with the sheer spontaneity of her apperception if she did not previously know that she was the subject behind the act of her own intellectual intuition?

3 Intransitive Self-Consciousness

Phenomenologists postulate a pre-reflexive, non-intentional form of access to oneself as a natural way out of this dilemma. In this sort of primary self-disclosure, one does not take oneself *as an object*, either of one's own inner perceptions or of one's own thought. Rather, primary self-disclosure is a non-objectifying and non-perceptual form of self-consciousness. Husserl was the first to claim that experience is self-conscious in the sense of being lived through (*erlebet*), rather than in the sense of being the object of reflection (1984: 669). But Sartre (1976) was certainly the philosopher from the phenomenological tradition who most contributed to the elaboration of what is called pre-reflexive self-consciousness today. For Sartre, only the necessity of syntax compels us to say that we are pre-reflexively aware *of* our experiences or *of* ourselves. In his words:

> "Thus reflection has no kind of primacy over the consciousness reflected-on. It is not reflection which reveals the consciousness reflected-on to itself. Quite the contrary, it is the non-reflective consciousness which renders the reflection

> possible; there is a pre-reflective cogito which is the condition of the Cartesian cogito." (Sartre, 1976: 19–20)

Heidegger presents the same idea of an intransitive self-consciousness:

> "Dasein, as existing, is there for itself, even when the ego does not expressly direct itself to itself in the manner of its own peculiar turning around and turning back, which in phenomenology is called inner perception as contrasted with outer. The self is there for the Dasein itself without reflection and without inner perception, *before* all reflection. Reflection, in the sense of a turning back, is only a mode of self-*apprehension,* but not the mode of primary self-disclosure." (Heidegger, 1989: 226)

The basic claim is that intentional or transitive consciousness of the world entails a pre-reflexive, intransitive, non-intentional self-consciousness. According to Sartre, one becomes pre-reflexively conscious of oneself in one's confrontation with what one is not. Therefore, there is no threat of infinite regress: "because consciousness has no need at all of a reflecting consciousness in order to become conscious of oneself" (1957: 45). The intransitive self-conscious is an immediate non-cognitive relation of the self to itself.

There is no space here to undertake an (necessary) exegesis of the main contributions of the vast phenomenological tradition to this notion. Therefore, I limit myself to summarizing those features that I consider to lie at its core. First, we are told that intransitive self-consciousness is a non-objectifying form of self-consciousness. By living through (*erbeben*), we do not see ourselves as objects of any intentional consciousness. Second, we can see the importance of idea that intransitive self-consciousness is the most primitive form of consciousness. Third, we can reflect on the idea that intransitive self-consciousness is a necessary condition for any intentional consciousness of the world. This suggests that intransitive self-consciousness is not the result of any self-awareness or self-observation. The fourth and last feature is closely connected to the second one: intransitive self-consciousness is omnipresent. Since I am awake and not in a coma, I am always intransitively self-conscious.

Recently, Zahavi (2006a: 280) has interestingly suggested that the key idea of a pre-reflexive or intransitive access to oneself (which does not present oneself as an object) could be understood as a form of self-reference without identification (in the way that Shoemaker (1968) characterizes immunity to error through misidentification relative to the first-person pronoun). In his own words:

> "Rather, my pre-reflective access to myself in first-personal experience is immediate and non-observational and non-objectifying. It involves what has more recently been called… "self-reference without identification" (Shoemaker, 1968). (2006a: 280)

Zahavi's reading is heavily based on the analogy between Husserl's and Sartre's characterizations of intransitive self-consciousness as non-observational and non-objectifying, and Shoemaker's (1968) and Wittgenstein's (1958) characterizations of self-reference without identification as being subjective (the use of the "I" as subject in contrast to the use of the "I" as object).

However, this striking terminological similarity is deceptive for at least three reasons. First, as Zahavi himself recognizes, the phenomenological tradition conceives pre-reflexive access to oneself in a non-intentional or non-referential way, and where there is no self-reference, there can be no *self-reference* without identification in the first place. In other words, whereas in

Shoemaker's account, the self as subject is part of the first-person content (2) [I-thought], Husserl's and Sartre's self as subject is never part of any content at all.

Second, Shoemaker's self-reference without identification is the only correct solution to the traditional problem of the Theory of Reflection. The only way of detaining an infinite regress (without avoiding presupposing the knowing self-reference that it must explain in the first place) is to assume that in limiting cases, such as those of self-ascription of mental predicates, self-reference does not require self-identification or is identification-free. In contrast, however interesting the phenomenological idea of an intransitive self-consciousness may be, it is not a real solution to the traditional problem of the Theory of Reflection. For one thing, if there is no self-reference in intransitive self-consciousness, it cannot account for knowing self-reference in the first place.

Moreover, Shoemaker's self-reference without self-identification does not capture any of the other key features of the phenomenological view, as described above. First, Shoemaker's account does not capture the phenomenological idea that pre-reflexive self-consciousness is primitive—while intransitive self-consciousness does not need to involve any self-concept, we certainly need a self-concept for cognitive self-reference, even when this self-reference is identification-free. Second, self-reference without identification sufficiently meets the criterion that intransitive self-consciousness is a condition for transitive consciousness. Obviously, Oedipus thinking (2) is not a condition for his representing anything else. Third, self-reference without identification cannot capture the idea that the intransitive self is omnipresent in all of the subject's experiences. We do not spend our waking lives thinking I-thoughts such as (2).

Now my first suggestion is to provide a positive description of the idea of an intransitive self-consciousness in an adverbial way that is analogous to the adverbial theory of perception. According to the adverbial theory of perception, we should think of sensory qualities as adverbial modifications of the experience itself rather than as properties instantiated by sense-data. Hence when someone has an experience of something blue, something like blueness is instantiated, but in the experience itself, rather than in its object. Thus, instead of saying that we experience blueness, we should say that we experience *blue-ly*. In a nutshell, the fundamental idea is to interpret the properties of the object of experience through adverbial modifications of the perceptual verbs (see Crane, 2015).

The adverbial theory of perception has two fundamental motivations. First, it aims to do justice to the phenomenology of experience while avoiding the commitment to Moore's ontology of the sense-data theory (Ducasse, 1942; Chisholm, 1957). The second motivation, closely connected to the first, was to reject the subject-object model of perception. Now I want to suggest that both motivations also link to the phenomenological idea of intransitive self-consciousness. First, phenomenologists want to avoid the reification of the subject as a substance. Second, phenomenologists also want to reject the subject-object model of self-consciousness. In that sense, their criticism of the Theory of Reflection goes deeper than Fichte, the Heidelberg school and the recent neo-Brentanean accounts (Horgan and Kriegel, 2007). It is not enough to reject the idea that self-consciousness emerges as the intellectual act of reflection on oneself. We must reject the idea that self-consciousness is always the object of the subject's thoughts, perception, proprioception, and so on. Therefore, the natural suggestion is to take self-consciousness as an adverbial modification of the representational content of experience: the subject's first-person way of entertaining the content of experience. It works just like an epistemic operator that prefixes any representational content:

(6) I think that *p*.

Where *p* can be the content of any experience the subject undergoes regardless of whether or not she self-refers in that experience.

4 Intransitive Self-Concernment

In this final section, I want to finish my alternative account. My starting-point is Perry's famous thought-experiment (1986). Perry invites us to consider Z-landers, a group or a tribe who live in complete isolation and have never left Z-land, the place where they live. What matters to us is the following. When Z-landers file weather reports such as "it is raining," "Z-land" is an argument role of a certain relation that never changes <rains; Z-land>. The correct conditions of its content certainly *involve* Z-land: the place where the Z-landers' weather report is filed. This content is correct or accurate if it is raining in Z-land at the time that Z-landers report it. However, Z-land is an argument role that never changes. Therefore, Z-landers do not need to worry about Z-land. According to Perry, Z-land is a so-called "unarticulated constituent" of the representational content "it is raining"; that is, a constituent of the representational content of their report that is neither verbally articulated by any utterance, nor mentally represented.

Now Perry's claim that Z-land is an unarticulated constituent of the content of Z-landers' weather reports is disputable. He supports his claim that, in such cases, the "argument role" is an unarticulated constituent of the content by arguing that the content of a thought would otherwise be incomplete, in the sense of being a proposition without a determined truth-value.

However, this last argument is also questionable. Within the framework of Kaplanian semantics (1989), a sentence *S* is true in its context of use *c if* the proposition *p*, expressed by *S* at *c*, is true at the default circumstance of evaluation, as determined by *c*. The default circumstances of evaluation are pairs of a world and a time, so proposition *p* is true for a given circumstance *if* the proposition is true in the world and time of that circumstance. If we are not afraid of incomplete contents, nothing prevents us from thinking of Z-land as *a further aspect of the circumstance of the evaluation* of the placeless content of Z-landers' reports, rather than as an unarticulated constituent of the content.

In opposition to Perry's account, I see no compelling reason against considering "the argument role that never changes" as an aspect of the wider circumstance of evaluating an incomplete, relative, and placeless proposition. There is also a simple argument against Perry's idea that "the argument role that never changes" is an unarticulated constituent of the content. It seems unlikely that they (the residents of Z-land) would have an idea or concept of Z-land in the first place. So in accounting for their communicative exchanges about the weather in Z-land, it would be more logical to assume that they are not implicitly referring to Z-land at all.

Perry's own further examples substantiate the same point. Time zones certainly are argument roles in any time report. However, before the Europeans' great discoveries of new continents, the argument roles of time reports never changed. Therefore, people never implicitly referred to time zones as unarticulated constituents of their time reports, because they did not have a concept of time zones in the first place. The most parsimonious part of the account of the Z-landers' weather report is the assumption that the concept of Z-land is an aspect of a wider circumstance of evaluation, rather than an unarticulated constituent of the placeless content itself.

Let us now suppose that anthropologists find the Z-landers. As usual, there are exchanges of gifts, and Z-landers receive cell phones from the anthropologists to communicate with their new acquaintances outside Z-land. Now things change. When they communicate the weather conditions in Z-land to the anthropologist outside Z-land, they must refer to Z-Land. I see no

reason why the new reference to Z-land must be verbally or mentally articulated in such a way as: "It is raining *in Z-land*." The reference to Z-land in their new reports can be implicit: "It is raining."

However, there are two crucial points that Perry does not mention in his famous paper. First of all, in order to refer to Z-land in reports, Z-landers must acquire a concept of Z-land. Without a concept of Z-land, Z-landers cannot refer to their land. In this regard, the reference to Z-land as the reference to a time zone is quite different from the reference to objects and properties within the subject's perceptual field. For one thing, like entities that are postulated by science (quarks, atoms, energy, photons, etc.), Z-land is never given as an object of perception (without a concept of Z-land, Z-land is ubiquitous; it is both everywhere and in no specific place at the same time). In these cases, references rely on, and are created by, concepts. Second, we must assume that the Z-landers are now referring to Z-land because this is the best available explanation for their intentional behavior of communicating with their new friends outside Z-land, reflecting their way of grasping the world.

Now let us apply this reviewed framework to our case in point: the phenomenological idea of an intransitive self-consciousness. First, as I have already suggested, the intransitive self is best understood in the adverbial fashion: the first-personal way that the subject entertains the content of her experiences without necessarily self-referring in that content. In that sense, the intransitive self is also "an argument role that never changes," and hence the subject does not have to worry about herself when she experiences something with certain content.

Now my proposal is as follows. Even though the content of experience is selfless in the sense that the subject is not represented, the subject is always present as the one who is entertaining the content of her own experience. This fixes the subject, the time, and the location where the experience takes places as the context relative to which the selfless content of experiences is meant to be evaluated: a world centered on the subject and the time and space when and where her experience takes place. I describe this view as intransitive self-concernment without self-reference.

However, as in the case of Z-landers, things change when the subject starts to communicate her experience to someone else with a different viewpoint. Let us suppose that someone asks the subject, "My dear, I am without my glasses. Do you see a cat in the tree over there?" Now the subject cannot help but refer either implicitly or explicitly to her own viewpoint in her answer to the person's question: "Yes, *I* do see a cat over there." Now in the face of this communicative exchange, the self, which was merely involved as the subject for whom there is something that is like "to be in a certain state," also becomes an essential part of the *de se* content.

References

Bermúdez, J. (1998). *Das Paradox des Selbstbewußtseins*, Cambridge: MIT Press.

Crane, T. (2015). 'The Problem of Perception.' *The Stanford Encyclopedia of Philosophy* (Summer 2015 Edition), E. N. Zalta (ed.), http://plato.stanford.edu/archives/sum2015/entries/perception-problem/.

Cramer, K. (1974). 'Erlebnis'. Thesen zu Hegels Theorie des Selbstbewußtseins mit Rücksicht auf die Aporien eines Grundbegriffs nachhegelscher Philosophie, *Hegel-Studien*, Beiheft 11: 537–603.

Heidegger, M. (1989). *Die Grundprobleme der Phänomenologie,* Gesamtausgabe Band 24, Vittorio Klostermann Verlag, Frankfurt a. M.

Heidegger, M. (1993). *Grundprobleme der Phänomenologie* (1919–1920), Gesamtausgabe Band 58,Vittorio Klostermann Verlag, Frankfurt a. M.

Henrich, D (1967). Fichtes ursprüngliche Einsicht, *in:* D. Henrich & H. Wagner (eds.), 'Subjektivität und Metaphysik. Festschrift für Wolfgang Cramer', Vittorio Klostermann Verlag, Frankfurt a. M., 188–233.

Henrich, D (1971). 'Self-consciousness, a critical introduction to a theory', *Man and World* **4(1)**: 3–28.

Horgan, T. & Kriegel, U. (2007). 'Phenomenal epistemology: What is consciousness that we may know it so well?', *Philosophical Issues* **17 (1)**: 123–144.

Husserl, E. (1984). *Einleitung in die Logik und Erkenntnistheorie, Husserliana XXIV,* Martinus Nijhoff, Den Haag.

Fichte, G. (1937). *Nachlass: Gesammelte Schriften, herausgegeben von der Königlich Preussischen Akademie der Wissenschaften (German Academy edition),* De Gruyter, Berlin.

Frank, M. (2004). 'Fragments of a history of the theory of self-consciousness from Kant to Kierkegaard', *Critical Horizons* **5(1)**: 53–136.

Kriegel, U. (2003). 'Consciousness as intransitive self-consciousness: Two views and an argument', *Canadian Journal of Philosophy* **33(1)**: 103–132.

Kaplan, D. (1977/1989). Demonstratives, *in* J. Almog, J. Perry & H. Wettstein (eds.), 'Themes From Kaplan', Oxford University Press, Oxford, 481–563.

Locke, J. (1959). *An Essay Concerning Human Understanding,* annotated by A. C. Fraser, Dover, New York.

Perry, J. (1986). 'Thought Without Representation', *Aristotelian Society Supplementary Volume,* **60**: 137–152.

Pothast, U. (1971). *Über einige Fragen der Selbstbeziehung,* Vittorio Klostermann Verlag, Frankfurt a. M.

Sartre, J. P. (1976). *L'être et le néant*, Gallimard, Paris.

Sartre, J. P. (1957). *The Transcendence of the Ego* (F. Williams and R. Kirkpatrick, trans.), Noonday Press, New York.

Shoemaker, S. (1968). 'Self-Reference and Self-Awareness', *Journal of Philosophy* **65**: 555–567.

Schear, J. K. (2009). 'Experience and self-consciousness', *Philosophical Studies* **144(1)**: 95–105.

Tugendhat, E. (1979). *Self-Consciousness and Self-Determination* (P. Stern, trans.:*),* MIT Press, Cambridge MA.

Rosenthal, D. (2004). 'Being conscious of ourselves', *The Monist* **87(2)**: 161–184.

Wittgenstein, L. (1958). *The Blue and Brown Books,* Blackwell, Oxford.

Zahavi, D. (2006a). Thinking about (Self-) Consciousness: Phenomenological Perspectives, *in*: U. Kriegel & K. Williford (eds.), 'Self-Representational Approaches to Consciousness', MIT Press, Cambridge MA, 273–295.

Zahavi, D. (2006b). *Subjectivity and Selfhood: Investigating the First-Person,* MIT Press, Cambridge MA.