Spring 5-15-2021

The Feeling Mind

Maria Doulatova
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Program in Philosophy-Neuroscience-Psychology

Dissertation Examination Committee:
Casey O'Callaghan, Co-Chair

John Doris, Co-Chair

Allan Hazlett

Ron Mallon

Laurie Paul

Elizabeth Schechter

The Feeling Mind

by
Maria Renatovna Doulatova

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2021
St. Louis, Missouri

Table of Contents

Acknowledgments

I would like to thank the Philosophy-Neuroscience-Psychology department for continued social and financial support. My committee members have gone above and beyond in making this manuscript a reality. I am humbled by their heartfelt dedication to my cause.

Finally, I would like to thank my family and especially my husband for always being there for me.

Maria Doulatova

Washington University in St. Louis

May 2021

Dedicated to my father.

ABSTRACT OF THE DISSERTATION

The Feeling Mind

By

Maria Renatovna Doulatova

Doctor of Philosophy in Philosophy-Neuroscience-Psychology

Washington University in St. Louis 2021

John Doris, Co-Chair

Casey O'Callaghan, Co-Chair

Elizabeth Schechter

Ron Mallon

Allan Hazlett

Laurie Paul

According to standard conceptions of agency, our reasons and intentions guide our actions. That is, goal-directed intentions play a key role in practical deliberation, planning, and execution of action. Furthermore, purposeful, goal-directed behavior warrants attributions of responsibility or "reactive attitudes" like resentment, anger, gratitude and forgiveness. However, recent developments of the dual-process theory of mind cast doubt on the empirical adequacy of this picture. While people take themselves to be responding to relevant reasons, they are often bypassed by irrelevant affective or automatic reactions. In this work I go beyond the dual-process theory of mind to offer a mechanistic account of the role of emotion in agency and practical deliberation. In particular, I show that emotion undermines our reflection by structuring our agential point of view, skewing our reactive attitudes, and preparing us for action at the expense of accurate self-awareness. I conclude by offering a way of developing agential skills without relying on accurate reflection. In particular, I show how increasing self-diversity can foster cognitive flexibility and the ability to form appropriate affective reactions.

Chapter 1: Introduction

According to everyday conceptions of agency, our reasons and intentions make a real difference to how we act (e.g., Malle 2004). According to this conception, there is a close connection between intentional action and acting for a reason (Anscombe 1957; Davidson 1963). Dating back to Aristotle, the standard conception of agency suggests that reasons-based changes to the world could be rationalized from the agent's point of view (i.e., by the agent's beliefs and desires that could be cited as justifications for her actions). That is, goal-directed intentions play a key role in practical deliberation, planning, and execution of action (e.g., Mele 1992, Bratman 1987).

How does an agent act from her point of view on the world? Many philosophers believe agency to be possible due to a single faculty of *practical reason.* For example, Aristotelians allot accurate *reflection* a key role in our ability to govern our lives in the development of virtue (e.g., Annas, 1993). Likewise, philosophers in the Kantian tradition celebrate the *reflective agent* as someone capable of self-direction (Wallace 2003; Velleman 1989). More recently, Frankfurt famously distinguished the *reflective agent* from a mere Wanton, who simply follows his unquestioned desires to action (Frankfurt, 1988).

In particular, Frankfurt delineates distinctly human agency with the concept *of reflective endorsement*. i.e., reflection that allows one to form second-order desires about one's first-order desires. In other words, what it means to endorse a desire is to reflect on the type of person one wants to be and form higher-order desires or volitions for a certain first-order desires to be effective (Frankfurt, 1988).

That is, unlike thoughtless automatons, humans engage in flexible, goal-directed behavior. Human behavior is flexible exactly because it is guided by accurate awareness of the relevant features of our psychology, as well as accurate awareness of the relevant features of the environment (e.g., Austin and Vancouver, 1996). For instance, a "person who does not know her motives is missing something she would need to deliberate well." (Velleman 2000, p. 12).

Conceptions of reflective agency permeate our everyday lives. For example, learning that someone accidentally stepped on your foot usually causes you to adjust your indignation. That is, purposeful, goal-directed behavior warrants attributions of responsibility or "reactive attitudes" like resentment, anger, gratitude and forgiveness (Strawson 1962; Vargas 2004, Watson 1993).

In sum, according to popular conceptions of agency, we act for reasons and have the capacity to make reasons-based changes to the world. Along the same lines, we have the capacity to discern relevant features of the world and relevant features of our psychology in following our reasons to action.

However, recent findings in neuroscience and cognitive science cast doubt on the empirical adequacy of this conception of agency. For example, recent challenges from *dual-process* theories of mind suggest that *System 1* automatic affective reactions drive judgment and action, while *System 2* reasoning merely offers post hoc rationalizations. In particular, dual-process theories of cognition suggest that two systems play a role in evaluation. While System 1 is unconscious, automatic, and affectively charged; System 2 is slow, conscious, and affectively neutral. Furthermore, System 2 cannot change concurrent automatic affective processes, but only offers post hoc rationalizations after the judgment has already been reached.

As a result, causes and reasons for action are often incongruent (Doris, 2015; Greene, 2008; Haidt, 2001; Prinz, 2009). That is, while people take themselves to be responding to relevant reasons in their psychology and the environment, they are often bypassed by irrelevant affective or automatic reactions (Cameron et. al., 2013). This kind of ignorance is particularly troublesome for the standard conception of agency if the agent would in fact disavow her motives, were she made aware of them.

Nevertheless, proponents of the standard conception of agency resist this skepticism on several grounds. First, mechanistic details behind the dual-process theory remain rather mysterious. Details about the role of emotion and automaticity remain highly disputed (e.g., Mugg, 2016). In particular, May argues that both affect and reasoning could be somewhat automatic and contain cognitive elements (May, 2020). Along the same lines, even if deliberation cannot always penetrate concurrent evaluation, perhaps there is a way for reasoning to offer a *top-down* influence on automatic processes through practice (e.g., Osman 2004; Keren and Schul, 2009).

In this work I consult empirical findings to offer several new arguments for incongruence not subject to these limitations. In chapter 2, I offer a detailed account on the role of consciousness in deliberation. In chapter 3, I outline the role of affect in our agential point of view on the world. In chapter 4, I outline the role of affect in attributing propositional attitudes like beliefs and desires towards oneself and others. In chapter 5, I argue against the view that emotions serve as epistemically significant representations of the world that could inform deliberation. I conclude by offering a way to develop agential skills without relying on accurate reflection. Let’s take a look at this outline in more detail.

In chapter 2, I argue that not all feelings represent changes in the world. Specifically, phenomenology of mental effort or the feeling of working hard mentally is not a matter of tracking features in a certain way. As a result, it does not appear that all phenomenal states represent features of the world.

This argument does not bode well for standard conceptions of agency. In particular, feelings of working hard mentally are commonly taken as indicators of intentional deliberation. Since mental effort comes and goes without our say, we do not always know whether we are performing mental tasks intentionally. Hence, our sense of agency in deliberating intentionally is under threat.

However, perhaps the claim that we do not know whether we're exerting conscious control can be reconciled with our commonsense conception of agency. That is, perhaps it is possible that the relevant mental attitudes could still play the right role in the exercise of agency, even if they are not consciousness accessed and marked by the phenomenology of mental effort (e.g., Pacherie 2008). In other words, it is possible that we can act for reasons without real-time conscious control and explicit deliberation. Nevertheless, the commonsense picture of the reflective agent deliberating hard in foraging self-direction is somewhat misguided.

In chapter 3, I offer a mechanistic account of the role of emotion in our point of view on the world. I argue that emotion plays a key role in ensuring a unified perspective on the world. In particular, emotion creates and sustains a center/periphery structure in our conscious perspective on the world by signaling orders of perceived importance as well as orders of perceived changeability (i.e., felt actionability).

This argument does not bode well for the standard conception of agency. In particular, the standard conception assumes that perception has a special epistemic significance in putting us in touch with the world. However, if not all consciousness is perceptual, not all consciousness works to put us in touch with the world. Hence, not all consciousness serves as a defeasible source of justification. In short, if consciousness motivates action but still does not serve as a defeasible source of justification, our capacity to act for reasons is under threat.

Furthermore, standard conceptions of agency assume that in order to make intentional, reasons-based changes to the world, we need to feel like our reasons for action are *actionable*. However, since affect determines which options we feel to be actionable, reasons could be epiphenomenal in effecting change in the world. Hence, our capacity for making reason-based changes to the world is under threat.

Nevertheless, these arguments still do not preclude the possibility of *top-down* influence on lower-level processes to create additional points of view on the world. In other words, even if certain reasons-based actions feel impossible, perhaps the agent could engage in emotion regulation to *animate* his reasons for action.

In Chapter 4, I argue that the affective nature of intersubjectivity creates a burden for emotion regulation. In turn, the capacity for attributing propositional attitudes to ourselves and others evolved to help us navigate this affective burden. However, while attributing propositional attitudes in navigating the social hierarchy provides temporary relief against intersubjective immersion, it spells trouble for the epistemic significance of propositional attitudes.

This argument does not bode well for the standard conception of agency. In particular, proponents of the standard conception of agency believe that making reasons-based changes to the world requires the ability to attribute appropriate propositional attitudes to oneself (ascribing the right beliefs, desires, and intentions). If attributing propositional attitudes to oneself is skewed by the need to navigate the perceived social hierarchy as well as the need for emotion regulation, our ability to make reasons-based changes to the world is under threat.

Similarly, the standard conception of agency requires the ability to form appropriate reactive attitudes towards oneself and others. Forming reactive attitudes requires the ability to attribute appropriate propositional attitudes towards oneself and others. However, if forming reactive attitudes is systematically skewed by irrelevant causes like perceived social rank and the need for emotion regulation, our status as moral agents is under threat.

In this chapter, the reply from *top-down* control of higher-order processes loses a bit of traction. Since the very formation of higher-order attitudes is itself skewed by irrelevant causes like the need to navigate perceived social ranking and the need for emotion regulation, our capacity to exert top-down control is under threat.

In Chapter 5, I argue against the view that emotions serve as epistemically significant representations of the world that could inform moral deliberation. In particular, I argue that emotions function to prepare us for action by activating other instrumental emotions. Since we cannot distinguish between “genuine” and instrumental emotions, and instrumental emotions remain resistant to reappraisal, the agential role of emotions undermines their epistemic significance. As a result, emotions’ ability to inform moral judgment is under threat.

This argument does not bode well for the standard conception of agency. In particular, some philosophers resist the skepticism from the dual-process theory of mind by arguing that emotions, as well as reasons, could serve as epistemically significant representations of the world (e.g., May, 2020). However, if emotions’ agential component undermines their epistemic significance, emotions indeed compromise our ability to makes reasons-based changes to the world.

As in chapter 4, the reply from *top-down* control loses a bit of traction. In particular, since the distinction between “genuine” mental states and instrumental mental states is not available to introspection, one cannot tell whether there is *any need* for top-down control. As a result, the ability to exert top-down control in deliberation is under threat.

I conclude by offering a way to cultivate agential skills without relying on accurate deliberation. In particular, I propose a way to increase cognitive flexibility by limiting the emotion regulation workload of instrumental emotions. I show that we can learn to limit the emotion regulation workload of a given set of instrumental emotions by cultivating a diverse set of values. For example, if ingroup pride is the only thing getting you through the day, it could be beneficial to form a number of other commitments to offer support while re-examining ingroup pride.

Importantly, the capacity for cognitive flexibility is distinct from the capacity for accurate deliberation. Since emotions prepare us for action by triggering instrumental emotions, even accurate beliefs and emotions could foster cognitive rigidity by playing instrumental roles. Hence, even accurate emotions and beliefs could fail to be appropriate is they compromise cognitive flexibility (i.e., play an inappropriate instrumental role in our mental economy).

Chapter 2: Tracking Intentionalism and the Phenomenology of Mental Effort

2.1 Introduction

Most of us are familiar with the phenomenology of mental effort accompanying cognitively demanding tasks, like focusing on the next chess move or performing lengthy mental arithmetic. In this paper, I argue that phenomenology of mental effort poses a novel counterexample to tracking intentionalism, the view that phenomenal consciousness is a matter of tracking features of one's environment in a certain way.

Intentionalists argue that, necessarily, any two mental states that differ in their phenomenal properties will differ in their representational properties.[1]A common way of arguing against intentionalism is to present experience pairs that differ in phenomenal character but fail to differ in representational content. In similar fashion, I argue that an increase in the phenomenology of mental effort (henceforth PME) does not accompany a change in any of the following candidate representational contents: a) representation of externally presented features, e.g. brightness, contrast, and so on b) representation of task difficulty, c) representation of the possibility of error, d) representation of trying to achieve some state of affairs, e) representation of bodily changes like muscle tension, or f) representation of change in cognitive resource availability and lost opportunity cost.

[1] Supervenience relation characteristic of weak intentionalism is a natural starting point because (i) any stronger relationship between the two will entail the supervenience relation (e.g., if these two types of properties turn out to be identical, this will entail that the phenomenal at least supervene on the representational), and (ii) any hopes of giving a naturalistic account of phenomenal properties requires delineating which facts about mental state form the minimal supervenience base

I focus on what is commonly taken to be the most promising reductionist intentionalist theory of phenomenal consciousness, i.e., *weak global* intentionalism conjoined with a tracking theory of intentionality, broadly defended by David Armstrong, Fred Dretske, Christopher Hill, William Lycan, and Michael Tye.[2] By offering a way to reduce *all* consciousness to tracking, tracking intentionalism offers a promising physicalist theory of consciousness.[3] Since many believe that this view offers the *best chance* of naturalizing consciousness, these arguments deserve detailed consideration (Pautz, 2013; Cutter and Tye, 2011).

2.1.1 What is Intentionalism?

Intentionalism posits a close relationship between phenomenal and representational properties. Phenomenal properties of mental states concern *what it is like* for a subject to experience certain phenomenal properties. For example, when I look at the gray cat on my desk, my visual experience has a distinctive phenomenal character. In virtue of what does my visual experience of the gray cat has its distinctive phenomenal character?

According to intentionalism, my experience of the gray cat has its distinctive phenomenal character in virtue of the way it *represents* the world as being (Cutter and Tye, 2011). Roughly, this reductionist program operates in the following way: all phenomenal consciousness is spelled

[2] The supervenience claim could take one of various forms: (a) local vs. global; and (b) intramodal vs intermodal. The first distinction concerns whether supervenience only holds for a certain class of mental states or for all mental states, e.g., whether it holds only for visual perceptual states or for all states including moods. The second distinction concerns whether supervenience holds only for pairs of states of the same sense modality, or for any random pair of states across modalities. Notably, as Speaks (2015) points out, these distinctions are a matter of degree. Since the modality of this phenomenology doesn't allow straightforward delineation, the intramodal/intermodal distinction won't provide any traction. For the same reason, since we are exploring a new type of phenomenology previously untouched in the debate, we are concerned with global rather than local intentionalism given that only global intentionalists promise to find a supervenience base for *any* pair of phenomenal experiences *whatsoever.*

[3] First, TI reduces phenomenal states to intentional states, which in turn are explained in terms of the tracking relation of broadly physical properties and states if affairs. The reduction of intentionality to tracking enjoys recent popularity (Bourget and Mendelovici, 2013).

out in intentionalist terms, while intentionality is spelled out in tracking terms (Bourget and Mendelovici, 2013). For example, tokens of a state S in an individual x represent that p in virtue of the fact that: under optimal conditions, x tokens S if p, and because p (Cutter and Tye, 2011).

2.1.2 Past Alleged Counterexamples to Intentionalism

The main counterexample strategy for intentionalism is to refute the supervenience claim, i.e. to find a change in phenomenal properties of a subject's experience that is *not accompanied* by a change in the way the world is represented. A common pattern in intentionalist replies is to find some, however small, difference in the way the world is represented along with the phenomenal change in question. After seeing how the intentionalist handles past alleged counterexamples to intentionalism, we will be in a good position to consider whether a change in PME is also vulnerable to the same style of refutation, i.e., whether a change in PME is in fact necessarily accompanied by a change in the way putatively external objects are represented.

Consider the putative counterexample of blurred vision. If you take off your glasses or simply stop focusing your eyes, you will experience certain blurriness in your visual field. In other words, you will see objects *in a blurry way* without seeing them *as* blurry. According to Boghossian and Velleman (1989), what the experience represents may well be the same before and after taking off the glasses, but the phenomenology is different.

The intentionalist response to this example is a familiar one: there is a change in the way features of the external objects are presented. With blurry vision, the edges of objects have indistinct contours. There is less information about the putative location of the edges in the blurry condition than in the focused condition. Thus, there is indeed a representational difference as well as a phenomenal difference between the two conditions (Tye, 2006).

This response exploits the *gestalt* function of attentional mechanisms (i.e., allowing the appearance of certain features to become brighter at the expense of the brightness of other features, or gaining more information about some features at the expense of information about others).[4] For example, as you focus on the computer monitor in front of you, you inevitably "zoom out" of your surroundings.

However, empirical findings suggest that the *gestalt* function of attention is not a necessary feature of attention. Another function of attentional mechanisms rarely discussed in the philosophical literature is to handle *interference* in processing (henceforth interference). As we will see in the following section, the subject does not notice the phenomenal change associated with this type of attentional control via *tracking* the way features are presented.

This is exactly what we need for a counterexample to go through. We need introspection to reveal a phenomenal difference between two mental states without an accompanying difference in the way the world is represented. In the next section, I propose to take a closer look at the *interference suppressing* role of attentional control mechanisms previously untouched by the philosophical literature. What will emerge is the following dissociation: the interference suppressing role of attention both (i) results in the phenomenology of mental effort, and yet (ii) does not change the way putatively external features are presented.

2.1.3 Why PME Poses a Problem for the Intentionalist

Cognitive neuroscience has long been interested in the way the brain handles challenging and non-routine situations. These cognitively taxing situations have been most studied in association

[4] Evidence for the gestalt function of attention could be found in studies showing that attention enhances the perception of low-level visual features including contrast (e.g., Carrasco, Ling, & Read, 2004).

with one of three contexts: (1) tasks that require the overriding of prepotent responses (2) tasks that require selection among a set of equally permissible or under-determined responses, or (3) tasks that involve the commission of errors. A conflict in attentional mechanisms can occur when an automatic response needs to be inhibited, or no automatic response is available (e.g. working memory tasks like problem-solving) (Botvinick et al., 2004).

Importantly, it has been shown that these types of mental processes are typically associated with a subjective feeling of mental effort (Naccache et al., 2005; Dehaene, Kerszberg et al., 1998). The main working paradigm for studying mental effort is to treat it as a result of interference in information processing. Certain processes interfere with one another for a variety of reasons. Some of these reasons are determined by the organization of our *cognitive* architecture. For example, the extent to which certain processes become automatized and are subsequently hard to override is determined by our cognitive organization.[5] Other reasons involve the organization of our *neural* architecture.

Van Veen et al. (2001) characterize the plurality of interference types in the following way:

> In cognitive psychology, information processing is often thought of as occurring at a number of different levels, which might correspond to the different phases of task processing, for example (i) stimulus encoding, (ii) target detection, (iii) response selection, and (iv) response execution. Theoretically, conflicts might occur at any or all of these levels (Van Veen et al., 2001, p.1302).

While (i), (ii), and (iv) might affect the way putatively external features are presented, (iii) does not (Milham et al., 2001). To make this more concrete, let's consider an example of PME where the *gestalt* features of attention are precluded by the conditions of the experiment.

[5] Botvinick et al. (2004), Cohen et al., (1990).

Saenz et al. (2003) explored the nature of information-processing interference while using the conflicting features of direction and speed.

Consider the description of the relevant condition in the experiment:

> Observers were instructed to divide attention equally across two stimuli placed to the left and right of a central fixation point. In the first experiment, each stimulus was a circular patch consisting of two transparently overlapping fields of upward and downward moving dots. Subjects concurrently performed a speed discrimination task on one field of dots from each side, either moving in the same direction (up or down on both sides) or in different directions (up on one side and down on the other). Thus, without changing the visual display or the spatial distribution of attention, subjects divided attention across stimuli composed of either a common feature or opposing features.

As you might have guessed, subjects experienced greater PME when they had to divide attention across different directions of moving dots than when they had to divide attention across the same direction of moving dots. Processing the speed of the targets feels more effortful when they move in different directions because speed and direction are processed by the same population of neurons (Maunsell and Newsome, 1987). To eliminate confounding variables, the authors made sure that no change in the way features are presented occurs:

> We used overlapping stimuli that were identical in all conditions so that differences in task performance could not be confounded with changes in the stimulus itself or with changes in the spatial distribution of attention (ibid, p.633)

Interestingly, what is considered easy in the ‘easy’ condition is that subjects have to track either *only* the upward moving dots or *only* the downward moving dots on either sphere. In order to do that, the subjects inevitably “zoom out” of the dots moving in the direction that is not relevant to the task at hand. Hence, they can simply learn to *not see* task irrelevant features. However, in the "difficult" condition, this "gestalt" switching does not take place since subjects have to track *both* the upward moving dots on the right side, and the downward moving dots on the left side. Since subjects are *primed* by each feature on the opposing side, "zooming out" is not possible. Subjects can't simply learn to ignore a certain feature (either upward motion or downward motion), since it is exactly what they are supposed to be tracking on the other side.

Importantly for our purposes, the phenomenological effect of interference is felt prior to a decrease in discriminatory performance. Hence, a change in phenomenology proceeds independently of a change in the way features are presented (Morsella et al., 2009, Milham et al., 2001; Van Veen et al., 2001).[6]

Saenz et al.'s findings provide some reason to doubt the necessity of the supervenience claim. A change in PME is not tracked via a change in the presentation of external features. Now that we have a handle of the type of phenomenology at stake, I will attempt to *further* the intentionalist account by offering ways of spelling out PME in intentionalist terms. If these accounts come up short, intentionalists face an uphill battle for finding a supervenience base for PME.

[6] Although the subjective ratings of mental effort were not the focus of the study, the validity of their measuring techniques has precedent in the empirical literature (Morsella et al., 2009). Specifically, to further explore the subjective dimension of cognitive effort, the Morsella et al. introduced the following paradigm for measuring subjective effects of interference. Subjects were trained to introspect the particular feeling associated with incongruent conditions of the Stroop task. This introspection training was done to ensure that "participants were introspecting the same thing during both flanker and Stroop tasks" (ibid, p.10). The experimenters found more subjective effects for incongruent conditions than for congruent conditions. Furthermore, these changes in phenomenology were reported prior to changes in response time. Interestingly, the phenomenology itself proved almost ineffable to the subjects (ibid, p. 16)

2.2 Obvious Intentionalist Replies

Before presenting more lengthy replies, I will present a few possible replies that could be handled with relative ease. First, one might wonder if the relevant PME change in the difficult condition supervenes on a change in judgment. A change in judgment would surely qualify as a change in the way the subject is representing the world. For example, as the subject finds herself experiencing more mental effort, she could go from judging that "this task is hard on level 1" to "this task is hard on level 2.".

Several replies are available. First, numerous experiments have shown that evaluation of task difficulty is dissociable from the experience of mental effort in *both* clinical and normal populations. If PME evolved to track task difficulty in the same way pain evolved to track potential or actual bodily damage, an absence of PME during cognitively demanding tasks should spell some type of malfunction. Consider the following characterization of TI: tokens of a state S (PME) in an individual x represent that p (task is difficult) in virtue of the fact that: *under optimal conditions*, x tokens S (PME) iff p (task is difficult), and because p. Hence, if subjects are tokening PME without representing the task to be difficult, or if subjects are not tokening PME if they're representing the task to be difficult, there must be a malfunction in the attentional mechanisms. Indeed, it is easy to see how subjects unable to feel pain are at an evolutionary disadvantage (Tye, 1995). Does an absence of PME during cognitively demanding tasks spell malfunction of the attentional mechanisms?

Naccache et al. (2005) show that a subject's attentional and cognitive abilities could be functioning normally despite (i) awareness of task difficulty level, as well as (ii), an absence of PME. The experimenters record the subject saying, "Yes, I can see how this task is a tricky one," but the subject still reports no feeling of PME.[7] Furthermore, optimal cognitive performance in normal subjects characterized by "flow states" involve awareness of task difficulty in the absence of PME (Nakumura and Csikszentmihalyi, 2014).[8]

Second, it appears that experiencing the task as hard is simply re-describing the phenomenology in question. Yes, you feel more effort, and this allows you to say that the task is hard, and then even harder, but this seems nothing more than introspecting on your phenomenology and finding words for it, i.e., "this feels a certain way, so it must be because the task is hard." However, if the intentionalist wants to provide a non-circular basis for a representational difference, it can't be a mere *re-description* of the phenomenology in question. Otherwise, there is a built-in response to *any* phenomenology presented as a counterexample. As Speaks (2015) points out, this built-in response risks trivializing intentionalism. Noticing a change in your phenomenology should not be the *basis* for delineating a representational change - it should be the other way around.

Relatedly, one might notice the amount of time it's taking to complete the task and thereby infer that it must be a difficult one. However, this appraisal could have nothing to do with the presence of PME. Nonchallenging, boring tasks tend to feel as if they're stretching on for a long time while both (i) not triggering PME, and (ii) leading you to infer that they might have been difficult.

[7] Naccache et al. write that "Unexpectedly, control abilities of patient RMB evaluated in various versions of this Stroop tasks were amazingly preserved… we could see the presence of an efficient dynamic regulation of control abilities as indexed by Gratton and proportion effects" (ibid, p.1319).

[8] Mental flow states are characterized by the experience of mastery or feeling "in the zone". For example, if playing a musical instrument and sight reading, the subject might be aware of how smoothly the experience is going. There might be associated feelings of control, i.e. being able to adjust to other players at short notice. In contrast, phenomenology of mental effort surfaces when one is aware that things are being presented smoothly, and yet there is an added awareness that it is costing some effort (even if you're not quite sure how to "apply" it). (Csikszentmihályi et. al, 2005). Thus, it is the complete opposite of the experience of transparency.

Another straightforward reply on the behalf of the intentionalist is that an increase in PME is accompanied by a change in judgement about the possibility of erring. Since PME is often felt during task performance, task parameters easily delineate possibilities of erring. This judgement change could be characterized in the following way: “Whoa! I was even closer to messing up this time, but fortunately, I skated by!”

Nonetheless, it appears that the phenomenology of judging the possibility of erring is distinct from the phenomenology of mental effort. Consider the following scenario: you are asked to predict whether a coin will land heads or tails. Having made three successful predictions so far, you are asked to call it once again. You experience an even greater feeling of the possibility of erring. Even if you are just as likely to get it wrong as you are to get it right, you feel that this time around your luck has run out and you’re sure to get it wrong. However, even if you feel that you’re very close to erring, this phenomenology seems distinct from the phenomenology of mental effort normally experienced during solving a difficult math problem. This resultant phenomenology will be something like excitement or dread, but not mental effort. Hence, it is unlikely that an increase in the phenomenology of mental effort is tracking the possibility of erring.

A related proposal with some intuitive appeal is that the phenomenology of mental effort tracks *trying* to achieve a yet unrealized outcome or state of affairs.[9] In exploring this proposal, a few points are worth making. First, empirical findings suggest a double dissociation between goal pursuit and PME. Goal pursuit often happens outside consciousness (Custers and Aarts, 2010;

[9] I thank an anonymous referee for this suggestion.

Zedelius et. al., 2012). This pursuit might trigger mental effort in the absence of goal representation.

On the other hand, conscious goal pursuit could happen in the absence of PME. For example, subjects experiencing mental flow are aware of pursuing and completing various goals but experience no accompanying PME (e.g. Csikszentmihályi et. al., 2005).

One way for the intentionalist to make sense of this data is to suggest that despite these cases, in the paradigm case, when a subject experiences mental effort, she represents herself as trying to achieve some goal. However, it seems that the phenomenology associated with representing oneself as trying to achieve some goal is distinct from the phenomenology of mental effort.

Going back to our previous example, I could represent myself as trying to predict whether the coin will land heads or tails. However, my *trying* is merely due to the uncertainty of the outcome and not the effort in bringing it about. Perhaps the intentionalist proposal could be refined to delineate the representational base of PME as "trying to achieve something difficult." However, the same type of reply is available. It seems that *trying* to predict whether the die will land on a certain number is difficult, but it does not involve any phenomenology of mental effort.

One might point out that what distinguishes these cases from instances of PME is the following: while you are a passive observer to the coin tosses, you're an active participant in exerting mental effort. That is, one might suggest that you're actively doing something to control the outcome when you're experiencing PME. Hence, PME is akin to the phenomenology of agency.

However, I would like to distinguish the phenomenology of mental effort from the phenomenology of agency. While the latter involves a feeling of *control*, the former does not.

Consider the following example of agentive phenomenology. As you apply more effort to pedaling, you're simultaneously aware of the effect this effort has on your muscles and the pace of the bike. Tracking the correspondence between effort and represented changes allows a feeling of control.[10]

On the other hand, no such correspondence exists between *mental* effort and the representation of worldly changes. Attempting to increase mental effort does not result in a reliable clarity increase. For example, a chess player could experience increasing levels of PME, but see no increase in the clarity of the problem. Try as we might, it is not always the case that we can simply "concentrate harder" when attempting a difficult mental problem. Unlike squinting, we can't simply "flex our mental muscles" in order to "see" the solution a little bit clearer.

Furthermore, even if no apparent change in the world is tracked, feelings of control in the phenomenology of agency still involve tracking some reliable effect our agential acts have on the represented state of the agent herself. Consider the following example: say you are trying to move an enormous boulder while noticing no accompanying change in the world (your efforts are futile since the boulder is too heavy). Even though no apparent changes in the world are tracked, you still experience a reliable correspondence between *trying itself* and some state of affairs. For example, you're aware that the harder you push the boulder, the greater the muscle burn, and so on. Furthermore, in experiencing agential phenomenology, you feel a sense of control in increasing your levels of effort appropriate to the task.

[10] Christensen et al. (2015) illustrate the nature of agentive phenomenology in professional mountain biking. They argue that complex action involves a complex parametric structure (p. 344). For example, the agent is (somewhat) aware that applying pressure to the bike brakes influences the "control parameter" to change the speed ("target parameter"), which in turn could influence another parameter (e.g. curvature around the turn). So, a skilled agent can navigate the upcoming sharp curve via manipulating the amount of pressure she applies in order to manipulate the speed needed to make the turn.

In contrast, in trying to solve a difficult math problem, you are unable to control "flexing your mental muscles" in increasing degrees. You have no feeling of control in trying to exert mental effort (level 1), and then mental effort (level 2). You might have no idea how you went from level 1 to level 2 and back to 1 again! That is, you might slip in and out of *flow states* without being aware of actively trying to change your levels of effort (e.g. Csikszentmihályi et. al., 2005).

In sum, it appears that the phenomenology of agency is distinct from the phenomenology of mental effort. While the former involves a feeling of control, the latter does not. Hence, it does not seem that PME tracks *trying* to achieve some state of affairs. Now that we've gotten some shorter replies out of the way, let's move on to considering more lengthy replies on the behalf on the intentionalist.

2.2.1 Is PME Another Type of *Bodily* Phenomenology?

Some tracking intentionalists might want to argue that the phenomenology of mental effort is just another type of bodily phenomenology, and thus could be handled in the same way. Some familiar examples of bodily phenomenology include the phenomenology of certain localized pains, muscle aches, and so on. For example, if you suddenly spill hot coffee on your leg, you experience a burning sensation on the skin surface of your leg. According to TI, your attention goes to the location of the bodily disturbance. That is, your pain appears to be on or in your body at location X.

Now, we could plausibly ask: "Could PME represent that there is a bodily sensation of type *d* at location *l*"? *Embodiment* arguments are frequently used for theories of emotion.[11] Prinz (2005)

[11] Prinz 2005, Damasio 1994b, and Tye 2008.

argues that bodily changes associated with emotions can, in fact, *induce* the feelings of emotions and thus could be taken to be *sufficient* for emotional experience (Ledoux, 1996). Could a similar line of argument be used for PME?

Let's take a closer look at the empirical backbone of the embodiment hypothesis. One common feature associated with mental hard work is bodily tension. While working on a difficult math problem, some might notice a familiar tension in their neck. Bodily muscle tension could thus serve as a supervenience base for the PME, i.e., changes in PME require changes in the representation of patterned bodily muscle tension.

Unfortunately, this candidate does not hold up to empirical scrutiny. Studies have shown that patients paralyzed below the neck still experience mental effort while engaging in effortful motor imagery like mental rotation of three-dimensional objects (Alkadhi et al., 2005). Similarly, Cramer et al. (2005) show that subjects experience PME during mental imagery despite having lost all feeling below the neck.

Perhaps the embodiment hypothesis could be refined to survive these findings. Even if bodily muscle tension does not play a key role in PME, a feeling of tightening in the *facial* musculature could serve as a supervenience base for PME. After all, knitted brows are a familiar sight in academic settings. The facial feedback hypothesis has enjoyed some standing in the empirical literature on emotion. Several studies have shown that when subjects are induced to make a certain emotion-specific facial expression (grimacing, frowning, etc.), they report experiencing the corresponding emotion (disgust, anger, and so on).[12]

[12] Duclos and Laird 2001; Duclosetal, 1989; Edelman 1984; Flack et al.1999; Kellerman and Laird 1982; for extensive review, see Laird and Bresler 1992; see also Niedenthal 2007; Niedenthal and Maringer, 2009).

Unfortunately, the facial feedback hypothesis has run into serious replication problems. The facial feedback hypothesis states that affective responses can be influenced by facial expressions, (e.g., smiling), even in the absence of emotional experiences. Specifically, Strack et al. (1988) had participants rate the funniness of cartoons while holding a pen in their mouths, thus inducing a “smile”. Surprisingly, ‘smiling’ subjects rated the cartoons as significantly funnier than when they held the pen with their lips (inducing a ‘pout’). However, this seminal study of the facial feedback hypothesis has not been replicated. In fact, the results of 17 independent direct replications of Study 1 from Strack et al. (1988), have failed to replicate (Wagenmakers et al., 2016).

Further trouble for the facial feedback hypothesis stems from case studies of subjects with facial paralysis. For example, Keillor et al. (2002) showed that a patient (F.P.) suffering from bilateral facial paralysis could still report normal emotional experiences despite her inability to convey emotions through facial expressions.

While the generalizability of results drawn from a few case studies pulls little weight, inference to the best explanation leans against the facial feedback hypothesis for PME. That is, our intuitions in the thought experiments, the replication troubles of the facial feedback hypothesis, as well as outcomes of individual case studies of facially paralyzed patients all suggest that the embodiment thesis is an unlikely option for the intentionalist.

2.3 Representing Scarce Resource Depletion and Opportunity Cost?

So far, I have argued that PME does not seem to be directed at any externally presented property. Perhaps tracking intentionalists could argue that PME represents a fact about the organism, e.g. cognitive or neural resource availability in relation to lost opportunity cost (Kurzban et al., 2013; Westbrook, 2015).

A clarificatory note is in order. So far we have honored tracking intentionalists' commitment to phenomenal externalism. According to phenomeal externalism, phenomenal properties are not in the mind but are out there in the world, (or on the subject who is in the world). This section will branch out to a more recent development of phenomenal externalism in the tracking intentionalist literature. That is, phenomenology doesn't only track seemingly external object features like bodily damage and brightness, but also tracks general *states of affairs.*

TI's ability to account for tracking states of affairs has enjoyed some recent success (e.g., Hill, 2009). Cutter and Tye (2011) argue that pain phenomenology, i.e., the experienced badness of pain, not only tracks bodily damage but also its threat to the organism's well-being:

> Our pain experiences do not just represent the presence of tissue damage, but also (roughly) represent our tissue damage as being bad for us to some degree. This view, we argue, is independently motivated by the phenomenology of pain experience, and we show how it is consistent with, and indeed predicted by, the tracking theory of intentionality (ibid, p. 91).

Analogously, intentionalists could argue that in representing cognitive resource depletion, PME tracks a state that is "bad for us to a certain degree." In this case, an increase in PME tracks the opportunity cost associated with continued resource expenditure.

Consider the following fodder for this hypothesis. Kurzban et al., (2013) argue that phenomenology of effort is associated with a cost/benefit computation. They theorize that given resource limitation; executive resource allocation should come with phenomenological tagging. If any resource is limited, its allocation carries an opportunity cost, i.e. the more resources are deployed in a particular task, the less resources could be deployed elsewhere. Hence, increased resource allocation should come with greater phenomenological reminder that perhaps these resources should be either (i) conserved, or (ii) redeployed elsewhere. Analogously, if physical energy expenditure and the related opportunity cost came with phenomenological tagging, it might increase one's chances of survival. If one could feel that his current task is using up the last of his energy resources, he would have a strategic advantage, e.g., entering confrontation with an inaccurate estimate of energy availability could prove deadly.

Perhaps intentionalists could extend this line of reasoning to PME. If mental resources are finite, their use should be phenomenologically marked in service of formulating competitive strategies. Hence, an increase in PME represents or tracks opportunity cost associated with continued resource expenditure.

Let's take a closer look at the predictions of this view. If PME tracks opportunity cost associated with continued resource expenditure, the greater the PME, the greater opportunity cost, as fewer resources remain for other tasks. Presumably, the longer this goes on, the harder it gets to recruit more resources to the task.

However, these predictions aren't exactly borne out. In fact, studies show a *decrease* instead of an *increase* in PME following ongoing resource allocation (Botvinick et al., 2001, Carter et al., 2007). Using the Stroop paradigm, Botvinick et al. (2001) showed that incongruent (i.e. cognitively demanding) trials induce *more* PME when such trials are *rare* in comparison to congruent trials in a given set. If hefty resource allocation is rare, they require a total of *less* resource allocation. If they require a total of *less* resource allocation, they should introduce *less* opportunity cost (since a lot more resources remain to potentially redeploy elsewhere), and trigger *less* rather than more PME.

Consider the following analogy. You are asked to jog lightly around the track with intermittent sprinting intervals. In scenario one, you jog lightly for two laps with only two 100-meter sprinting intervals at the end of each lap, totaling in 600 meters of light jogging and 200 meters of sprinting. In scenario two, you sprint every other 100 meters, totaling in 400 meters of sprinting and 400 meters of jogging. Which scenario would produce greater feelings of effort: 600 meters of jogging combined with 200 meters of sprinting or 400 meters of jogging combined with 400 meters of sprinting? I can attest that the second scenario would induce greater amount of effort phenomenology than the first.

In sum, if PME is supposed to track opportunity cost associated with resource expenditure, and fewer resources are used if a smaller number of demanding tasks are being performed during a set interval of time, then subjects should be experiencing *less* rather than *more* PME during sets with *fewer* mentally demanding trials.

Bayne and Levy's (2006) account of mental effort provides yet another option for the intentionalist. Bayne and Levy categorize mental effort as a component of agentive experience. According to the authors, "the experience of mental effort involves a *representation* of the utilization and progressive fatigue of mental muscles" (ibid, p.17). In other words, while undergoing PME, we represent our use of "mental strength" and the effect this use has on our remaining mental resources. Bayne and Levy characterize the phenomenology of mental effort in the following way:

> Anyone who has struggled with a difficult conceptual issue has experienced the effort involved in thinking a problem through. It gives rise to characteristic feelings of tiredness and a growing urge to stop (ibid, p.13).

This thesis fits well with the opportunity cost view outlined above. After all, resource finitude is an essential component of opportunity cost. If the resources are not limited, their allocation does not involve an opportunity cost.

Before going over the empirical details behind this hypothesis, two distinctions should be made clear. First, as discussed in the previous section, likening PME to agentive phenomenology appears to be on the wrong track. While the latter involves a feeling of control, the former does not. Second, it is plausible to distinguish between *fatigue* phenomenology and *effort* phenomenology. The former involves an awareness of current resource availability, while the latter does not. In particular, when you are feeling fatigued, you are aware that you are close to not being able to carry on. On the other hand, when you experience effort, you might be unaware that you can't carry on. For instance, say you are running and decide to suddenly increase your pace considerably. Unless you're not used to running, you could very well be aware of applying effort, but remain ignorant that you can't keep up this pace for a long time. On the other hand, if you are running and suddenly experience fatigue, you immediately feel that you cannot keep up for much longer. Similarly, the phenomenology of mental effort does not have to be accompanied by an awareness that you will not be able to carry on for much longer. For example, you could feel pleasantly challenged by a crossword puzzle and remain optimistic about your ability to carry on indefinitely.

Nevertheless, for the purposes of this discussion, I am willing to set this distinction aside and discuss Bayne and Levy's thesis as a thesis about the phenomenology of mental effort. To review, Bayne and Levy argue that mental effort phenomenology is a type of agentive phenomenology that involves “a representation of the utilization and progressive fatigue of mental muscles”.

While this characterization seems like an attractive option for delineating the representational content of PME, it still falls short. Consider the following summary of its shortcomings:

> (a) Unlike the individuals who are unable to feel pain, those who lack PME function perfectly well. These cases suggest that the phenomenology of mental effort did not evolve to track cognitive resource depletion (i.e., alert the organism of the utilization and progressive fatigue of mental muscles).
> (b) Unlike physical energy use, frequent “mental energy” use in a given interval of time produces *less effort* than infrequent energy use.
> (c) No systematic relation indicative of *tracking* could be established between cognitive resource depletion and PME.

Let us take a closer look at (a). An increase in PME is supposed to track increased mental resource expenditure because PME *evolved* to signal resource expenditure.[13] Hence, an absence of PME would spell malfunction. However, as discussed in the previous section, studies show that subjects lacking PME function well.[14] Critchley et al. (2003) found that the patients missing PME had well-preserved general intellectual functions. They performed generally well on many demanding clinical tasks sensitive to frontal executive functions. In sum, while nothing precludes us from speculating that PME might turn out to have other (possibly derived) adaptive value, tracking cognitive resource depletion does not seem to be it.

[13] Recall that according to TI, representation is grounded in evolutionary histories. For example, pain experience tracks bodily damage and badness because it was selected to carry information about bodily damage and badness by reliably correlating with these features of the environment (Cutter and Tye, 2011).
[14] To review, Naccache et al. (2005) show that subjects’ attentional and cognitive mechanisms function normally in the absence of PME.

We have already encountered (b) in discussing the opportunity cost account above. Briefly, if PME tracks mental resources expenditure, an increase in PME should signal more resource expenditure. However, studies show a decrease instead of an increase in PME following ongoing resource allocation (Botvinick et al., 2001, Carter et al., 2007).

A further hurdle for Bayne and Levy's hypothesis involves narrowing down the notion of cognitive resources. After all, a *tracking* relation should be associated with some measure of systematicity. If the concept of cognitive resources defies systematic treatment, then Bayne and Levy's intentionalist characterization of PME is on the wrong track. The systematic nature of the tracking relation is nicely summarized by Hilbert and Klien (2014) in the following way.

> Given very general assumptions about physiology and the evolution of nervous systems, what is to be expected is that the internal states that *track environmental features* will have some systematic structure (ibid, p. 300)

How are we to understand the notion of "cognitive resources" and their systematic depletion? One non-metaphorical answer is that cognitive resources could be understood in neural terms. In other words, cognitive resources are neural changes underlying cognitive resource depletion.

Is there a systematic relationship between neural resource depletion and PME? In order to answer this question, we need to take a closer look at how the current scientific community is talking about concepts like "finite capacity" and "neural resources." What exactly is a "limited cognitive resource" and in what way can it become depleted?

Explaining degrading performance with increased task difficulty has typically been explained by metaphorical allusions to "finite working memory capacity" that is spread more "thinly" with increased "task load" (Baddeley, 1996). However, any systematic or *structural* relationship between these placeholders is yet to be empirically demonstrated. Many researchers suggest that what creates a tax on attention control processing varies across contexts (Franconeri et al., 2013). Perhaps the limitation comes from the nature of neural processing prevalent throughout the cortex, i.e., surround inhibition of an activity 'peak' of some neural processes on other processes in their neighborhood. Franconeri et al. summarize this point in the following way:

> Items interact destructively when they are close enough for their activity profiles to overlap, due to the inhibition zone that typically surrounds each activity peak. These suppressive surrounds sharpen the activity profiles of single items and resolve inter-item competition–a critical step especially when unitary actions are needed (e.g., a saccade to a single location) (ibid, p.3).

In this two-dimensional 'map' architecture, capacity is not fixed but flexible, determined by a number of fluctuating variables, i.e., the space taken by the activity profile of certain items on the map, how these items interact with one another (e.g., whether they interact constructively or destructively is partly determined by the inhibition zone that surrounds each activity peak), the spacing on the items on the map, and so on. All these factors contribute to the competition between items at the neural level. As this competition is resolved, a more pronounced activity peak takes place. Furthermore, the way these competitive interactions resolve is not fixed but flexible. Hence, alluding to a limited capacity that somehow becomes systematically depleted with use is misleading. This organization gives the brain an adaptive edge, but it also has side effects. Perhaps PME is one such side effect, alerting the organism to no particular state of affairs.

The upshot is that cognitive neuroscience does not provide evidence of a match in structural relations between PME and neural changes in the way we have seen between pain phenomenology and somatosensory cortex activation (e.g., Price et al.,1994, illustrate the structural relationship between pain intensity and neuron firings in S1). Since tracking relations are characterized by a match in systematic relations (e.g., similarity, differences, equal intervals, proportions), it does not appear that PME tracks cognitive resource depletion.

Inzlicht and Marcora (2016) summarize the criticisms of the resource depletion model (i.e., “the central governor”) in the following way:

> There is no credible evidence that mental effort actually consumes inordinate amounts of energy that are not already circulating in the brain. Recent modifications of the model make the central governor appear like an all knowing homunculus and unfalsifiable in principle, thus contributing very little to our understanding of why people tend to disengage from effortful tasks over time (*ibid,* p.1)

Hence, it does not appear that Bayne and Levy have provided sufficient reason to think that the phenomenology of mental effort tracks mental resource depletion.

2.4 Conclusion

To sum up, let us briefly review the main argument for TI and how it fares against PME. Arguments from introspective difference state that, necessarily, if there is an introspectable difference in the two phenomenal properties of subjects, then there is a difference in the objects and properties those subjects *represent as* in their environment. However, a difference in PME does not accompany a difference in either a) representation of the way external features appear, i.e., brighter, with more contrast and so on, b) representation of task difficultly, c) judgement of the possibility of error, d) representation of trying to achieve some state of affairs, e) representation of bodily changes like muscle tension, f) representation of cognitive resource depletion and opportunity cost.

While local intentionalism about some phenomenal experiences like pains might obtain, it does not look like it obtains for *all* phenomenal experiences. This puts the intentionalist in an uncomfortable position of trying to explain why *some* phenomenal experiences have representational content and not others. Moreover, even if as much as one type of phenomenal experience doesn’t have representational content, reductionist theories of consciousness are under threat.

References

Aarts, H., Bijleveld, E., & Custers, R. (2010). Unconscious reward cues increase invested effort, but do not change speed-accuracy tradeoffs. *Cognition*, 115, 330-335.

Aarts, H., Bijleveld, E., Daunizeau, J., Veling, H., & Zedelius, C.M. (2012). Promising high monetary rewards for future task performance increases immediate task performance. *Plos One*, 7.

Alkadhi, H., Brugger, P., Boedermaker, S.H., Crelier, G., Curt, A., Hepp-Reymond, M. C., and Killias, S. S. (2005). What disconnection tells about motor imagery: evidence from paraplegic patients. Cereb. Cortex 15, 131-140.

Baddeley, A. (1996). Exploring the central executive. *Quarterly Journal of Experimental Psychology*, 49, 5±28.

Bayne, T., Levy, N. (2006) N. Sebanz and W. Prinz (eds.) *The Feeling of Doing: Deconstructing the*

Bicknell, K., Christensen, W., Mcllwain, D., & Sutton, J. (2015). The sense of agency and its role in strategic control for expert mountain bikers. *Psychology of Consciousness: Theory, Research and Practice*, 2, 340-353.

Boghossian, P. and Velleman, J.D. (1989). Color as a secondary quality. *Mind* 98: 81–103.

Botvinick, M., Carter, C. S., Braver, T. S., Barch, D. M., Cohen, J. D. (2001) Conflict monitoring and cognitive control. *Psychological Review*, *108*, 624–652.

Botvinick, M., Cohen, J. D., & Carter, C. (2004). Conflict Monitoring and anterior cingulate cortex: an update. *Trends in cognitive sciences*, *8*(12), 539–546.

Botvinick, M.M, Braver, T.S., Barch, D.M., Carter, C.S., Cohen, J.D. (2001). Conflict monitoring and cognitive control. *Psychol. Rev*, 108, pp. 624-652.

Bresler, C. & Laird, J.D. (1992). *The process of emotional feeling: A self-perception theory*. In M. Clark (Ed.) Emotion: review of personality and social psychology (223-234). Newbury Park, CA: Sage.

Bush, F.M., Francis, M., Harkins, S.W., Long, S., & Price, D.D. (1994). A comparison of pain measurement characteristics of mechanical visual analogue and simple numerical rating scales. *Pain*, 56, 217-226.

Carrasco, M., Ling, S., & Read, S., (2004). Attention alters appearance. *Nature Neuroscience*, 7(3): 308–31

Carter, C. S., van Veen, V. (2007). Anterior cingulated cortex and conflict detection: An update of theory and data. *Cognitive Affective and Behavioral Neuroscience*, *7*, 367–379.

Cavallaro, L.A., Flack, W.F., & Laird, J.D. (1999). Separate and combined effects of facial expressions and bodily postures on emotional feelings. *European Journal of Social Psychology*, 29, 203-217.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychological review*, *97*(3), 332.

Cramer, S. C., Lastra, L., Lacourse, M.G., and Cohen, M.J. (2005) Brain motor system function after chronic, complete spinal cord injury. *Brain* 128, 2941 -2950.

Critchley, H.D. (2003) Human cingulate cortex and autonomic cardiovascular control: converging neuroimaging and clinical evidence. Brain 216.2139-2152.

Csikszentmihályi, M. (1992). Flow: The Psychology of Happiness. Rider. ISBN 978-0-7126-5477-7.

Cutter, B. and Tye, M. (2011). Tracking Representationalism and the Painfulness of Pain. *Philosophical Issues,* 21, 90-109

Damasio, A. R. (1994). Descartes' Error: Emotion, Reason and the Human Brain. New York: Gossett/Putnam.

Dehaene, S., Chengeux, J.P., & Kerzberg, M. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences of the United States of America*, 95, 14529-14534.

Duclos, S.E., & Laird, J.D. (2001). The deliberate control of emotional experience through control of expressions. *Cognition and Emotion*, 15, 27-56.

Duclos, S.E., Laird, J.D., Schneider, E., Sexter, M., Stern, L., & Van Lighten, O. (1989). Emotion-specific effects of facial expressions and postures on emotional experience. *Journal of Personality and Social Psychology*, 57, 100-108.

Franconeri, S.L. et al. (2012). Flexible cognitive resources: competitive content maps for attention and memory. Trends in Cognitive Sciences

Hilbert, D., Klien, C., (2014). No Problem. ed. Brown, R. Consciousness Inside and Out: Phenomenology, Neuroscience, and the Nature of Experience, Studies in Brain and Mind. p. 229-306.

Hill, C. S., (2009). Consciousness. Nova York: Cambridge University Press.

Inzlicht, M., and Marcora, S. M. (2016). The central governor model of exercise regulation teaches us precious little about the nature of mental fatigue and self-control failure. *Front. Psychol*. 7:656.

Keillor, J., Barrett, A., Crucian, G., Kortenkamp, S., Heilman, K. (2002). Emotional experience and perception in the absence of facial feedback. *Journal of the International Neuropsychological Society*, 8(1), 130-135.

Kellerman, J.M., & Laird, J.D. (1982). The effect of appearance on self-perceptions. *Journal of Personality*, 50, 296-351.

Kurzban, R., Duckworth, A., Kable, J. W., Myers, J. (2013).An opportunity cost model of subjective

LeDoux, J.E., (1996). The Emotional Brain: The Mysterious Underpinnings of Emotional Life, New York: Simon and Schuster.

Maringer, M., & Niedenthal, P.M. (2009). Embodied emotion considered. *Emotion Review*, 1, 122-128.

Martin, & Stepper 1988. Perspect. Psychol. Sci. 11(6):917–28.

Maunsell, J. H., & Newsome, W. T. (1987). Visual processing in monkey extrastriate cortex. *Annual review of neuroscience*, *10*(1), 363–401.

Mendelovici, A. & Bourget, D. (2014). Naturalizing Intentionality: Tracking Theories Versus Phenomenal Intentionality Theories, *Philosophy Compass* 9(5): 325–337.

Milham, M.P., Banich, M.T, Webb, A., Barad, V., Cohen, N. J, Wszalek, A.F. Kramer (2001) The relative involvement of anterior cingulate and prefrontal cortex in attentional control depends on nature of conflict. Cognitive Brain Research 12. 467-473

Morsella, E., Wilson, L.E., Berger, C.C., Honhongva, M., Gazzaley, A., Bargh, J.A. (2009) Subjective Aspects of Cognitive Control at Different Stages of Processing. *Atten Percept Psychophys.* 71(8): 1807–1824

Naccache, L., Dehaene, S., Cohen, L., Habert, M. O., Guichart-Gomez, E., Galanaud, D., Willer, J. C. (2005). Effortless control: executive attention and conscious feeling of mental effort are dissociable. *Neuropsychologia*, *43*(9), 1318–1328.

Nakamura, J., & Csikszentmihályi, M. (2014). The concept of flow. "Handbook of positive psychology," 89-105. New York, NY: Oxford University Press.

Niedenthal, P.M. (2007). Embodying emotion. *Science*, 316, 1002-1005.

Pautz.A. (2013) The real trouble for phenomenal externalists. In *Consciousness Inside and Out: Phenomenology, Neuroscience, and the Nature of Experience*. ed. T. Brown. Springer. pp. 237-298.

Phenomenology of Agency in *Disorders of Volition*. Cambridge, MA: MIT Press, 49–68.

Prinz, J. J. 2005. Are emotions feelings? Journal of Consciousness Studies 12 (8–10): 9–25

Sàenz, M., Buraĉas, G. T., & Boynton, G. M. (2003). Global feature-based attention for motion and color. *Vision research*, *43*(6), 629–637.

Speaks, J. (2015). The Phenomenal and the Representational, Oxford University Press.

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, 54, 768–777.

Tye, M. (2006). The puzzle of true blue. *Analysis*, 66, 173-178.

Tye, M. (1995). *Ten problems of consciousness: a representational theory of the phenomenal mind.* Cambridge, Mass.: MIT Press.

Tye, M. (2008). The Experience of Emotion: An Intentionalist Theory. *unpublished.* URL = <https://webspace.utexas.edu/tyem/www/emotions.pdf>.

van Veen, V., Cohen, J.D., Botvinick, M. M., Stenger, V.A., Carter, C.S. (2001). Anterior cingulate cortex, conflict monitoring, and levels of processing. *Neuroimage*. 14(6):1302-8.

Wagenmakers EJ, Beek T, Dijkhoff L, Gronau QF, Acosta A, et al. 2016. Registered replication report: Strack,

Westbrook, A., Braver, T. S. (2015). Cognitive effort: a neuroeconomic approach. Cogn. Affect. Behav. Neurosci. 15:395—415.

Chapter 3: Emotion's Role in the Unity of Consciousness

3.1 Introduction

The question of what it is for consciousness to be *unified* has been at the forefront of philosophical debates before the advent of empirical perspectives (James, 1890; Kant, 1781). A recently revived debate on the unity of consciousness in the split-brain has given new life to this investigation (Schechter, 2018; Pinto et. al., 2017; Corballis et. al., 2018). In subjects with the split-brain syndrome, some conscious states appear to be *disunified* (e.g., visual states), while others remain *unified* (e.g., affective states). While placing emphasis on the disunities, disunity accounts conclude that split-brain subjects have two subjective perspectives and not one (Schechter, 2018; Gazzaniga & LeDoux, 1978). That is, there is *something that it's like* to be the left hemisphere, and *something that it's like* to be the right hemisphere, but *nothing that it's like* to be the entire subject of experience.

In this work I argue that affective unity is more important than perceptual disunity in delineating our subjective perspective. Hence, split-brain patients retain a unified subjective perspective on the world. Unlike enjoying an *objective* perspective, enjoying a *subjective* perspective entails experiencing aspects of your phenomenal field in terms of their overall relation to *you*. That is, what it's like to be you at any given time entails experiencing certain aspects of the phenomenal field as peripheral to others. Since emotion plays a greater role than perception in creating and sustaining this periphery/center structure, affective unity is more important than perceptual disunity in delineating *what it's like to be you* at any given time.

Emotion creates and sustains a center/periphery structure in our phenomenal field by signaling orders of *perceived importance* as well as orders of *perceived changeability* (e.g., to someone feeling grief, the experience of getting a parking ticket is felt as *peripheral/unimportant* to the experience of grief; to someone feeling depressed, the experience of reaching a hilltop is felt as *peripheral/unattainable* to the experience of laying down to rest) (Oatley and Jenkins, 1996; Schnall et. al., 2008).

While a lack of interhemispheric integration in the split-brain might be damaging to the subjective perspective, it is not enough to split the subjective perspective into two. The resultant verdict helps explain some of the main objections to disunity models, i.e. the relative lack of impairments suffered by split-brain subjects in everyday life, as well as the eventual dissipation of apparent disunity under experimental conditions.[i]

3.2 The Split-Brain Syndrome and the Duality Account

The split-brain syndrome results from a procedure severing the corpus collosum between the two hemispheres in varying degrees.[ii] While the surgery fulfills its goal in reducing seizures, it has subtle effects on the patient's consciousness. These effects are subtle in that everyday life seems largely unaffected.

The effects of the surgery are tested using an experimental set up designed to prevent the two hemispheres from interacting in an *indirect* way.[iii] Roughly, an indirect exchange of information is a way in which two distinct minds might interact. Indirect interaction between mental states would entail something like tracing out the information presented to one hemisphere with one's hand or using perceptual cues to get the information across to the other hemisphere. Since we are interested in the way the subject's (putatively single) mind might process information, the design includes perceptual lateralization, response control, and prevention of cross-cuing. Consider the following description of a typical split-brain experiment:

> Two stimuli are presented to the patient in such a way that one will be processed by the left hemisphere and the other by the right hemisphere. For example, the word 'keyring' might be projected such that 'key' is restricted to the patient's left visual field (LVF), and 'ring' is restricted to the patient's right visual field (RVF). The contralateral structure of the visual system ensures that stimuli projected to the LVF are processed in the right hemisphere and vice-versa. Other perceptual systems can be studied in a similar manner. For example, tactile perception is examined by asking the patient to compare an object presented to the right hand with one presented to the left (Bayne, 2008, p.278).

This experimental set up allegedly reveals two types of disunities in split-brain patients, i.e. behavioral disunities and representational disunities (Bayne, 2008). When carefully prompted (as to prevent cross-cuing), split-brain patients are behaviorally disunified in that a patient will only verbally report what was presented to the left hemisphere, e.g., the word 'ring'; but will use her left hand to pick out what was presented to the right hemisphere, e.g., the picture of a 'key'.[iv] Moreover, patients appear to be representationally disunified in that the representations of 'key' and 'ring' are never combined into a unified representation of 'key-ring'. While experimental findings suggest that both representations are individually conscious, they are nonetheless not *co-conscious*.

Importantly for our purposes, theorists have been careful to note that emotional states remain unified even under experimental conditions. Sperry describes this phenomenon in the following way:

> Unlike other aspects of cognitive function, emotions have never been readily confinable to one hemisphere…the emotional effects spread

> rapidly to involve both hemispheres, apparently through crossed fiber systems in the undivided brain stem (Sperry, 1982, p.1225).

For example, if the right hemisphere is shown a disturbing film from the subject's left visual field, the entire subject experiences fear. One participant describes her experience after viewing the film in the following way: "I don't really know why I'm kind of scared. I feel jumpy…." (Gazzaniga, 1985, p. 76).

Consider a similar experiment. Geometric shapes were shown to either the right hemisphere (on the left visual field) or the left hemisphere (on the right visual field).

> One young male patient was tested with the hemi-field tachistoscopic technique which allows visual presentation of stimuli to only one visual half-field at a time. With his gaze focused on a central fixation point on the screen, different simple configurations were flashed quickly one at a time either to the left or to the right side of the point and the task was to name them. He had no difficulty in naming those flashed on the right side of the fixation point (the information was transmitted to the left, speaking hemisphere) but he was unable to name simple geometrical configurations on the left (information was transmitted only to the right, mute hemisphere). He attempted guesses or simply said he was unable to name the image. "One of the configurations that was projected on the left was that of a swastika. Unlike any of the previous reactions in earlier trials, he immediately sat back in his chair exclaiming. 'What was this that you just showed me!'. What do you think it was, asked the experimenter. He replied, 'A terrible thing, an awful thing.' You did not like it, stated the experimenter. 'No, I didn't,' he replied, shaking his head. Was it a good thing or a bad one, probed the experimenter (who did not anticipate strong reactions to any of the items in the set) 'Bad, very bad,' replied the patient. He was never able to name it…" (Zaidel, 1994: 171).

While only the right hemisphere was shown the swastika; the left hemisphere reported the affective phenomenology, suggesting that affect was likely experienced by the *entire subject.* In the following sections, I propose to take a closer look at emotion's role in the unity of consciousness.

3.3 What is a Subjective Perspective?

Before we analyze experimental findings in more detail, it would be useful to get a handle of the notion of a subjective perspective. To have a subjective perspective on the world is for there to be *something that it's like* to be that being (Nagel, 1974). To review, according to the disunity accounts, there is something that it’s like to be the left hemisphere, and something that it’s like to be the right hemisphere, but nothing that it’s like to be the entire subject of experience.

How does a subjective perspective differ from an objective perspective? Using a visual analogy, a perspective is a *point of view* that establishes spatial relations in terms of their overall relation to the perceiver.[v] That is, every visual perspective has a structure where some parts of the scene are experienced as *peripheral* to others. For example, in focusing on the computer screen in front of me, I automatically “zoom out” from the photograph on the right side of my desk. Watzl describes this phenomenon in the following way: “Attending to something consists in *structuring* the stream of consciousness into center and periphery” (Watzl, 2010; p.7).

A similar line of thought could be applied to our *overall* subjective perspective on the world. Like visual attention, emotion plays a structuring role in our overall point of view. It does so by purporting to signal *relations of relative importance* from our point of view. For example, as you’re suddenly hit with feelings of grief, you experience the rest of the world *in relation* to your grief. Other everyday concerns like getting a parking ticket fade into the periphery of your grief. In other words, undergoing an emotional experience *creates* and *sustains* central/peripheral relations between aspects of your experience (e.g., Oatley and Jenkins, 1996). Emotion’s structuring role on our point of view also shows up in colloquial expressions. For example, an emotional subject with a momentarily narrow perspective on the world could be described as "blind with rage" or "blinded by love".

Notably, since emotions could be misleading, they only *purport* to signal the relative importance of things from our point of view. For example, someone losing their cool on the road might momentarily feel like a victim of an important injustice, only to realize his overreaction a second later.

In short, our point of view on the world is inherently bound up with our emotional experiences. Emotion *structures* our phenomenal field by bringing certain elements into the center and others into the periphery.

One might object that emotion is not the only type of phenomenology with a power to structure our point of view. After all, vision also creates and sustains phenomenal relations between items in your phenomenal field. Going back to the example above, focusing on the computer screen in front of me pushes the rest of the room into the periphery. Since experimental findings suggest that visual states are indeed disunified in the split brain, why should we prioritize emotional unity over visual disunity?

To understand why emotion plays a greater role than vision in structuring our subjective perspective, consider the following example. Say you're visually attending to the television screen in front of you that takes up the majority of your visual field. Suddenly, from the very corner of your eye, you spot your partner's terrified face. Even without moving your eyes, your entire subjective perspective on the world shifts. Despite the fact that the television screen still takes up the majority of your *visual* perspective, *what it's like to be you* is largely determined by your affective experience. In other words, even though your visual attention is fixed on the screen, your affective experience causes you to experience the screen *as peripheral* to the terrified face (Ohman et. al., 2001a). The TV screen is no longer important since your partner might be facing a dangerous threat. In short, while vision has the capacity to structure your *visual* perspective, your *overall* subjective perspective on the world is largely determined by affect.

A further objection could be spelled out in the following way. If one is accustomed to the misleading nature of one's emotions, one's point of view may no longer be structured by these emotions. For example, someone struggling with depression might no longer experience these feelings from the first-person point of view. Instead, one could learn to *distance* oneself from these feelings and gain a certain level of *dissociation.* For instance, the onset of these emotions could trigger the following feelings in a weary subject: "I am so tired of feelings this way… this is not *really me!*" or "Here I go again focusing on my failings, but I know from experience that things aren't *really* as bad as they *seem*!" In other words, even if depressive emotions push the experience of one's failures into the center and achievements into the periphery, one could still occupy a point of view *outside* these experiences. Hence, if one experiences these emotions from the third person point of view, one's subjective perspective would no longer be structured by these emotions.

Several replies are available. First, it is notable that *feeling* dissociated is still a type of emotional experience. In this example, the subject struggles with depression and over time has developed certain feelings towards his illness. Having grown "sick and tired" of feeling a certain way, he has developed negative emotions about his emotional tendencies. Hence, it is still the case that emotions structure his point of view, even if these emotions have other emotions as their objects.

Secondly, the very notion of "struggling" implies that his depressive feelings still very much structure his subjective perspective on the world. Unable to just wish them away, their onset still pushes certain elements into the foreground and others in the periphery. Hence, it is still the case that emotions structure his point of view, even if he wishes that weren't the case.

Lastly, feeling dissociated from one's emotional experiences is not the norm for most people. Rather than constantly double-checking whether they're being led astray, non-clinical populations experience emotions more or less full-heartedly. Hence, it is still the case that emotions structure one's point of view, even if these emotions are more complex than previously suggested.

3.3.1 Subjective Perspective and Feelings of Agency

So far, we have seen that emotion structures our point of view by bringing seemingly important elements into the center and other seemingly unimportant elements into the periphery. In this section I further this argument by outlining affect's role in feelings of agency. In particular, in addition to structuring our phenomenal field by *orders of perceived importance*, emotion also structures our phenomenal field by *orders of perceived changeability* (e.g., Gibson, 1979; Noe 2004; O'Regan, 2011). For example, a depressed traveler might perceive a tree at the top of a steep hill as unreachable, while an excited child might perceive it as just "a hop away" (Schnall et. al., 2008). In other words, the depressed traveler's emotions almost "push" the tree further into the periphery by determining his *felt inability* to reach it. This line of thought is confirmed by empirical findings. Specifically, according to the *affective coding* hypothesis, affect modulates feelings of (i) prospective agency, (ii) real-time agency, and (iii) post-hoc judgements of agency (Gentsch and Synofzik, 2015).[vi]

Hence, if split-brain patients remain *affectively unified* under experimental conditions, they also remain *agentially unified.* Recent findings from the split-brain paradigm lend support to this hypothesis. In their recent work “Split-Brain: Divided Perception but Undivided Consciousness”, Pinto et. al. show that while visual perception could be split and processed independently by each hemisphere, consciousness is nonetheless unified. The authors summarize their findings in the following way:

> The canonical idea of split-brain patients is that they cannot compare stimuli across visual half-fields (left), because visual processing is not integrated across hemispheres. This is what we found as well. However, another key element of the traditional view is that split-brain patients can only respond accurately to stimuli in the left visual field with their left hand and to stimuli in the right visual field with their right hand and verbally. This is not what we found. Across a wide variety of tasks, we observed that split-brain patients could reliably indicate presence, location, orientation and identity of stimuli throughout the entire visual field regardless of how they responded *(ibid,* p.1232).

It appears that contra previous findings of behavioral disunity, split-brain patients are indeed agentially unified in that they’re able to respond accurately to stimuli presented *anywhere* in the visual field, using *any* response modality (e.g., both left and right hands as well as verbal report). These findings of agential unity in the split-brain lend support to our hypothesis that affective unity structures our point of view by determining our *felt ability* to act on the world. Since their experience of agency is already determined by the shared affect, split-brain subjects under experimental conditions do not experience themselves to be *passive* with respect to actions that might have originated in one opposite hemisphere.

Someone might object to this notion of structuring in the following way. According to this proposal, emotion doesn’t only structure elements in the phenomenal field by orders of perceived importance, but also by orders of perceived changeability. For the example, a depressed traveler might feel the hilltop to be unsurmountable and thus experience it *as peripheral* to the nice comfy bench right next to him. However, it is not always the case that things that are experienced as unchangeable or unattainable occupy the periphery of our consciousness. In fact, someone feeling depressed might become excessively preoccupied with things she finds unattainable (e.g., having a stress-free workday). In other words, experiencing things as unattainable might actually push them into the center and not the periphery of one's phenomenal field.

This example features an interesting interplay between the two notions of structuring (i.e., structuring in order of perceived importance and structuring in order of perceived changeability). In particular, it appears that one's phenomenal field could feature two layers. In this case, the first layer is structured by orders of perceived changeability. In other words, one's depression has moved imagined experiences like having a stress-free workday into the periphery of one's phenomenal field. Given one's illness, having an "episode free" workday just seems unattainable and really "far away". The second layer is structured by one's affective attitude towards this attainability. In particular, one's affective attitude or felt preoccupation towards this attainability has pushed it into the center of one's consciousness.

3.4 Objection from Cortical Integration

So far, I have argued for the following conclusion. Emotion structures our subjective perspective by pushing certain elements to the center and others to the periphery. Furthermore, affect has the power to structure our subjective perspective by determining our felt ability to make changes to these elements. In order to develop this view further, I propose to turn our attention to several objections from the disunity account.

According to the integration model of consciousness, integration is necessary for a unified subjective perspective on the world. Since split-brain subjects suffer a lack of cortical integration due to a severed corpus collosum, split-brain subjects do not enjoy a unified subjective perspective on the world.

In particular, Schechter's duality account makes use of the *global workspace* model of consciousness (e.g., Dahaene et. al., 2011). According to this model, what makes a mental state conscious is that its content becomes available to multiple reasoning systems simultaneously.

Speaking somewhat metaphorically, we can say that two centers of consciousness were formed the moment the right hemisphere became conscious of the percept "ring", and the left hemisphere become conscious of the percept "key". That is, while each bit was integrated *within* each hemisphere (e.g., two separate percepts gave rise to two separate beliefs), the absence of a corpus collosum prevented the entire content "key-ring" from being integrated interhemispherically.

Nevertheless, it is not the case that cortical integration houses the only attentional mechanisms in the brain. In fact, several lines of evidence suggest that low-level affect plays an *attention-like* role in organizing sensory, cognitive, and agentive elements into a seemingly coherent whole (Pourtois et al., 2012; Vuilleumier and Schwartz, 2001b; Fox, 2002; Grabowska et al., 2011).

The first line of evidence comes from Pourtois' *Multiple Attention Gain Control* model. According to this model, sub-cortically realized affective states compete with attentional mechanisms realized by cortical integration (Pourtois et. al, 2012; Gentsch and Synofzik, 2015; Corballis et. al., 2018). This model details systematic dissociations between cortically realized attentional mechanisms and sub-cortically realized affective mechanisms in organizing conscious states (Pourtois et al., 2012). Since the two depend on *non-overlapping* circuits, attentional and affective organizing effects could be pitted against one another in experimental conditions (Corbetta and Shulman, 2002). According to this model, "emotion signals can shape perception by mechanisms that do not overlap with other (e.g., endogenous or voluntary) attentional processes" (ibid, p. 505). These affective mechanisms operate *very similarly* to low-level organizing principles like attention and structure perceptual inputs into recognizable configurations (Vuilleumier et. al., 2001; McMains and Kastner, 2011).

Furthermore, given that these effects are impenetrable to manipulations of attentional control mechanisms (e.g., either endogenous or exogenous selective attention), these affective mechanisms could operate *independently* of attentional mechanisms. Consider the following summary of the dissociation data:

> Patients with neglect or visual extinction suffer from selective damage to fronto-parietal networks controlling spatial (endogenous and/or exogenous) attention and show severe deficits in orienting their attention towards the contralesional side of space, but emotional biases in spatial orienting may still occur despite the overall neglect biases (Vuilleumier and Schwartz, 2001b; Fox, 2002; Grabowska et al., 2011).

Additional evidence for the independence of affective and cortical attentional mechanisms could be found in neuroimaging studies on hemispatial neglect following parietal damage (Vuilleumier et al., 2002; Grabowska et al., 2011). Patients with hemispatial neglect fail to orient to stimuli in the (usually left) space opposite to their (usually right) brain lesion, due to a destruction of brain networks controlling spatial attention towards that side (Driver and Vuilleumier, 2001). However, even while exogenous and endogenous attentional mechanisms are impaired, emotional stimuli *still* serve *attention-like* role in orienting subjects to stimuli in the space opposite to their brain lesion (Vuilleumier and Schwartz, 2001a; Fox, 2002; Lucas and Vuilleumier, 2008; Grandjean et al., 2008; Grabowska et al., 2011).

Furthermore, it appears that sub-cortically realized affective mechanisms arrive on the scene *prior to* attentional mechanisms of cortical integration (Pourtois et al., 2010b; Luo et al., 2010; Brosch et al., 2011; Ciesielski et al., 2010). Consider the following summary of the two key claims of the model (my emphasis):

> (i) The time-course of emotional effects suggests a distinctive spatio-temporal dynamic as compared with other attentional modulations (in fronto-parietal areas), with relatively early responses observed in some limbic regions, such as the amygdala or orbitofrontal cortex (Kawasaki et al., 2001), which might then act to *gate* sensory processing in distant regions at later latencies.
>
> (ii) These emotional attention effects may occur in parallel to other gating effects mediated by fronto-parietal attention networks (Amaral et al., 2003; Krolak-Salmon et al., 2001; Vuilleumier and Pourtois, 2007; Pourtois et al. 2010a, [b] 2010) and thus be partly independent of (or even competing with) any concomitant modulation by the latter systems (ibid, p.7).

Further evidence of affect's *attention-like* role in organizing our conscious experience comes from findings on inter-modal integration. In particular, it is commonly assumed that inter-modal integration is a perceptual phenomenon that takes place more or less automatically (Alais and Burr, 2004). For example, the "ventriloquist effect" is believed to be an automatic, cross-modal phenomenon. In experiencing this effect, we attach the perceived location of an auditory stimulus to a concurrently presented visual stimulus. The magnitude of the effect is supposed to reflect the extent of this audiovisual binding.

However, Maiworm et. al., (2012) demonstrated emotion's orchestrating role in this multi-sensory integration. The authors showed a reduction of the ventriloquist effect following emotion stimulus manipulation (ibid, p.102). The control conditions revealed that it is unlikely that the effect resulted in merely enhanced processing of the auditory information. Instead, the affective stimulus present in the auditory modality modulated the multimodal integration in favor of audition, reversing the previously assumed automaticity of visual dominance.

In sum, it appears that affect plays an *attention-like* role in directing how each modality interacts with another or if a given modality gets to contribute at all. Hence, since *affective* attentional mechanisms are indeed preserved in the split-brain, a lack of interhemispheric integration is not enough to split the subjective perspective into two.

3.5 The Compatibility Objection

While Schechter grants that some affective states might be shared between the two hemispheres, she argues that shared affect is *compatible* with the disunity account. This objection could be broken down into two parts: (i) argument from architectural assumptions of the mind, and (ii) argument from the type/token distinction. According to the former, "all the shared affect in the world cannot dissolve the boundaries of individual psychologies" (ibid, p.110). Schechter formalizes these architectural assumptions in the following way:

> At a minimum, the objection to the duality account should show that R's mental activities interact with L's substantially directly, rather than mainly in a way that multiple minds characteristically interact, that is via paired re/action and sensation/perception. Ideally, the objection would also show that R and L do not separately but together meet the architectural assumptions. It would show that R and L's activities interact in a way we would expect the mental states of a single mind to interact – for instance, with LH percepts leading to the formation of RH perceptual beliefs (ibid, p.110).

In sum, even if affect remains unified in the split-brain, lack of direct formation of perceptual beliefs from the opposite hemisphere's perceptions violates basic "architectural assumptions" of the mind.

The second strand of this objection makes use of the type/token distinction to neutralize the significance of shared affect in the split-brain. In particular, Schechter argues that while the affective state *type* is shared, each hemisphere creates its own conscious experience token by providing its own "readout" of the shared state.[vii] In other words, the general affective type like negative mood does not belong to *any* subjective perspective prior to becoming available for "broadcast" within each hemisphere.

Several replies are available. First, the architectural characterization makes use of the assumption that activities within a single mind are organized via attentional mechanisms realized by cortical integration. However, we've seen some reasons to doubt that cortical integration is the *only* means for a single mind to ensure internal interaction. In fact, there is *already* some integration of information going on via sub-cortically realized affective mechanisms, and a lack of cortical interhemispheric integration is not enough to split this unified perspective into two.

Along the same lines, it is difficult to understand why two independent subjective perspectives would be created when there is *already* a perfectly good subjective perspective on the scene. Specifically, the *independence* and *unique timescale* of affective and attentional mechanisms detailed in the previous section suggests that cognitive interpretations of affective phenomenology may proceed *independently* of these low-level organizing effects and create *additional* mental states without *canceling out* the initial phenomenal feels in the split-brain. Thus, it is not the case that shared affect does not belong to *any* subjective perspective prior to intra-hemispheric cortical integration, i.e. that there is *nothing that it's like* to be the entire subject of experience of these low-level affective states, even if the later interpretations by each hemisphere may create *additional* mental states.

Furthermore, contra the type/token characterization, even if the later interpretations provided by each hemisphere are capable of creating additional conscious states, it does not mean that these two new states would now belong to two independent subjective perspectives. In particular, we have no reason to suspect that the right hemisphere's new elaborate affective state would interact with the left hemisphere's new elaborate affective state "in a way we would expect *two distinct* minds to interact" (ibid, p. 110). Instead, we would see that these two new states would "interact in a way we would expect the mental states of a *single mind* to interact – for instance, with LH's *emotions* leading to the formation of RH's *emotional* beliefs" (ibid, p.110).[viii]

For instance, say that both hemispheres experience a general shared negative affect. However, while the left hemisphere is primed with pictures of financial loss, the right hemisphere is primed with pictures of health troubles. So while the left hemisphere thinks that it feels terrible because of financial loss, the right hemisphere thinks that it feels terrible because of health troubles. Nevertheless, if the right hemisphere were asked whether it would like to donate a large sum to charity, it would likely decline; and if the left hemisphere were asked whether it would like to decrease the deductible on health insurance, it would likely agree.

Hence, the emotional state of the right hemisphere would indeed give rise to (emotional) beliefs and decisions of the left hemisphere and the emotional state of the left hemisphere would indeed give rise to (emotional) beliefs and decisions of the right hemisphere. Thus, by Schechter's own architectural standards, each hemisphere's conscious states would interact with the other hemisphere's conscious states in a way we would expect two states of a single, *albeit fuzzy*, mind to interact.

Final reasons to doubt that the left hemisphere's and the right hemisphere's two elaborate affective states would belong to *two* independent subjective perspectives comes from our pre-theoretical intuitions about the structure of the subjective perspective. Since your affective experiences largely determine what *it's like* to be you at any point in time, further cognitive interpretations will not create *two new* subjective perspectives, but merely change details in the point of view already anchored by affect.

3.5.1 Objection from Visual Representational Disunity

So far, we have seen some reasons to believe that split-brain patients retain a unified subjective perspective on the world. If so, how do we make sense of the findings indicating a lack of representational integration between the visual representations “key” and “ring” under experimental conditions? In other words, if affect has the power to organize one’s subjective perspective into a seemingly coherent whole, why doesn’t it “pull” visual representations of “key” and “ring” into a unified whole?

To understand why visual representational disunity is *compatible* with a unified subjective perspective, consider the distinction between *phenomenal* and *access* consciousness. In particular, a lack of visual representational disunity only shows a lack of unified *access* consciousness and not a lack of unified *phenomenal* consciousness (Bayne and Chalmers, 2003). Furthermore, not only does phenomenal unity survive access disunities in the split-brain, it might benefit at their expense (Jolij & Lamme, 2005).

Many philosophers distinguish between access consciousness and phenomenal consciousness (Block, 1995; Bayne and Chalmers, 2003). Roughly, a state is access conscious when its content is available for verbal report and rational inference. On the other hand, a state is phenomenally conscious when there is something that it’s like to be in that state, but its content is not available for verbal report, reasoning systems, and rational behavioral control. According to Block, states could be phenomenally conscious without being access conscious and vice versa. For example, someone intensely focused on his conversation might be *phenomenally* conscious of the background noise in his surroundings, but only become *access* conscious of the noise once the conversation dies down (Block, 1995).

Let's take a look at Bayne and Chalmers' proposal in more detail. Even in non-clinical subjects, access to experiences could be limited by limitations in attentional resources. Consider the following example of this limitation:

> Perhaps the clearest example of such a bottleneck is given by a famous experiment by George Sperling (1960). In Sperling's experiment, a subject is presented with a matrix consisting of three rows with four letters each. The matrix is flashed only briefly, for 250 milliseconds. After the matrix vanishes, a tone sounds, indicating whether the subject is to report the contents of the first, second, or third row. When subjects are required to report the contents of the top row, on average they correctly report 3.3 of the four letters in that row. The same goes when they are required to report the contents of the middle row, or of the bottom row. But when subjects are asked to report the contents of the entire matrix, on average they correctly report 4.5 of the twelve letters. So, to simplify a little, it seems that the subject has access to the information in any single row, but the subject does not have joint access to the information in all three rows (ibid, p.15).

This distinction could carry itself to different types of *co-consciousness* in the split-brain. In particular, co-consciousness could amount to the following relations: co-accessibility and co-phenomenality (e.g., Schechter, 2018; Bayne and Chalmers, 2003; Bayne, 2010). Two states are *access-unified* if the conjunction of their contents is available for verbal report, reasoning, and rational behavioral control. For example, if mental state G has content P and mental state R has content Q, these states will be individually access-conscious if the information 'that P' is available for report and rational control, and if the information 'that Q' is available for report and rational control. They will be *access-unified*, if the information "that P&Q" is available for report and rational control (Bayne and Chalmers, 2003, p.8). On the other hand, if two experiences are co-phenomenal, there is something that it's like for the subject to undergo them simultaneously as opposed to on separate occasions. For example, there is something that it's like to hear a guitar and smell a rose together at t1, as opposed to hearing a guitar at t1 and smelling a rose at t2. According to Chalmers and Bayne, what it means for a subject's phenomenal consciousness to be unified is for all the relevant phenomenal states to be *subsumed* under a total phenomenal state or phenomenal field. This total state amounts to something that it's like to be the subject of experience at t1 or the subject's subjective perspective. Specifically, experiences are phenomenally unified just in case they stand in this subsumption relation to the total phenomenal state. States A and B are unified just in case there is one total state (M) that subsumes A and B. For example, the total phenomenal state M (hearing a guitar and smelling a rose together) subsumes the phenomenal states of A (smelling a rose) and B (hearing a guitar).

Notably, while my definition of a unified phenomenal perspective slightly diverges from the one offered by Bayne and Chalmers, the main distinction between *access* and *phenomenal* consciousness still applies.[ix] The standard 'key-ring' experiment described in the beginning of this paper seems to provide prima facie evidence that split-brain subjects do not enjoy unified *access* consciousness. It appears that one hemisphere is *access conscious* of the word 'key', while the other is *access conscious* of the word 'ring'; but no one is *access conscious* of the unified word 'key-ring'. The disunity accounts conclude that perhaps there is someone (the right hemisphere) undergoing what it's like to see 'key', and someone (the left hemisphere) undergoing what it's like to see 'ring', but no one is undergoing the unified perception 'key-ring'. Hence, it appears there are really two subjective perspectives and not one.

However, Bayne and Chalmers point out that simply because the joint content "key-ring" is not available for report, it does not show that the subject does not enjoy one total phenomenal state that subsumes the phenomenology of "key" and "ring". The authors conclude by emphasizing that "there is nothing paradoxical or contradictory about this…this is just what we might expect" (*ibid*, p.16).

Nevertheless, Schechter is not impressed with this treatment of the split-brain data. According to Schechter, not only does the appeal to unified phenomenal consciousness seem incomplete, it also comes at a cost to intelligibility. Schechter emphasizes the role of access conscious in our introspective experience (my emphasis):

> It is not immediately clear what it would mean for consciousness to be unified while *conscious perception* was split, since *so much of conscious experience* (or perhaps *all* of it; Carruthers, 2011) *is perceptual in nature* (ibid, p.40).

In the previous sections we've seen some reasons to doubt that *all* consciousness is perceptual in nature. In fact, some conscious states, in particular *affective phenomenal* states do not appear to be perceptual in nature.

Why do phenomenal unity and access unity come apart? Philosophers theorize that access could be limited by limitations in attentional resources (Bayne and Chalmers, 2003; Block; 2005). There are simply not enough attentional resources to ensure that all conscious states get broadcasted to all reasoning systems. However, the existence of bottleneck does not explain the exact relationship between access and phenomenal consciousness. The bottleneck explanation seems to suggest that allocation of states into access and phenomenal consciousness is a mere *side-effect* of the attentional bottleneck. In other words, there is nothing about access disunity *itself* that triggers phenomenal unity. Furthermore, affect is *just another phenomenal* state like visual phenomenology that could either make it into access consciousness or not.

I suggest that affective states play a special role in this dissociation. Moreover, there is something about access disunity itself that triggers phenomenal unity. Studies on affective blindsight show that it is only when affectively salient stimuli are shown below the threshold associated with cortical integration and access consciousness, that subjects experience an intense phenomenal state (e.g. Killgore and Yurgelun-Todd, 2004). In particular, Jolij and Lamme (2005) show that blocking the cortical route associated with attentional integration and access consciousness (interfering with V1 activity) via TMS greatly activates the subcortical pathway to the amygdala (likely via the midbrain and the thalamus), thereby blocking perception but *facilitating* affective discrimination (e.g. Killgore & Yurgelun-Todd, 2004).[x] However, if the affective stimulus is shown just above the threshold associated with access consciousness and cortical integration, (i.e. when the stimuli are clearly presented), intense affective phenomenology and affective discrimination are absent.

Surprisingly, there appears to be a *causal* relationship between the quality of *access* consciousness and the quality of *phenomenal* consciousness. Subjects report a greater affective phenomenology following poor quality of access consciousness.

The causal relationship between access disunity and phenomenal unity makes sense in evolutionary terms. That is, when affectively tagged information is not fully available to reasoning systems, the brain reacts by upping the feeling of unified subjectivity and agential unity, marshalling greater effort into securing a unified front for facing potential threats. Since the sub-cortical system is evolutionarily prior, it is likely more apt for detecting coarse changes in the environment, and coarse changes are likely affectively tagged (Corballis et. al., 2018).

In sum, it is unlikely that a lack of unified access consciousness would result in two subjective perspectives and not one. Since perception takes a backseat to emotion in structuring our point of view, our subjective perspective can survive perceptual disunities.

3.6 Conclusions

In this work I have offered several reasons to suspect that split-brain patients enjoy a unified subjective perspective on the world. In particular, since affective unity is more important than perceptual disunity in delineating our subjective perspective, split-brain subjects remain phenomenally unified and only suffer disunities in access consciousness.

Affective unity is more important than perceptual disunity since emotion plays a structuring role in our subjective perspective. Emotion structures our subjective perspective by bringing certain elements into the center and others into the periphery. It does so by signaling orders of perceived importance and perceived changeability.

Arguably, perceived importance and changeability these are the most important aspects of a subjective point of view or *what it's like to be you* at any given time. In other words, if someone is tasked with figuring out what it's like to experience the world from X's point of view, they would do well by first ascertaining what X finds important and what X finds changeable.

Hence, as long as the structur*ing* states remain unified, our subjective perspective can survive disunities in the details. While a lack of interhemispheric integration in the split-brain might be damaging to the subjective perspective, it is nonetheless not enough to split the subjective perspective into two. The resultant verdict helps explain some of the main objections to disunity models, i.e., the relative lack of impairments suffered by split-brain subjects in everyday life, as well as the eventual dissipation of apparent disunity under experimental conditions.

References

Alais, David, and David Burr. 2004. “The Ventriloquist Effect Results from near-Optimal Bimodal Integration.” Current Biology: CB 14 (3): 257–62.

Amaral, D.G., Behniea, H., Kelly, J.L., 2003. Topographic organization of projections from the amygdala to the visual cortex in the macaque monkey. Neuroscience 118 (4), 1099–1120.

Anderson, A.K., Phelps, E.A., 2001. Lesions of the human amygdala impair enhanced perception of emotionally salient events. Nature 411 (6835), 305–309

Bayne, T. (2008). The unity of consciousness and the split-brain syndrome. The Journal of Philosophy, 105, 277–300.

Bayne, T. (2010). The unity of consciousness. Oxford: Oxford University Press

Bayne, T., & Chalmers, D. (2003). What is the unity of consciousness? In A. Cleeremans (Ed.), The unity of consciousness: Binding, integration and dissociation (pp. 23–58). Oxford: Oxford University Press.

Beckers, Tom, Jan De Houwer, and Paul Eelen. (2002). “Automatic Integration of Non-Perceptual Action Effect Features: The Case of the Associative Affective Simon Effect.” Psychological Research 66 (3): 166–73.

Block, N. 1995. On a confusion about the function of consciousness. Behavioral and Brain Sciences.

Brosch, M., Selezneva, E. and Scheich, H. (2011). “Representation of Reward Feedback in Primate Auditory Cortex.” Frontiers in Systems Neuroscience 5 (February.

Carruthers, P., 2011. The opacity of mind, Oxford: Oxford University Press.

Corballis, M. C., Corballis, P. M., Berlucchi, G., and Marzi, C. A. (2018). Perceptual unity in the split brain: the role of subcortical connections. Brain 141: e46.

Corbetta, M., Shulman, G.L., 2002. Control of goal-directed and stimulus-driven attention in the brain. Nat. Rev. Neurosci. 3 (3), 201–215.

Dehaene S, Changeux JP. 2011. Experimental and theoretical approaches to conscious processing. Neuron 70:200227.

Driver, J., Vuilleumier, P., 2001. Perceptual awareness and its loss in unilateral neglect and extinction. Cognition 79 (1–2), 39–88.

Droit-Volet, S., Meck, W.H., 2007. How emotions colour our perception of time. Trends Cogn. Sci. 11 (12), 504513.

Eder, Andreas B., and Karl Christoph Klauer. (2007). "Common Valence Coding in Action and Evaluation: Affective Blindness towards Response-Compatible Stimuli." Cognition and Emotion 21 (6). Routledge: 1297–1322.

Eder, Andreas B., Jochen Müsseler, and Bernhard Hommel. (2012). "The Structure of Affective Action Representations: Temporal Binding of Affective Response Codes." Psychological Research 76 (1): 111 18.

Elsner, B., and B. Hommel. (2001). "Effect Anticipation and Action Control." Journal of Experimental Psychology. Human Perception and Performance 27 (1): 229–40.

Ekman, P., Davidson, J. (1994) The Nature of Emotion, Oxford University Press

Ellenberg, L., & Sperry, R. (1979). Capacity for holding sustained attention following callosotomy. Cortex, 15, 421–438.

Ellenberg, L., & Sperry, R. (1980). Lateralized division of attention in the commissurotomized and intact brain. Neuropsychologia, 18, 411–418.

Fox, E., 2002. Processing emotional facial expressions: the role of anxiety and awareness. Cogn. Affect. Behav.
Neurosci. 2 (1), 52–63.

Gazzaniga, M., & LeDoux, J. (1978). The integrated mind. New York: Plenum Press.

Gazzaniga, M. (2000). Cerebral specialization and interhemispheric communication: does the corpus callosum enable the human condition? Brain, 123, 1293–1326.

Gentsch, A., & Synofzik, M. (2014). Affective coding: the emotional dimension of agency. Frontiers in Human Neuroscience, 8, 608.

Grabowska, A., Marchewka, A., Seniow, J., Polanowska, K., Jednorog, K., Krolicki, L., et al., 2011. Emotionally negative stimuli can overcome attentional deficits in patients with visuo-spatial hemineglect. Neuropsychologia 49, 3327–3337.

Grandjean, D., Sander, D., Lucas, N., Scherer, K.R., Vuilleumier, P., 2008. Effects of emotional prosody on auditory extinction for voices in patients with spatial neglect. Neuropsychologia 46 (2), 487–496.

Hommel, B., Müsseler, J. Aschersleben, G, and Prinz, W. (2001). "Codes and Their Vicissitudes." The Behavioral and Brain Sciences 24 (05). Cambridge Univ Press: 910–26.

Hommel, B., & Musseler, J. (2006). Action-feature integration blinds to feature-overlapping perceptual events: Evidence from manual and vocal actions. Quarterly Journal of Experimental Psychology, 59, 509– 523.

James, W. (1890). The principles of psychology. New-York: Holt.

Jolij J, Lamme VA (2005) Repression of unconscious information by conscious processing: evidence from affective blindsight induced by transcranial magnetic stimulation. Proc Natl Acad Sci U S A 102: 10747–10751.

Kant, Immanuel (1781/1787/1998). Critique of pure reason, ed. and trans. P. Guyer and A.W. Wood. Cambridge: Cambridge.

Kawasaki, H., Kaufman, O., Damasio, H., Damasio, A.R., Granner, M., Bakken, H., et al., 2001. Single-neuron responses to emotional visual stimuli recorded in human ventral prefrontal cortex. Nat. Neurosci. 4 (1), 15–16.

Killgore, W.D., Yurgelun-Todd, D.A., 2004. Activation of the amygdala and anterior cingulate during the processing of nonconscious sad versus happy faces. NeuroImage 21, 1215 – 1223.

Kitamura, iho S., Katsumi Watanabe, and Norimichi Kitagawa. (2016). “Positive Emotion Facilitates Audiovisual Binding.” Frontiers in Integrative Neuroscience 9: 66.

Krolak-Salmon, Pierre, Catherine Fischer, Alail Vighetto, and François Mauguiere. (2001). “Processing of Facial Emotional Expression: Spatio-Temporal Data as Assessed by Scalp Event-Related Potentials.” The European Journal of Neuroscience 13 (5). Wiley Online Library: 987–94.

Krolak-Salmon, Pierre, Marie-Anne Hénaff, Alain Vighetto, Olivier Bertrand, and François Mauguière. (2004). “Early Amygdala Reaction to Fear Spreading in Occipital, Temporal, and Frontal Cortex.” Neuron 42 (4). Elsevier: 665–76.

Lucas, N., Vuilleumier, P., 2008. Effects of emotional and non-emotional cues on visual search in neglect patients: evidence for distinct sources of attentional guidance. Neuropsychologia 46 (5), 1401–1414.

Luo, Q., Holroyd, T., Jones, M., Hendler, T., Blair, J., 2007. Neural dynamics for facial threat processing as revealed by gamma band synchronization using MEG. NeuroImage 34 (2), 839–847.

Luo, Q., Holroyd, T., Majestic, C., Cheng, X., Schechter, J., Blair, R.J., 2010. Emotional automaticity is a matter of timing. J. Neurosci. 30 (17), 5825–5829.

Maiworm, Mario, Marina Bellantoni, Charles Spence, and Brigitte Röder. (2012). “When Emotional Valence Modulates Audiovisual Integration.” Attention, Perception & Psychophysics 74 (6): 1302–11.

McMains, S., Kastner, S., 2011. Interactions of top-down and bottom-up mechanisms in human visual cortex. J. Neurosci. 31 (2), 587–597.

Muhle-Karbe, P. S., and Krebs, R. M. (2012). On the influence of reward on actioneffect binding. Front. Psychol. 3:450.

Müsseler, J. & Hommel, B. (1997a) Blindness to response-compatible stimuli. Journal of Experimental Psychology: Human Perception and Performance 23(3):861–72. (1997b) Detecting and identifying response-compatible stimuli. Psychonomic Bulletin and Review 4:125–29.

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

Naikar, N., & Corballis, M. C. (1996). Perception of apparent motion across the retinal midline following commissurotomy. Neuropsychologia, 34, 297–309.

Öhman, A, Flykt A, Esteves F (2001a) Emotion drives attention: detecting the snake in the grass. J Exp Psychol Gen. 130:466–478.

Panksepp. J. Affective Neuroscience: The Foundations of Human and Animal Emotions (Series in Affective Science) 1st Edition.

Pinto Y, Neville DA, Otten M, Corballis PM, Lamme VA, de Haan EH, et al. Split brain: Divided perception but undivided consciousness. Brain 2017.

Pourtois, Gilles, Laurent Spinelli, Margitta Seeck, and Patrik Vuilleumier. (2010a). “Temporal Precedence of Emotion over Attention Modulations in the Lateral Amygdala: Intracranial ERP Evidence from a Patient with Temporal Lobe Epilepsy.” Cognitive, Affective & Behavioral Neuroscience 10 (1): 83–93.

NA. 2010b. “Modulation of Face Processing by Emotional Expression and Gaze Direction during Intracranial Recordings in Right Fusiform Cortex.” Journal of Cognitive Neuroscience 22 (9): 2086–2107.

Rotshtein, P., Richardson, M.P., Winston, J.S., Kiebel, S.J., Vuilleumier, P., Eimer, M., et al., 2010. Amygdala damage affects event-related potentials for fearful faces at specific time windows. Hum. Brain Mapp. 31 (7), 1089 1105.

Sergent J. (1987). A new look at the human split brain. Brain; 110: 1375–92.

Schechter, E. (2018). Self-consciousness and “split” brains. Oxford: Oxford.

Schnall S, Harber KD, Stefanucci JK, Proffitt DR. Social support and the perception of geographical slant. Journal of Experimental Social Psychology. 2008; 44:1246–1255.

Sperry, R. (1982). Some effects of disconnecting the cerebral hemispheres. Science, 217, 1223-1226.

Sperling, G. (1960) The information available in brief visual presentations. Psychological Monographs 74:1–29.

Tsuchiya, N., Moradi, F., Felsen, C., Yamazaki, M., Adolphs, R., 2009. Intact rapid detection of fearful faces in the absence of the amygdala. Nat. Neurosci. 12 (10), 1224–1225.

Vuilleumier, Patrik, and Gilles Pourtois. (2007). "Distributed and Interactive Brain Mechanisms during Emotion Face Perception: Evidence from Functional Neuroimaging." Neuropsychologia 45 (1): 174–94.

Vuilleumier, P., N. Valenza, and T. Landis. (2001). "Explicit and Implicit Perception of Illusory Contours in Unilateral Spatial Neglect: Behavioural and Anatomical Correlates of Preattentive Grouping Mechanisms." Neuropsychologia 39 (6): 597–610.

Vuilleumier, P., Schwartz, S., 2001b. Emotional facial expressions capture attention. Neurology 56 (2), 153–158

Vuilleumier, P., Schwartz, S., 2001a. Beware and be aware: capture of spatial attention by fear-related stimuli in neglect. Neuroreport 12 (6), 1119–1122.

Vuilleumier, P., Richardson, M.P., Armony, J.L., Driver, J., Dolan, R.J., 2004. Distant influences of amygdala lesion on visual cortical activation during emotional face processing. Nat. Neurosci. 7 (11), 1271–1278

Vuilleumier, P., Armony, J.L., Clarke, K., Husain, M., Driver, J., Dolan, R.J., 2002. Neural response to emotional faces with and without awareness: event-related fMRI in a parietal patient with visual extinction and spatial neglect. Neuropsychologia 40 (12), 2156–2166.

Watzl, S. 2010. The Significance of Attention. Columbia University PhD Thesis. 2011. Attention as Structuring of the Stream of Consciousness. In Attention: Philosophical and Psychological Essays. Edited by C. Mole, D. Smithies, and W. Wu. New York: Oxford University Press.

Zaidel, W. D., "A View of the World from a Split-Brain Perspective," in E.M.R. Critchley, ed. The Neurological Boundaries of Reality (Northvale, NJ: Aronson, 1995), pp. 161–74; S.M.

Chapter 4: Mindreading, Emotion-Regulation, and Oppression

4.1 Introduction

Folk psychology refers to the commonsense cognitive framework people use in order to “comprehend, predict, explain, and manipulate” behavior and mental states (Churchland, 1994, p. 308). Theorists interested in folk psychology commonly accept that people coordinate their lives via attributing propositional attitudes like beliefs and desires.[15] Until recently, most theorists have argued that people attribute propositional attitudes (henceforth APA) in order to secure *epistemic* benefits like prediction and explanation. These abilities are realized in our psychology via either *theoretical* (e.g., Carruthers, 1996) or *simulative* mechanisms (e.g., Goldman, 2006).

Recently, some theorists have shifted the focus away from *epistemic* towards *practical* functions of APA (e.g., McGeer, 2015; Zawidzki, 2008). According to the *mindshaping* account, we do not attribute propositional attitudes for the sake of prediction and explanation, but in order to *shape* mental states in accordance with social norms. According to these theorists, APA abilities evolved to promote cooperation via a bi-directional exchange of justifications (McGeer, 2015; Zawidzki, 2013).

[15] Most theories draw a distinction between “high-level” and “low-level” mindreading (Goldman, 2006; Waytz and Mitchell, 2011). Roughly, “low-level” mindreading is “mirror-based” mindreading of emotions, intentions, and sensory states (e.g. he is angry at the referee, he’s reaching for the ball); while the “high-level” mindreading involves attributions of propositional attitudes (e.g. she believes that I’m trying to deceive her).

In this work I incorporate mechanistic details behind *theorizing* and *simulation* to outline two underappreciated practical functions APA. In particular, I present converging empirical findings to show that theorizing and simulation evolved for navigating interpersonal emotional struggles.

To start, folk psychologists have underappreciated the need for emotion regulation in interpersonal coordination (Singer, 2006; Decety, 2014). Unregulated interpersonal emotions drain attentional resources necessary for interpersonal coordination (Decety, 2014, p.103). Secondly, contra the mindshaping account, findings suggest that our cognitive mechanisms evolved to ensure a *hierarchical* and not a *bi-directional* exchange of justifications (e.g., Cummins, 1999; Cheng and Tracy, 2014; Hawley, 1999). In particular, those with dominant access to group resources work to maintain the *status quo* by collecting justifications from their low-ranking counterparts, and not vice versa (e.g., Cummins, 1999; Cheng and Tracy, 2014; Hawley, 1999).

Looking closely at the mechanistic details behind theorizing and simulation reveals *how* social coordination survives these affective burdens. Findings suggest that while theorizing fosters emotional distance by "reframing" affective cues from a 3rd person point of view, simulation fosters interpersonal intimacy (Gross and Thompson, 2007; Liotti and Gilbert, 2010; Galinsky et al., 2005). As a result, theorizing and simulation likely evolved for navigating the affective burdens of emotion regulation and social inequality. While theorizing allowed dominant individuals to manipulate norm violators without succumbing to interpersonal emotions, *simulation* allowed the oppressed to form intimate alliances amongst themselves.

This account not only heeds insights from both epistemic and mindshaping accounts, but also promises important ramifications for therapeutic interventions.

4.2 Epistemic Models of Mindreading

Folk psychology purports to explain how ordinary agents coordinate their everyday lives. Most philosophers agree that this coordination proceeds via attributions of *propositional attitudes* like beliefs and desires (e.g., Churchland, 1994). In order to show that people understand and predict mental states and behaviors via attributing beliefs and desires, theorists use examples like the following. Seeing your colleague leave with her coffee mug, you try to predict whether she will get coffee upstairs or across the street. While *you* know that the upstairs dispenser is broken, you still attribute to her the *desire* to go upstairs and the *belief* that there's coffee there. How are you able to discern that her perspective on the world does not contain information that appears immediately obvious to you? That is, how are you able to attribute the *false belief* that there is coffee upstairs?

The above scenario is a version of a paradigmatic task used to study mindreading - the false belief task (Wimmer and Perner, 1983; see Wellman et al. 2001 for a meta-analysis). In this task, the participant (three-year-old) observes two dolls as one of them (Sally) puts a toy in location X and leaves the room. While Sally is out of the room, the other doll (Anne) switches the toy from location X to location Z. As Sally returns, the participant is asked whether Sally will look for the toy in location X (when she thought she left it) or location Z (the toy's current whereabouts).

Until recently, theorists interested in folk psychology have focused on the *mechanisms* underlying our predictions and explanations. While theory theory (henceforth TT) explains propositional attitude attributions in terms of tacit theoretical reasoning, simulation theory (henceforth ST) appeals to tacit *perspective-taking* of the target. Furthermore, some argue that TT but not ST accords *meta-representation* a key role in APA. According to TT, one needs to be able to represent the target’s faulty representation of the world (Perner, 1991). This representation is constructed with the aid of tacit general theoretical knowledge about human psychology (Carruthers, 1996). According to TT, children can eventually pass the false belief test as their conceptual repertoire grows and holds an increasing number of *folk rules* or principles connecting mental states with sensory stimuli, behavioral responses, and other mental states.

In contrast, according to ST, mental state attributions involve a type of *self-projection* into the target’s shoes. By projecting herself into the target’s shoes, the simulator gets on to *what it’s like* to be in the target’s situation (e.g., Goldman, 2009). For example, in trying to discern whether your friend likes his baked potatoes, you project yourself in his shoes while abstracting away from the fact that you yourself love baked potatoes. Simulation theorists do not think that you need much *theoretical knowledge* to engage in mindreading. The simulator “rents out” her own mind to arrive at the attribution. There is no need to engage in *theoretical* reasoning about human psychology because the simulator happens to be such a human herself (Harris, 1992; Heal, 1996). According to ST, children are able to pass the false belief test once they become “more adept at imaginatively identifying with other people and at imagining counterfactual situations” (Davies & Stone, 1995b, p. 6).

Most theorists of mindreading distinguish between *low-level* and *high-level* mindreading (e.g., Goldman, 2009). This distinction was first introduced in the context of clarifying various types of simulation. In particular, while *low-level* simulation refers to “unmediated resonance” or mirroring of emotions and intentions (e.g., being hit with a pang of pain while seeing your partner in pain), high-level simulation refers to attribution of propositional attitudes. High-level simulation (ST) entails projecting oneself into the target's shoes and then feeding the resultant 'pretend' beliefs and desires into our own decision-making process.

Notably, theorists have now introduced a *hybrid* account of mindreading (e.g., Goldman, 2006; Botterill & Carruthers, 1999, Nichols & Stich, 2003). Hybrid accounts still allot the main role to their preferred method of APA while granting certain limited capacities from the opposing method. For example, Botterill and Carruthers argue that TT still plays the default role in APA, but some limited capacity for simulation might be present as well. Likewise, Goldman argues that simulation does the "heavy lifting" in APA, but some type of theorizing may supplement simulation from time to time.

In sum, according to *epistemic* models of high-level mindreading, the nature of folk psychological practices is clear. Folks achieve social coordination by *explaining* and *predicting* mental states and behavior in terms of propositional attitudes like beliefs and desires. The main disagreement centers on the nature of cognitive mechanisms responsible for these attributions (Von Eckardt, 1997).

Shifting the focus away from *epistemic* functions of APA, *mindshaping* theorists argue that we do not attribute propositional attitudes for *prediction* and *explanation*, but for *shaping* mental state in accordance with social norms. According to these theorists, mindshaping evolved to promote cooperation and cognitive homogeneity. While low-level mindshaping ensures rule following via mimicry, imitation, intention perception, and so on; high-level mindshaping ensures rule following via a bi-directional exchange of justifications (McGeer, 2015; Zawidzki, 2013).

4.3 Low-Level Mindshaping

Low-level mindshaping pervades everyday life. Methods of low-level mindshaping ensure cognitive homogeneity through "irresistible conformism and pedagogy" necessary for quick cultural transmission (Zawidzki, 2008, p. xvii). Internalizing our cultural repertoire allows us to understand that certain types of situations call for certain types of beliefs, emotions, and behavioral responses.

Low-level mindshaping methods are "simple" in not relying on attributions of propositional attitudes. Instead, low-level mindshaping relies on more or less direct perception of mental states, character traits, behavioral patterns, and so on. In particular, some mental states like emotions and intentions are directly perceivable (e.g., Gallagher, 2016; Rochat, 2009). For example, the ability to directly experience my dance partner's intentions and emotions allows us fluid coordination. Gallagher summarizes the nature of this engagement in the following way:

> We often understand the actions, responses, intentions and emotions of others, in their embodied comportments – their postures and movements, facial expressions, eye direction, gestures and vocal intonation, as well as their speech, in contexts of our dynamic interactions, all of which happen in the rich pragmatic and social situations of everyday life (ibid, p. 453)

Furthermore, perception of others' mental states is supplemented by the intentional stance, i.e., an attitude of parsing others' behavior into goals and appropriate ways of pursuing them (Zawidzki, 2013). The intentional stance allows us to internalize societal norms via enacting cultural scripts. For example, I need very little information to smoothly coordinate with my chess opponent. After all, there are only a handful of appropriate moves open at any time (McGeer, 2015).

While enacting cultural scripts constrains my social interactions, neither my onlookers nor I make use of propositional attitudes to make sense of these interactions. For example, my interviewer would explain my behavior in terms of *traits* associated with my social role rather than in terms of my beliefs and desires (e.g., she is smiling because she is agreeable and nurturing) (Uleman, 2015; Spaulding, 2016; Bermúdez 2006a; Von Eckardt, 1997). Stereotypical traits are attributed via perceptions of age, gender, and social status. These low-level perceptions activate quick in-group/out-group categorizations and action-guiding stereotypes (Weber and Zou, 2012; Ames et al., 2012; Vorauer et al., 2000).[16]

If everyday coordination can easily be accomplished with *low-level* mindshaping, what are the *practical* dividends of attributing propositional attitudes? This concern has motivated many theorists to question the need for APA (Bermúdez 2006a; Von Eckardt, 1997; Gallagher, 2015). For instance, Bermudez argues that we simply rely on heuristics and pattern detection in social coordination, leaving little need for propositional attitudes (Bermudez, 2006a, p.11).

4.3.1 Practical Dividends of APA

Nevertheless, McGeer and Zawidzki argue that attributing propositional attitudes indeed plays an important practical role in everyday life. That is, instead of aiding in prediction and explanation, APA facilitates *bi-directional* exchanges of *justifications*.

This non-epistemic role is evident when we look at people’s reactions to norm violations. When puzzling over thoughts and behaviors out of step with social norms, we react very differently

[16] Furthermore, we often view norm-governed behavior as expressive of fixed biological essences. Expressing traits particular to societal roles is perceived to be an unintentional result of nature, a consequence of membership in that kind (Mallon, 2016). For example, a man raised in a Hispanic machismo culture would not be able to help but feel “enraged” if his masculinity were threatened in some paradigmatic way. He has internalized the appropriateness of these feelings by observing their paradigmatic displays in his male family members, peers, and popular culture. This expression of his masculinity trait is taken to be “natural” and involuntary. Hence, he is able to acquire societal benefits from its expression (e.g. get excused for punching his offender in the face) (e.g., Griffiths 1997).

from scientists puzzling over unobservables. Instead of adjusting our theory like good epistemologists, we proceed by demanding justifications (McGeer, 2015, p. 166).

This non-epistemic role is further evident when we look at the causal powers of social expectations. In particular, attributing (even false) propositional attitudes to others is akin to *shaping* them into existence. For example, happily believing that your daughter likes practicing the piano eventually causes her to like practicing the piano. The child responds to encouragement and eventually comes to hold the relevant attitudes.[17]

Zawidzki argues that abilities to exchange of justifications allowed our ancestors to reach new levels of cooperation (e.g., Sterelny, 2006). Zawidzki describes this evolutionary pressure in the following way (my emphasis):

> When it comes to normative sanctions, propositional attitude attribution is part of a negotiated *give-and-take* aimed at determining the normative status of an interpretive target, and hence whether or which sanctions are appropriate (ibid, p. 222).

In particular, APA allows the receiver of a promise to question the veracity of others' commitments in the face of aberrant behavior (e.g., "I know she said she is committed to this cause, but is that how she *really* feels given her recent behavior?"). Furthermore, it allows the maker of the promise to search through her own propositional attitudes in order to find credible justifications for her aberrant behavior (e.g., "I know I promised to bring two bushels of wheat from the outpost, but I misunderstood your request."). As a result, APA pays practical dividends by allowing us to discern committed partners as well as negotiate our failings (Zawidzki, 2008; Sellars, 1997; Brandom, 1994; Frankish, 2004).

[17] Common examples of self-fulfilling prophesies include both negative and positive outcomes (Biggs, 2013).

4.4 The Affective Burdens of Social Selves

Now that we have outlined the main tenets of the mindshaping account, it's time to take a closer look at its shortcomings. According to the mindshaping account, APA evolved to ensure a *bi-directional* exchange of justifications. In particular, given the efficacy of low-level mindshaping in everyday social coordination, APA serves a distinct practical role by facilitating the "give and take" in the "court of law". That is, partners in cooperation attribute propositional attitudes in order to establish the mental reality behind explicit commitments (e.g., "I know he said he's committed to contributing, but is that how he *really* feels?").

Nevertheless, there are reasons to doubt that APA evolved to ensure a bi-directional exchange of justifications. In particular, this account mischaracterizes the burdens of interpersonal coordination. Findings suggest that the primary practical burdens of interpersonal coordination are *affective* in kind (Decety, 2014; Cummins, 1999; Cheng and Tracy, 2014; Hawley, 1999). That is, any account of APA needs to explain *how* attributing propositional attitudes helps us navigate the affective burdens of emotion regulation and hierarchical power relations. Let's take a closer look at each burden in turn.

4.4.1 The Struggle for Emotion Regulation

As we've seen in the previous section, low-level mindshaping ensures a measure of automatic affective resonance (Gallagher, 2016, p. 263). For example, we automatically respond to others' intentions and affective states by tracking their eye movements, posture, intonations, and so on (ibid, p. 453). Interpersonal affective resonance emerges at a very young age. Infants respond to their immediate environment via mimicry and somato-sensorimotor resonance (Lodder et al., 2014; Rochat, 2009). For example, infants become instantly distressed as soon as another infant starts to cry (Dondi et al., 1999; Martin & Clark, 1987).

To a lesser extent, this type of emotional attunement is also present in adults. Findings on mirror neurons support the existence of primitive mimicry or re-experiencing others' motor intentions and emotions (e.g., Lacoboni et al., 2005; Kaplan & Iacoboni, 2006). Given a large cortical overlap between regions responsible for processing firsthand and others' somatic and affective experience, we re-experience a large portion of others' low-level states (Gallese et al. 1996; Rizzolatti et al. 1995; Grezes and Decety 2001; Singer et al. 2004; Keysers et al. 2004; Wicker at al. 2003).

In the same vein, studies on emotion contagion detail our disposition to "catch" others' emotions (Hatfield et al., 1994). For example, Hsee et al. (1990) found that participants who watched a videotape of a target describing sad or happy moments proceeded to experience these emotions themselves. Similarly, participants watching others make disgust faces proceeded to experience disgust themselves.

In short, people are *affectively immersed* in their interactions. We automatically respond to our interlocutors via mimicking their facial expressions, re-experiencing their affective states, and so on (Lodder et al., 2014; Rochat, 2009; Lacoboni et al., 2005; Kaplan & Iacoboni, 2006; Hatfield et al., 1994).

Nevertheless, there are plenty of everyday situations in which this type of *emotional immersion* would severely hinder social coordination. As many family holiday survivors know, being emotionally attuned to one's family members could be disastrous for successful social coordination. Consider the following example of emotional attunement. As your partner exclaims in distress, you immediately "catch" their general emotion and become distressed yourself (Elfenbein, 2014; Hatfield et al., 1994). If this automatic reaction is not inhibited, you will fail to respond appropriately in service of social coordination (e.g., your own distress will now prevent nuanced coordination) (e.g., Singer, 2006). Decety summarizes this regulatory need in the following way:

> If not regulated, affective resonance can be costly, both physiologically and cognitively, and can eventually conflict with the observer's capacity to be of assistance to the other (Decety and Lamm, 2009b). Difficulty inhibiting or reducing an emotional response may deplete the resources available for other aspects of self-regulation…, hindering the ability to function adaptively and appropriately (Decety, 2014, p.103).

In sum, philosophers interested in folk psychology have underappreciated the affective burdens plaguing interpersonal coordination. Hence, accounts detailing the adaptiveness of APA need to explain *how* these mechanisms allow us to overcome the affective burdens of interpersonal coordination.

4.4.2 Struggles Navigating Social Hierarchies

Mindshaping theorists argue that APA evolved to ensure cooperation via a bi-directional exchange of justifications. These theorists make use of the *social exchange* hypothesis (Cosmides and Tooby 1992, 1994). According to social exchange hypothesis, cognitive mechanisms evolved in the context of cooperation. In this context, reciprocity could only be rewarded if humans evolved cognitive mechanisms for cheater detection (i.e., detection of parties who enjoy group benefits without contributing). According to the mindshaping account, high-level mindshaping evolved to facilitate *bi-directional* cheater detection (i.e., helping cooperative parties track *justifications* for aberrant behavior).

Nevertheless, findings reveal that our mental architecture is attuned to a *hierarchical* and not *bi-directional* exchange of justifications. In particular, according to the *social dominance* hypothesis, our mental architecture evolved in the context of within group competition *under the guise* of cooperation (e.g., Cummins, 1999; Cheng and Tracy, 2014; Hawley, 1999). Contra the mindshaping account, detecting norm violations is not a concern plaguing *all* interpersonal coordination. In fact, it is a concern *specific* to those monopolizing access to group resources. Those with dominant access to group resources work to maintain the *status quo* by collecting justifications from possible norm violators, and not vice versa (e.g., Cummins, 1999; Cheng and Tracy, 2014; Hawley, 1999). Cummins (1999) summarizes this hypothesis in the following way:

> Violation-detection is implicated in the acquisition and maintenance of dominance rank. Low-ranking individuals attempt to improve their access to competitive resources through acts of cheating and deception. Dominant individuals attempt to maintain priority of access to resources by detecting and thwarting acts of cheating and deception (ibid, p. 231).

In other words, collecting justifications is not a way of ensuring egalitarian cooperation but a way of keeping the dominant in charge. Those benefitting from the current hierarchical organization have a special interest in maintaining the *status quo* by punishing those trying to change their access to resources by cheating and deception.

Converging lines of findings support the social dominance hypothesis. If our cognitive mechanisms evolved to facilitate a *bi-directional* exchange of justifications, social rank should make no difference to their functioning. However, it appears that social rank plays a key role in our ability to track norm violations and their respective justifications.

Awareness of social rank emerges early in cognitive development. Specifically, children as young as 24 months organize their interactions according to implicit rules of social dominance

(Frankel and Arbel 1980; Hold-Cavell and Borsutzky 1986; La Freniere and Charlesworth 1983; Rubin and Caplan 1992). Furthermore, it appears that people retain significantly better memories of *low-ranking* norm violators than their *high-ranking* counterparts (e.g., Mealey et al.,1996).

In order to further test the social dominance theory, Cummins (1999) broke up participants into two groups. The first group was assigned to a cheater-detection task, while the second group was assigned to a truth testing task. Consider the following description of the two tasks:

> In the cheater-detection version of the task, reasoners were told that there was an important rule in the dormitory, namely, that *if someone is assigned to tutor a study session, that person is required to tape record the session.* The reasoners were then shown four cards. One side of each card indicated whether or not the person in question had been assigned to tutor a particular study session and the other side indicated whether or not the person had in fact tape recorded the session. The faces of each card showed, respectively, "Assigned to tutor the session," "NOT assigned to tutor the session," "Taped the session," and "Did NOT tape the session." Reasoners were instructed to *select the card or cards that need to be turned over to determine whether or not the person followed the rule.* In the truth-testing version of the task, reasoners were told that study sessions took place in the dorm, but no mention was made of a rule concerning them. Instead, they were asked to imagine that they had overheard someone say, "If I'm assigned to tutor a session, I always tape record the session." They were then shown same four cards described earlier and were asked to select the card or cards that need to be turned over to determine whether or not the person told the truth (ibid, p. 234).

These two tasks were performed across four conditions:

> - High ranking: The reasoner adopted the perspective of a high-ranking individual (Resident Assistant) checking on low-ranking individuals (Students).
> - Low ranking: The reasoner adopted the perspective of a Student checking on Resident Assistants.
> - Equally high ranking: The reasoner adopted the perspective of a Resident Assistant checking on other Resident Assistants.
> - Equally low ranking: The reasoner adopted the perspective of a Student checking on other Students (ibid, p. 234).

Experiments revealed that individuals asked to assume socially dominant roles were significantly more vigilant for norm violations than individuals who assumed lower-ranking roles. Justifications were demanded from the lower-ranked individuals and not vice versa. To isolate the effects of social rank, experimenters showed that this asymmetry was not maintained for other forms of social reasoning tasks such as truth-testing.

Now that we have detailed two underappreciated struggles plaguing interpersonal coordination, it's time to take a closer look at the mechanistic details behind theorizing and simulation. If the mechanistic details behind theorizing and simulation reveal their adaptiveness in navigating these particular struggles, we can infer the practical functions of APA.

4.5 The Emotion-Regulatory Role of APA

As detailed in the previous sections, automatic interpersonal resonance could hinder nuanced social coordination (e.g., Decety, 2014). Theorists interested in the adaptive functioning of APA need to explain how interpersonal coordination survives this obstacle. To meet this challenge head on, I propose to look back at the mechanistic details offered by the epistemic account of mindreading.

Findings suggest that both ST and TT involve *selective inhibition* of one's current mental states in service of mental state attribution (Aboulafia-Brakha et al., 2011; Singer, 2006). The regulatory function is revealed in imaging findings. In particular, OFC (orbitofrontal cortex) and vmPFC (ventromedial prefrontal cortex) responsible for attributing propositional attitudes function as top-down mediators of sub-cortically realized affective responses (e.g., Zelazo et al., 2008; Diamond, 2002; Decety & Michalska, 2010; Cheng et al., 2007).

In particular, Cheng et al. (2007) found that medical practitioners are particularly adept at exploiting the emotion-regulatory role of APA. Specifically, fMRI scans on medical practitioners viewing patients being pricked with needles revealed activation of regions responsible for attributing propositional attitudes (e.g., vmPFC and OFC), as well as increased connectivity between these regions and regions responsible for low-level empathic responses. Increased connectivity suggests frequent use of this regulatory network. Unusually frequent use of this regulatory network is to be expected given the unusual everyday demands of the medical profession.

In short, given the hazardous nature of *low-level* mindshaping and findings outlined above, attributions of propositional attitudes play an important emotion-regulatory role in social coordination.

4.5.1 Unique Emotion-Regulatory Benefits of Theorizing and Simulation

Since leading theorists now accept that both simulation and theorizing play some role in APA (e.g., Goldman, 2006; Botterill & Carruthers, 1999, Nichols & Stich, 2003), looking at the emotion-regulatory functions of APA might shed light on the exact nature of this combinatorial account. That is, how do the regulatory benefits ensured by simulation differ from those ensured by theorizing? Furthermore, how do these benefits fit with what we know about the evolutionary context of social dominance? That is, if cognitive mechanisms of APA evolved in the context of social dominance, how did various methods of APA allow the dominant individuals to maintain preferential access to resources and the oppressed to navigate the social hierarchy?

In this section I detail findings suggesting the following division of labor between theorizing and simulation. While theorizing fosters emotional distance by “reframing” affective cues from a 3rd person point of view, simulation fosters interpersonal intimacy (Gross and Thompson, 2007; Liotti and Gilbert, 2010; Galinsky et. al., 2005). Hence, it is reasonable to hypothesize that while theorizing allowed the dominant to manipulate potential norm violators without succumbing to costly interpersonal resonance, simulation allowed the oppressed to form intimate alliances amongst themselves.

Skilled manipulation requires a measure of *emotion insulation*. For instance, in order to deceive others that the *status quo* is in fact the best arrangement for *all* parties, dominant individuals would need to be able to downregulate emotion contagion from exploited individuals (e.g., Bilewicz, 2016). On the other hand, the exploited would need to be able to build social bonds with similar others to ensure feelings of intimacy in hopes of overcoming their circumstance (e.g., Galinsky et al., 2005).

Recall that TT but not ST allots *theorizing* a key role in mental state attribution. Theorizing allows a certain type of “reframing” of low-level affective cues. This reframing allows subjects to change the phenomenological impact of their emotions (e.g., Gross and Thompson, 2007; Gross 2015; Williams et al., 2009). For example, in regulating his test anxiety, a student “steps back” from his emotional experience and attempts to *re-interpret* his anxious cues as “excitement about an opportunity to excel.” This psychological maneuver allows a measure of emotion insulation (i.e., the sense of urgency and the nature of the affective state is changed). In using TT, one uses knowledge of psychological principles and folk rules to re-interpret affective cues to a palatable form.

In contrast, since ST requires the agent to “rent out” her own mind in arriving at the attribution, this type of emotional distancing is no longer available. Self-projection during simulation creates a greater vulnerability for *emotion contagion* than theorizing (Gallagher, 2012). For example, in simulating the target’s distress, the manipulator might “catch” the target’s emotion and have trouble going through with his manipulative action. Even if the attributer “abstracts away” from his personal preferences in the simulation procedure, he still has to *undergo* the experience from the first-person point of view.

If simulation is so hazardous, how does this vulnerability serve the emotion regulation needs created by low-level resonance? While this vulnerability could be a liability during manipulation, it could also be invaluable during bonding. Studies have shown that perspective taking creates and maintains social bonds and increases feelings of psychological closeness (e.g., Galinsky et al., 2005; Cialdini et al., 1997). In particular, simulating others’ mental states fosters feelings of intimacy and mutual understanding. In turn, these feelings provide emotion regulation benefits by fostering comforting feelings of belongingness and acceptance (Galinsky et al., 2005). For example, simulating your partner strengthens your intimate bonds and feelings of mutual understanding.

Using the wrong APA mechanism would be highly disadvantageous. For example, using simulation and not theorizing during manipulation would impede the ability to undertake utilitarian actions (e.g., Majdandžić et al., 2012). Similarly, using theorizing and not simulation would hamper interpersonal intimacy (e.g., Koenigsberg et al., 2011). Let’s consider these findings in more detail.

Majdandžić et al. (2012) investigated the effects of simulation on our abilities to undertake utilitarian actions. In the experimental priming condition, participants were prompted to simulate the target’s mental states. In the control priming condition, participants were not prompted to imagine themselves in the targets’ shoes. Afterward, participants were asked to make a classic trolley moral dilemma decision (e.g., participants were asked whether they would sacrifice the previously simulated target). Previous simulation was significantly correlated with failing to authorize a utilitarian response to the moral dilemma using the target.

On the other hand, using theorizing and not simulation during interpersonal bonding could undermine feelings of intimacy (e.g., Koenigsberg et al., 2009). In particular, Koenigsberg et al. (2009) had subjects attribute mental states to targets using something like folk-theory. Remember that according to TT, subjects sub-consciously "piece together" various clues in theorizing about the target's mental state from the 3rd person point of view (e.g., Carruthers, 2011). Likewise, in this study, subjects were explicitly asked to theorize about the targets' mental states from the 3rd person point of view (Koenigsberg et al., 2009; Ochsner & Gross, 2005; 2008). Researchers found that this strategy impeded feelings of interpersonal intimacy between the subjects and their targets.

In sum, we have seen some reasons to suspect that each attribution method provides unique ways of processing interpersonal emotions. While ST facilitates feelings of intimacy, TT allows a level of emotional insulation often seen in social manipulation (Galinsky et al., 2005; Cialdini et al., 1997).

Are these predictions borne out? Indeed, it appears that participants with a low subjective status are, in fact better at simulating *what it's like* to be in others' shoes than their high-ranking counterparts (Kraus et al., 2010).[18] In contrast, participants asked to assume *high-ranking* roles are indeed better at using TT to represent others' faulty representations of the world and worse at using ST than their low-ranking counterparts (Blader et al., 2016 Fiske, 1993; Galinsky et al., 2006; Lammers et al., 2008; Tjosvold and Sagaria, 1978; Schmid-Mast et al., 2009).

[18] Subjective social status (SSS) is often defined as one's belief about one's location in a status order" (e.g., Singh-Manoux et al., 2005).

4.5.2 Self-Directed APA

In this section I propose to expand this account to self-directed attribution of propositional attitudes. Looking closely at the use of these mechanisms in the modern context of social oppression might help us outline important therapeutic interventions.

Getting onto our own propositional attitudes takes a bit more than being hit with a pang of hunger or becoming aware of an itch. While many theorists argue that we have more or less direct access to our sensory states, direct access to our propositional attitudes is a matter of some debate.[19] While some theorists argue that we use different psychological mechanisms for self-directed and other-directed APA, respectively (e.g., Goldman, 2009), others propose that we use the same psychological mechanism in both self and other-directed APA (Carruthers, 2011). In particular, Carruthers argues that we use a single sub-personal "interpretive" or non-direct process (taking sensory information as input and producing attitudes as output) for all APA (i.e., attributing attitudes to all others as well as the self).

I do not hope to resolve the debate on the existence of "direct" introspection here. Instead, I outline distinct emotion-regulative differences between ST and TT, respectively. In particular, while self-directed simulation fosters feelings of *diachronic continuity*, self-directed theorizing fosters a sense of *self-control*.

[19] Access is said to be "direct" if it allows us knowledge of our own mental states that is itself not based on knowledge of other things in the world (Schwitzgebel, 2019). Roughly, *introspection* is typically defined as a mental process that allows us *direct* access to our currently ongoing, or very recently past, mental states or processes. Notably, the existence of these self-regulative differences does not indicate the existence of direct introspective access. That is, simply because either simulation or theorizing could be used in self-directed attribution of propositional attitudes, it does not rule out the existence of a third self-directed method like direct introspection.

One way to figure out whether I hold the attitude that “gymnastics is important to me” is to gauge my reaction to various gymnastics related hypothetical scenarios. If my heart fills with horror at the prospect of forgetting to watch gymnastics, I can immediately sense its importance in my life. This simulative process is commonly referred to as *mental time travel* and activates the same neural network responsible for simulation.[20] This process is perspectival, phenomenologically rich, and has a distinct *autonoetic* component or “a unique awareness of re-experiencing in the here and now” (Tulving, 1985; Hassabis and Maguire, 2007). That is, when I imagine forgetting to watch gymnastics, I experience a distinct phenomenological awareness of what it would be like to undergo that experience.

Another way to figure out whether I hold a particular attitude is via self-directed theorizing. According to Carruthers (2011), our sub-conscious interpretive mechanism takes various pieces of evidence (e.g., extended collection of gymnastics memorabilia) and applies the same general folk-psychological theory it would apply to anyone: “anyone who acts this way must really like gymnastics; hence, I like gymnastics.”

Simulative and interpretive methods of self-directed APA offer distinct emotion-regulative benefits and pitfalls. In the previous section we’ve seen that other directed theorizing allows dominant to avoid interpersonal affective resonance. In this way, these individuals could avoid encountering the emotional aftermath of their actions. Likewise, self-directed theorizing also helps subjects avoid volatile intrapersonal emotions.

[20] The default neural network is activated during simulation of mental states, episodic memory, and simulation of hypothetical scenarios (Hassabis et al., 2014; St. Jacques, et al., 2014; e.g. De Brigard, 2014; Klein, 2015; Tulving, 1985; Hassabis and Maguire, 2007).

During self-directed theorizing, one tacitly applies different models or “theories” to one’s current mental state. In self-directed theorizing we “step-back” from the experience and interpret it from the 3rd person point of view (e.g., “anyone who acts this way must really like gymnastics; hence, I like gymnastics.”). Once we look at our experience from the 3rd person point of view, we tacitly shift our attention away from actually *undergoing* the experience. For example, a dominant individual might re-interpret his perception of others’ facial expressions as not responsive to his unjust actions but as responsive to other extraneous circumstances (Gross, 1998).

In the same vein, Robinaugh and McNally (2010) show that adopting the observer perspective helps an experience become “incongruent with one’s sense of self” (ibid, p. 650). In other words, using self-directed theorizing would allow the dominant individual to re-interpret aversive interpersonal emotions as *irrelevant* to his sense of identity. In fact, researchers found that adopting the observer perspective is a mechanism of *cognitive avoidance*, allowing a measure of emotional distance (Wilson and Ross, 2003; Lemogne et al., 2009). In this way, dominant individuals could avoid confronting potentially uncomfortable ramifications of their actions during self-reflection (e.g., Dubois et al., 2015).

In contrast, extended use of self-directed simulation in the context of social oppression could be maladaptive. For example, shame is an interpersonal emotion triggered by real or imagined disapproval of others with high social standing (Tangney, 1992; Averill, 1982; Tangney, 1992; Andrews et al., 2000; Bennett et al., 2005; Morrison et al. 2001). Given the volatile nature of this interpersonal emotion, self-directed simulation is not the best method for self-directed attitude attribution. Besides its rich phenomenological component, mental time travel is rigged by one’s current emotional state (D’Argembeau and Van der Linden, 2004). If one does engage in self-directed simulation while experiencing shame, one will likely project oneself to phenomenologically similar scenarios. For example, in experiencing shame about my obtuse contribution to the discussion, I inevitably project myself to similar scenes of social rejection (e.g., being picked last for a team, getting stood up for a date, and so on). These "affect-colored glasses" present one of the biggest obstacles in cognitive behavioral therapy, i.e., depressed patients simply cannot imagine or "see" things getting better due to their current mood (e.g., Uher et al., 2013). Likewise, oppressed individuals internalize negative self-conscious emotions and have trouble "seeing" past their current circumstances (Chung et al., 2011).

To make matters worse, findings suggest that neural activation associated with self-direction simulation may directly predispose the brain towards a non-attentive state and could be responsible for poor performance on attention-heavy tasks (D'Argembeau et al., 2005; Greicious et al., 2007).

In sum, while targets of oppression might often use simulation to form intimate bonds with similar others, extending this method of APA towards the self incurs additional psychological costs to an already vulnerable population. As a result, therapeutic interventions for individuals with low subjective social standing would include de-emphasizing the use of simulation in everyday life. Instead, practicing assuming high-ranking positions could help build up theorizing skills beneficial for emotional insulation and self-control.

Are these predictions borne out? Indeed, findings suggest that practicing self-directed theorizing provides greater emotion regulation benefits to participants with lower socioeconomic status than to their high-status counterparts (e.g., Troy et al., 2017). This difference makes sense within our account. In particular, high-status individuals might benefit less from practicing self-directed theorizing due to a ceiling effect. That is, since high-ranking individuals have *already* benefitted from self-directed theorizing throughout their life span, additional practice in this method would not make as much of a difference to their emotion regulation skills as it would to the skills of their low-status counterparts.

4.6 Conclusion

I started out by outlining two underappreciated obstacles in interpersonal coordination: obstacles in emotion regulation and obstacles navigating the social hierarchy. I then incorporated mechanistic details behind *theorizing* and *simulation* to explain *how* these obstacles are negotiated in everyday life. While theorizing allowed dominant individuals to manipulate norm violators without succumbing to interpersonal emotions, simulation allowed the oppressed to form intimate alliances amongst themselves.

As a result, we have outlined two novel practical functions of APA with important ramifications for therapeutic interventions. While there is no denying that APA might also be used in prediction, explanation, and mindshaping, these results cannot be achieved without first overcoming the affective obstacles at hand.

References

Ames, D. R., Weber, E. U., & Zou, X. (2012). Mind-reading in strategic interaction: The impact of perceived similarity on projection and stereotyping. *Organizational Behavior and Human Decision Processes, 117*(1), 96-110.

Andrews, B., Brewin, C. R., Rose, S., & Kirk, M. (2000). Predicting PTSD symptoms in victims of violent crime: the role of shame, anger, and childhood abuse. *Journal of Abnormal Psychology*, 109, 69–73.

Averill, J. R. (1982). Anger and aggression: An essay on emotion. New York: Springer-Verlag.

Baron-Cohen, S., Jolliffe, T., Mortimore, C. & Robertson, M. (1997) Another advanced test of theory of mind:

Evidence from very high functioning adults with autism or Asperger syndrome. *Journal of Child Psychology and Psychiatry* 38:813–22

Bermúdez, J. L. (2005). Philosophy of Psychology: A Contemporary Introduction. London, Routledge.

Bermúdez, J. L. (2006a). "Commonsense psychology and the interface problem: Reply to Botterill." SWIF

Philosophy of Mind Review 5(3): 54–57.

Benson, P., (1990), "Feminist Second Thoughts About Free Agency," *Hypatia,* 3: 47–64.

Bennett DS, Sullivan MW, Lewis M., (2005) Young children's adjustment as a function of maltreatment, shame,

and anger. *Child Maltreat*. 2005;10(4):311–323

Botterill, G. and Carruthers, P. (1999). The Philosophy of Psychology. Cambridge University Press.

Bushman BJ, Moeller SJ, Crocker J (2011) Sweets, Sex, or Self-Esteem? Comparing the Value of Self-Esteem

Boosts With Other Pleasant Rewards. *Journal of Personality* 79: 993–1012.

Brewer, M. B., & Brown, R. J. (1998). Intergroup relations: McGraw-Hill.

Carruthers, P. (1996). "Simulation and self-knowledge." In P. Carruthers and P. K. Smith (eds.), Theories of

Theories of Mind. Cambridge, Cambridge University Press, pp. 22–38.

Carruthers, P. (2011). The Opacity of Mind: The Cognitive Science of Self-Knowledge. Oxford, Oxford University Press.

Crothers, M., & Warren, L. (1996). Parental antecedents of adult codependency. *Journal of Clinical Psychology*, 52(2), 231–239.

Currie, G., & Ravenscroft, I. (2002). Recreative minds. Oxford, UK: Oxford University Press.

Clement, R. W., & Krueger, J. (2002). Social categorization moderates social projection. *Journal of Experimental Social Psychology*, 38(3), 219-231.

Cialdini, R. B., Brown, S. L., Lewis, B. P., Luce, C., & Neuberg, S. L. (1997). Reinterpreting the empathy-altruism relationship: When one into one equals oneness. *Journal of Personality and Social Psychology*, 73, 481-494.

Christman, J., (2004). Social and Political Philosophy: A Contemporary Introduction, London: Routledge.

D'Argembeau, A., & Van der Linden, M. (2004). Phenomenal characteristics associated with projecting oneself back into the past and forward into the future: Influence of valence and temporal distance. *Consciousness and Cognition*, 13, 844–858.

Davies, M. and Stone, T (1995b), Mental Simulation: Evaluations and Applications—Reading in Mind and Language, Oxford: Blackwell Publishers.

De Brigard, F. (2014). Is memory for remembering? Recollection as a form of episodic hypothetical thinking. *Synthese,* 191, 1–31

Eisenberger, N. I., Taylor, S. E., Gable, S. L., Hilmert, C. J., & Lieberman, M. D. (2007). Neural pathways link social support to attenuated neuroendocrine stress responses. *NeuroImage*, 35(4), 1601–1612.

Epley, N., & Waytz, A. (2010). Mind perception. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), Handbook of Social Psychology (5th ed., Vol. 1, pp. 498-451). Hoboken, NJ: Wiley.

Fischer, J. L., Spann L., & Crawford, D. (1991). Measuring codependency. Alcoholism Treatment Quarterly, 8, 87-100. controlled trial. *Behavior Therapy*, 43, 666–678.

Friedman, M., (1997) "Autonomy and Social Relationships: Rethinking the Feminist Critique," in Meyers, ed. (1997), pp. 40–61.

Fonagy, P., Gergely, G., Jurist, E. L., & Target, M. (2002). Affect regulation, mentalization, and the development of self. New York: Other Press

Frankfurt, H., (1988c), "Freedom of the Will and the Concept of a Person," in Frankfurt 1988a, 11–25.

Gallagher, S., and Varga, S. (2015). Social cognition and psychopathology: A critical overview. World Psychiatry, 14, 5–14.

Galinsky, A. D., Ku, G., & Wang, C. S. (2005). Perspective-taking and self-other overlap: Fostering social bonds and facilitating social coordination. Group Processes and Intergroup Relations, 8, 109–124.

Gallagher, S., and D. Zahavi, 2008, The Phenomenological Mind: An Introduction to Philosophy of Mind and Cognitive Science, New York: Routledge.

Gallese, V., & Goldman, A. (1998). Mirror Neurons and the Simulation Theory of Mind-Reading. *Trends in Cognitive Sciences*, 2, 493-501.

Goldman A.I. (2006). Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading. New York: Oxford Univ. Press.

Grecucci, A., Thneuick, A., Frederickson, J., and Job, R. (2015b). "Mechanisms of social emotion regulation: from neuroscience to psychotherapy," in Handbook on Emotion Regulation: Processes, Cognitive Effects and Social Consequences, ed. M. L. Bryant (New York, NY: Nova Publishing), 57–84.

Gross, J.J., & Thompson, R.A. (2007). Emotion regulation: Conceptual foundations. In J.J. Gross (Ed.), Handbook of emotion regulation (pp. 3–26). New York: Guilford.

Gross JJ. (2015). Emotion regulation: current status and future prospects. *Psychol. Inq*. 26:1–26

Harris, P. L. (1992). From simulation to folk psychology: The case for development. Mind and Language, 7, 120-144.

Hassabis D., Maguire E.A., (2007). Deconstructing episodic memory with construction. *Trends Cogn Sci* 11:299– 306

Hassabis, D., Spreng, R. N., Rusu, A. A., Robbins, C. A., Mar, R. A., and Schacter, D. L. (2013). Imagine all thepeople: how the brain creates and uses personality models to predict behavior. Cereb. Cortex doi: 10.

Heal, J. (1996). Simulation and cognitive penetrability. Mind and Language 11:44– 67

Heal, J. (2003). Mindreading: an integrated account of pretence, self-awareness, and understanding of other minds. *Mind* 114(453):181-184.

Heatherton TF, Wyland CL, Macrae CN, Demos KE, Denny BT, Kelley WM (2006): Medial prefrontal activity differentiates self from close others. *Soc Cogn Affect Neurosci* 1:18–25.

Iacoboni M, Molnar-Szakacs I, Gallese V, Buccino G, Mazziotta JC, Rizzolatti G (2005) Grasping the Intentions of

Others with One's Own Mirror Neuron System. PLoS Biol 3(3): e79.

Jenkins, A. C., Macrae, C. N. & Mitchell, J. P. (2008) Repetition suppression of ventromedial prefrontal activity during judgments of self and other. *Proceedings of the National Academy of Sciences* USA 105:4507 – 12.

Kaplan JT, Iacoboni M. 2006. Getting a grip on other minds: mirror neurons, intention understanding and cognitive empathy. Soc. Neurosci. 1:175–83

Kelley, W. M., Macrae, C. N., Wyland, C. L., Caglar, S., Inati, S., & Heatherton, T. F. (2002). Finding the self? An event-related fMRI study. J. Cogn. Neurosci., 14, 785– 94

Koenigsberg, H. W., Fan, J., Ochsner, K. N., Liu, X., Guise, K. G., Pizzarello, S., … Siever, L. J. (2009). Neural correlates of the use of psychological distancing to regulate responses to negative social cues: a study of patients with borderline personality disorder. *Biological psychiatry*, 66(9), 854–863.

Kring, A. M., & Werner, K. H. (2004). Emotion regulation in psychopathology. In P. Philippot & R. S. Feldman (Eds.), The regulation of emotion (pp. 359 –385). Mahwah, NJ: Erlbaum

Kurzban, R., & Neuberg, S. (2005). Managing Ingroup and Outgroup Relationships. In D. M. Buss (Ed.), The handbook of evolutionary psychology (pp. 653-675). Hoboken, NJ, US: John Wiley & Sons Inc.

Klein, SB. (2015). What memory is. WIRES Cognitive Science. 6(1):1–38.

Langone, M. D., (1995). An investigation of a reputedly psychological abusive group that targets college students (Tech. Rep)., Boston: Boston University, Danielson Institute.

Lemogne, C., Bergouignan, L.,Piolino, P., Jouvent, R., Allilaire, J.F.,& Fossati, P. (2009).Cognitive avoidance of intrusive memories and autobiographical memory: specificity, autonoetic consciousness, and self-perspective. Memory,17, 1e7.

Lodder, P., Rotteveel, M., & Van Elk, M. (2014). Enactivism and neonatal imitation: Conceptual and empirical considerations and clarifications. *Frontiers in Psychology*, 5.

Mackenzie, C. & N. Stoljar (2000a). Introduction: Autonomy refigured. In Mackenzie & Stoljar (eds.), 3–31. Majdandziˇ c J, Bauer H, Windischberger C, Moser E, Engl E, Lamm C. (2012). The human factor: behavioral and neural correlates of humanized perception in moral decision making. PLoS ONE 7:e47698.

Malle BF, Hodges SD, (2005). Other Minds: How Humans Bridge the Divide Between Self and Others. New York: Guilford

Mallon, R., (2016), The Construction of Human Kinds, Oxford: Oxford University Press.

McGeer, V. (2015) Mind making practices: the social infrastructure of self-knowing agency and responsibility. Philosophical Explorations 18(2) 251 – 289.

Mele, A. (2012d). Autonomy and Neuroscience. In Radoilska, L. (ed.) Autonomy and Mental Health. New York: Oxford University Press.

Meyers, D. (2000a). "Feminism and Women's Autonomy: The Challenge of Female Genital Cutting," *Metaphilosophy*, 31: 469–491.

NA. (1989), Self, Society and Personal Choice, New York: Columbia University Press.

Michael, J., Christensen, W., & Overgaard, S. (2014). Mindreading as social expertise. Synthese, 191(5), 817–840.

Mitchell, J. P., Banaji, M. R., & MacRae, C. N. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of cognitive neuroscience*, 17(8), 1306-1315.

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron*, 50(4), 655-663.

Mitchell, J.P., (2009). Inferences about mental states. Philos Trans R Soc Lond B Biol Sci., 364:1309–1316.

Morrison, D., & Gilbert, P. (2001). Social rank, shame and anger in primary and secondary psychopaths. *Journal of Forensic Psychiatry*, 12, 330 –356.

Morton, A. 1996. Folk psychology is not a predictive device. Mind 105: 119–37.

Nahmias, E., Coates, D. Justin, Kvaran, Trevor, 2007. "Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions," *Midwest Studies in Philosophy*, 31: 214–242.

Navarette, C. D., Kurzban, R., Fessler, D. M. T., & Kirkpatrick, L. A. (2004). Anxiety and intergroup bias: Terror management or coalitional psychology. *Group Processes and Intergroup Relations*, 7, 370–397

Nichols, S. and Stich, S. P., (2003). Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding of Other Minds, Oxford: Oxford University Press.

Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain: A meta-analysis of imaging studies on the self. *NeuroImage*, 31, 440–457

Ochsner, K. N., & Gross, J. J. (2008). Cognitive Emotion Regulation: Insights from Social Cognitive and Affective Neuroscience. Current directions in psychological science, 17(2), 153–158.

Perner, J., and Kühberger, A., (2005). “Mental Simulation: Royal Road to Other Minds?”, in Bertram F. Malle and Sara D. Hodges (eds.), Other Minds: How Humans Bridge the Divide Between Self and Others, New York: Guilford Press, pp. 174–187.

Perner, J. (1991). Understanding the Representational Mind, MIT Press.

Pettigrew, T. F. (1979). The ultimate attribution error: Extending Allport's cognitive analysis of prejudice. *Personality and Social Psychology Bulletin*, 5(4), 461- 476.

Phan, K.L., Wager, T.D., et al., 2004b. Functional neuroimaging studies of human emotions. CNS *Spectr*. 9 (4), 258 – 266.

Robinaugh, D. J., & McNally, R. J. (2010). Autobiographical memory for shame or guilt provoking events: Association with psychological symptoms. *Behaviour Research and Therapy*, 48, 646-652.

Rochat, P. (2009). Others in mind: Social origins of self-consciousness. New York, NY, US: Cambridge University Press.

Spaulding, S. (2016). Mind Misreading. *Philosophical Issues*, 26

Schwitzgebel, E., (2010). “Acting contrary to our professed beliefs, or the gulf between occurrent judgment and dispositional belief”, Pacific Philosophical Quarterly, 91: 531–553.

Southgate, V., Senju, A., & Csibra, G., (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science,* 18, 587–592

St Jacques PL, Szpunar KK, Schacter DL. (2017). Shifting visual perspective during retrieval shapes autobiographical memories. *Neuroimage*. 148:103–114.

Tangney, J. P., Wagner, P. E., & Gramzow, R. (1992). Proneness to shame, proneness to guilt, and psychopathology. *Journal of Abnormal Psychology*, 103, 469–478.

Tamir, D. I., Mitchell, J., P., (2010). Neural correlates of anchoring-and-adjustment during mentalizing. *Proc. Natl. Acad. Sci*. USA, 107 (2010), pp. 10827-10832

Tulving, E. (1985). Memory and consciousness. Canadian Psychology/ Psychologie Canadienne, 26, 1–12.

Twenge, J. M., Catanese, K. R., & Baumeister, R. F. (2003). Social exclusion and the deconstructed state: Time-perception, meaninglessness, lethargy, lack of emotion, and self-awareness. *Journal of Personality andSocial Psychology*, 85, 409 – 423.

Velleman, J.D. (2005), Self to Self: Selected Essays, Cambridge: Cambridge University Press.

Velotti, P., Elison, J., & Garofalo, C. (2014). Shame and aggression: Different trajectories and implications. *Aggression and Violent Behavior*, 19, 454–461.

Vorauer, J. D., Hunter, A., Main, K. J., & Roy, S. A. (2000). Meta-stereotype activation: evidence from indirect measures for specific evaluative concerns experienced by members of dominant groups in intergroup interaction. *Journal of Personality and Social Psychology*, 78(4), 690.

Waytz, A., & Mitchell, J. P. (2011). Two mechanisms for simulating other minds: Dissociations between mirroring and self-projection. Current Directions in Psychological Science, 20(3), 197-200.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: representation and the containing function of wrong beliefs in young children's understanding of deception. *Cognition,* 13, 103±128

Williams, M. 2001. In whom we trust: Group membership as an affective context for trust development. *Academy of Management Review*, 26: 377–396.

Wilson, A.E., and Ross, M. (2003). The identity function of autobiographical memory: time is on our side. Memory,11, 137e149.

Whitsett, D., & Kent, S. A. (2003). Cults and families. *Families in Society*, 84(4), 491-502.

Zawidzki, T. (2008), 'The Function of Folk Psychology: Mind reading or mind shaping?', Philosophical Explorations, 11, pp. 193–210.

Chapter 5: The Tension Between the Agential and the Epistemic Roles of Emotion in Moral Deliberation

5.1 Introduction

The relation between reasoning and emotion in moral judgement has been at the forefront of empirical and philosophical debates. Challenges from *dual-process* theories of mind suggest that *System 1* automatic affective reactions drive moral judgment, while *System 2* reasoning merely offers post hoc rationalizations.[21] As a result, causes and reasons for moral judgement are often incongruent (Doris, 2015; Greene, 2008; Haidt, 2001; Prinz, 2009). That is, while people take themselves to be responding to morally relevant reasons for moral castigation of the target, they are often driven by morally irrelevant affective reactions (Cameron et. al., 2013). If affective reactions can bypass moral judgement, their epistemic significance is under threat.

Nevertheless, this argument for incongruence has faced some recent challenges (e.g., May, 2019; Landy and Goodwin, 2015). To start, it is not clear whether System 1 is particularly "affectively laden", while System 2 "affectively neutral". The old "passions versus reason" dichotomy becomes particularly troublesome if we consider that emotions are not necessarily devoid of cognitive elements (e.g., Railton, 2017; May, 2019; Rozin et. al., 1997). If emotions are not strictly System 1 responses, their ability to bypass moral judgment is not obvious (May, 2020, p. 8).

[21] Fiery Cushman, Liane Young, and Joshua Greene, "Multi-System Moral Psychology," in The Moral Psychology Handbook, ed. J. Doris (Oxford: Oxford University Press, 2010), 47–71, 67.

In this work I offer a new argument for incongruence. In particular, I argue that (i) emotions function to prepare us for action, and (ii) this agential role undermines their epistemic significance. As a result, emotions' ability to inform moral judgment is under threat.

The agential nature of emotions is evident in everyday life. After all, in experiencing an emotion, we are not simply making note of features in the world, but also feeling ourselves prepare for dealing with them (e.g., Prinz, 2005; Ledoux 1996; Deonna and Teroni, 2015). For instance, in experiencing an emotion, we experience "readiness to move away, towards or against a given object" (Deonna and Teroni, 2015, p.17).

Nevertheless, while it's easy to notice *bodily preparedness* like readiness to move away, *emotional preparedness* is more elusive to introspection. Specifically, converging empirical findings suggest that we tacitly prepare for dealing with the world by exploiting the *instrumental* role of emotions. For example, athletes tacitly drum up feelings of rage to optimize their work outs, artists drum up feelings of anxiety to heighten creativity, and so on (e.g., Tamir et al., 2011; Butler and Gross, 2009; Netzer et al., 2015; Ma et al., 2018, Leung et al., 2015). Similarly, people tacitly enact episodes of anger to regulate feelings of shame, and episodes of moral disgust to regulate feelings of anxiety (e.g., Pascual-Leone et al., 2013; Navarrete et al., 2004).

Unfortunately, the *instrumental* role of emotions in everyday life undermines their *epistemic* significance in moral deliberation. In particular, while instrumental emotions are not necessarily System 1 responses, their instrumentality helps them to bypass moral judgement. For instance, if feelings of anger work to keep my anxiety at bay, they will appear especially epistemically secure and resistant to reappraisal (i.e., doubting their epistemic significance during moral deliberation will leave my anxiety running rampant) (e.g., Greenberg and Watson 2006; Paivio and Pascual-Leone 2010; Greenberg and Goldman 2008). In short, while people take themselves to be responding to morally relevant reasons for moral castigation of the target, they are often driven by morally irrelevant emotion regulation needs.

I conclude by offering a way of partially guarding against this type of incongruence. In particular, I propose a way to limit the emotion regulation workload of instrumental emotions by cultivating self-diversity. For instance, if both nationalistic pride and athletic pride serve to regulate anxiety in everyday life, re-examining one while leaving the other intact might prove to be more manageable (e.g., Renaud and McConnell, 2002; Rafaeli-Mor and Steinberg, 2002).

5.2 Resisting The Skeptical Challenge

Recent theories of moral evaluation present the following skeptical challenge. Sometimes causes and reasons for moral judgement are incongruent. That is, while we think we are making judgements for morally relevant reasons, we are in fact driven by morally irrelevant affective causes (Doris, 2015; Greene, 2008; Haidt, 2001; Prinz, 2009). In particular, *dual-process* theories of moral cognition suggest that two systems play a role in moral evaluation. While System 1 is unconscious, automatic, and affectively charged; System 2 is slow, conscious, and affectively neutral. Furthermore, System 2 cannot change concurrent automatic affective processes but only offers post hoc rationalizations after the judgment has already been reached.

Disgust reactions provide classic examples of incongruence. In particular, studies suggest that disgust reactions greatly amplify negative moral judgments. For example, drinking a disgusting drink causes people to come up with harsher moral judgements (Eskine et al., 2011). Along the same lines, viewing a disgusting film causes amplified condemnation of purity violations (Horberg et al., 2009). In these cases, the causes and reasons for moral judgement are incongruent. While subjects take themselves to be responding to morally relevant reasons for the moral castigation of the target, they are in fact driven by morally irrelevant disgust reactions (Cameron et. al., 2013).

Nevertheless, some theorists have started to question these findings on empirical and theoretical grounds (e.g., May, 2014; Landy and Goodwin, 2015). In particular, recent meta-analyses have revealed that both the control and manipulation groups rate the morality of the target just about the same (May, 2019, p. 3). While the differences are statistically significant, they may be too small to ground serious concerns about incongruence in everyday life.

Furthermore, while dual-process theories of cognition have offered some insight into the nature of moral evaluation, details about the role of *emotion* and *automaticity* remain highly disputed (e.g., Mugg, 2016). To review, the original skeptical challenge states that System 1 automatic affective reactions drive moral judgment, while System 2 effortful reasoning merely offers post hoc rationalizations.

However, it is not clear whether System 1 is particularly "affectively laden" while System 2 "affectively neutral". Likewise, it is not clear whether affect is particularly "automatic" while reasoning "slow" and "effortful". As a result, affect's ability to bypass moral reasoning is not obvious.

In particular, May argues that *both* affect and reasoning could be somewhat automatic and contain cognitive elements (May, 2020). He summarizes this worry in the following way:

> Like paradigm emotional processing, reasoning can be rapid and relatively inaccessible to consciousness. And emotions, like paradigm reasoning, aid both conscious and unconscious inference, as they provide us with relevant information (Dutton & Aron 1974; Schwarz & Clore 1983), often through gut feelings about which of our many options to take (Damasio 1994) (May, 2020; p. 6).

In other words, both emotions and cognitions could be quick and yet provide relevant information about the world. If emotions offer a way of seeing the world a certain way, they could be a source of knowledge and justified belief. While May does not offer his own theory of emotions, I propose to turn our attention to theorists of emotion who defend the epistemic significance of emotions.

5.2.1 The Perceptual Analogy and the Epistemic Significance of Emotions

Many theorists of emotion argue that emotions are epistemically significant representations of the world (e.g., Deonna 2006; de Sousa 1987; Döring 2003, 2007; Milona 2016; Pelser 2014; Roberts 2013; Tappolet 2000, 2016). In particular, emotions have a representational structure by *presenting* the world as being a certain way. Emotions present the world as being a certain way by allowing us attribute properties to objects. For example, vegetarians' disgust reactions towards meat consumption attribute disgust-worthy properties, (e.g., systematized animal torture), towards the practice of meat consumption. Since the content of the emotion has an object-property structure, it could easily be checked for veridicality (i.e., is the world really the way it is being presented by a given emotion?) If the practice of meat consumption really instantiates the relevant evaluative properties, the vegetarians' disgust reactions could be described as *fitting*. In turn, these features could be cited as justifications for the particular emotion at hand. If emotions are a source of knowledge and justified belief, they are indeed epistemically significant representations of the world.

In order to promote this *epistemological* agenda, many naturalistically inclined philosophers adopt the 'Perceptual Analogy' (e.g., Deonna 2006; de Sousa 1987; Döring 2003, 2007; Milona 2016; Pelser 2014; Roberts 2013; Tappolet 2000, 2016). Since perceiving is a non-conceptually demanding way of representing the word (e.g., one can perceptually represent the ball as round without having the concept of roundness), theorists inclined to grant emotions to conceptually unsophisticated creatures argue that emotions are similar to perceptual experiences of evaluative properties. Much like perceptions, emotions have a vivid phenomenology, appear to be about the world, and at times remain impenetrable to judgement. That is, like perceptions, emotions have distinct phenomenal properties that concern *what it is like* to experience them. Furthermore, like perceptions, emotions have *exteroceptive intentionality* or appear to be directed outward toward particular objects in the world (e.g., Mitchell, 2020). Finally, some perceptual illusions appear to be just as recalcitrant as emotional ones. For example, one can't help but remain frightened of the harmless spider much in the same way one can't simply "unsee" the Muller-Lyer illusion (e.g., Tappolet, 2016). Simply judging the spider to be harmless and the two lines to be identical does not really change the relevant feelings.

In sum, proponents of the perceptual analogy argue that since perceptions are epistemically significant representations of the world, and emotions are similar to perceptions, emotions are also epistemically significant representations of the world.

5.3 The Agential Component of Emotions

Nevertheless, some theorists argue that likening emotions to perceptions fails to account for the distinctive *agential* phenomenology of emotional experiences (Deonna and Teroni, 2015; Kriegel, 2017; Naar, 2020). In particular, emotional episodes feature feelings of *preparedness* or *unpreparedness* for dealing with the world. After all, in experiencing an emotion, you are not simply making note of evaluative features through perception or judgement, but also feeling yourself prepare for dealing with them.

Theorists often characterize these agential components in *bodily* terms. For instance, Deonna and Teroni argue that emotions are experienced with the following feelings of agency:

> …felt readiness to move away, towards or against a given object, to contemplate it, to submit to it, to be attracted by it, to disengage from it or even to suspend any kind of interaction with it (Deonna and Teroni, 2015, p.17).

In contrast, perceptual representations are not necessarily *agential* in nature. For example, I could perceive a slight without getting angry. In fact, anyone who has ever worked a minimum wage job could attest to perceiving multiple slights a day without getting enraged. Likewise, it is reasonable to assume to that people could perceive dangers without feeling afraid. For instance, it is likely that experienced fire fighters notice danger lurking at every corner without losing their cool (Kriegel, 2017).

Could emotion theorists maintain the *epistemological* agenda inspired by the Perceptual Analogy while finding a place for agential phenomenology? After all, while perceptions work to put us in touch with the world, actions carry no accompanying epistemological significance. Naar summarizes this worry in the following way:

> Given that it is arguably not in the nature of action in general to put us in touch with aspects of the world, i.e. actions in general are not epistemically special (at least the way perception is), the claim that emotions play some special kind of epistemological role becomes something we have no particular reason to hold (Naar, 2020, p. 4-5).

In trying to account for the agential nature of emotions while retaining the epistemological agenda, *motivational attitudinal* theories of emotion argue that emotions are not simply perceptions of evaluative properties, but ways or modes of relating to these properties. Specifically, "emotions are bodily experiences of being disposed or tending to act in a differentiated way vis-à-vis a given object or event" (Deonna and Teroni, 2015, p.16).

Just like we can have different attitudes towards the same content, Deonna and Teroni argue that we could have different emotions towards the same content. For example, just like you can doubt that it's raining and I can be sure of it; you can feel mirth at the funny joke, while I can recognize its merit without feeling mirth after hearing it one too many times.

Notably, despite acknowledging the agential nature of emotions, Deonna and Teroni still uphold some key tenets of the epistemological agenda. While emotions do prepare us for the world, they only do so *in virtue of* correctly representing it. Just like the proponents of the Perceptual Analogy, Deonna and Teroni argue that affective attitudes are object directed and serve to provide important information about the world. In particular, emotions are "evaluative attitudes towards intentional contents provided by other mental states—their *cognitive bases*, such as judgements, imaginations, and perceptions" (Mitchell, 2020, p.1). For example, fear triggers bodily readiness to move away from an object that is represented *as* dangerous by perception or judgment.

5.3.1 How Do Emotions Prepare Us for Dealing with the World?

According to the attitudinal motivational theory, emotions prepare us for dealing with the world in a bodily way. However, since bodily readiness only happens in virtue of correct representation, we can try to account for the agential component of emotions without compromising the epistemological agenda.

In this section I argue that emotions' agential component extends beyond preparing the body for action. Oftentimes, emotions prepare us for dealing with the world by activating other emotions with *instrumental* value. In turn, this type of preparedness indeed compromises the epistemological agenda. While instrumental emotions might not necessarily be System 1 responses, their instrumentality allows them to bypass moral judgement.

While bodily preparedness is available to introspection, real world emotional experiences invite further types of preparation not readily available to introspection. In particular, it is often the case that emotions prepare us for dealing with the world by activating other emotions. Many researchers believes that this interplay compromises the very distinction between emotion *generation* and emotion *regulation* (e.g., Frijda, 2007; Mesquita, 2003; Crone, 2009; Baker et al., 2004). Mesquite and Frida argue that "emotional events in real life involve multiple emotions that regulate each other (Mesquite and Frida, 2011, p. 783).

Converging empirical findings suggest that we tacitly prepare for dealing with the world by exploiting the *instrumental* role of emotions. For example, athletes habitually drum up feelings of rage to optimize their work outs, artists drum up feelings of anxiety to heighten creativity, and so on (e.g., Tamir et al., 2011; Butler and Gross, 2009; Netzer et al., 2015; Ma et al., 2018, Leung et al., 2015).

In order to flesh out the nature of instrumental emotions, Tamir et al., (2014) propose the *expectancy-value* model of emotion regulation. According to this model, people prefer to enact emotions they expect to be useful to them. For example, people are motivated to increase feelings of anger in situations they expect anger to be useful, "even in the context that bears little relevance to anger-related appraisals or goals" (Tamir et al., 2014, p.1).

The authors tested people's preference for inducing emotions they believed to be useful in particular circumstances, even if they previously believed these circumstances to not merit the relevant emotions. They found that participants preferred to read anxiety-inducing material if they believed anxiety would be useful to them, even if they previously indicated that the circumstances did not merit anxiety. Likewise, participants preferred to read anger-inducing material if they believed that anger would be useful to them, even if they previously indicated that the circumstances do not merit anger.

The authors theorize that people often enact instrumental emotions without explicit awareness of doing so. Tamir et al. summarize their findings in the following way:

> Our findings demonstrate that the expected usefulness of emotions can be manipulated relatively easily, even when they contradict previous evaluative beliefs… Indeed, such expectancies may not require conscious awareness. The proposed model, therefore, can potentially inform the understanding of both adaptive and maladaptive emotion regulation. To the extent that people expect anxiety or anger to be useful in particular contexts, they might find themselves trying to increase or sustain such feelings, *without necessarily knowing why* (Tamir et al., 2014, p. 13*, my emphasis*).

If people are unaware of enacting instrumental emotions as means to an end, they would have little basis for distinguishing "genuine" from instrumental emotions in everyday life. After all, once enacted, both genuine and instrumental have a vivid phenomenology and appear to be about the world.

Pascual-Leone et al., (2012) offer a detailed model of our tacit use of instrumental emotions in emotions regulation. The authors argue that emotions are often elicited to regulate other less palatable emotions. This pattern of regulation becomes *habitual* through negative and positive reinforcement. For example, feelings of anger regulate feelings of shame by displacing painful experiences (negative reinforcement), as well as by increasing positive feelings of preparedness or control (positive reinforcement). While feelings of anger feel like genuine emotions, they are often brought about to regulate uncomfortable emotions like shame.

The instrumental role of emotions in emotion regulation is nicely demonstrated in the following study. Prior to asking participants to report their feelings about ingroup rule violations, Navarrete et al. primed one group of participants with vignettes of being burglarized, and another with neutral content (Navarrete et al., 2004). Those primed with contemplating threatening events like being burglarized experienced heightened feelings of pride towards ingroup rules and derogation towards perceived violators of these rules.

These results fit the model outlined above. It appears that feelings of outgroup derogation play a regulating role to less palatable feelings of fear and anxiety. Feelings of outgroup derogation displace painful feelings of anxiety (negative reinforcement) and increase positive feelings of preparedness or control (positive reinforcement).

There are reasons to suspect that the instrumental role of emotions in everyday life thwarts the epistemological agenda and the perceptual analogy. After all, unlike perceptions, emotions are not triggered in response to evaluative properties, but in response to the emotion regulation needs of the subject. Going back to the example above, it appears that outgroup anger is not triggered in response to the evaluative properties of slights, but in response to the emotion regulation needs of the subject.

Nevertheless, a proponent of the perceptual analogy might respond by alluding to the response-dependent nature of evaluative properties. Specifically, there may be something in the individual values of the subject who regularly responds to feelings of vulnerability with anger. After all, a person who responses with outgroup disgust after being primed with feelings of vulnerability might interpret his phenomenology of vulnerability in terms of bigoted values (e.g., "This used to be a safe country, but now that its borders are open, it's gone to the dogs."). So, while his anger might seem to be unfitting from an objective point of view, it might in fact be an appropriate response in light of his own values.

Unfortunately, this objection does not save the epistemological agenda. While the epistemological agenda may tolerate value-depended emotional responses, it still requires the ability to form beliefs that could be questioned for justification. In other words, to avoid the threat of incongruence, the subject enacting emotions of outgroup derogation should be able to interrogate his emotional responses for moral relevance. However, if he is unable to distinguish between emotions enacted for instrumental ends and "genuine" emotions, he cannot interrogate them for moral relevance.

In the next section, I propose to take a closer look at why instrumental emotions thwart our ability to interrogate them for moral relevance.

5.4 The Price of Instrumental Emotions

So far, we have seen that the agential component of emotions extends well beyond bodily preparedness. Converging empirical findings suggest that we often tacitly prepare for dealing with the world by enacting instrumental emotions.

In this section I argue that this agential component undermines emotions' epistemic significance in moral deliberation. In particular, while instrumental emotions might not necessarily be System 1 responses, their instrumentality allows them to bypass moral judgement. Specifically, instrumental emotions appear to be both (i) epistemically secure, and (ii) resistant to reappraisal. Hence, if one has no reason to suspect that his emotions might be misleading and, upon closer inspection, further maintains that they are not misleading, the epistemological agenda is under threat.

Why do instrumental emotions seem secure and resistant to reappraisal? Once emotions assume instrumental roles in regulating other emotions, we start to rely on them for regular relief. Some psychologists refer to this process of reliance as potentially *addictive* (e.g., Korman 2005;

Linehan 1993; Tangney et al. 1992). For example, if my feelings of anger work to keep my anxiety at bay, re-examining them will leave my anxiety running rampant (Pascual-Leone et al., 2012).

Difficulties in reappraising instrumental emotions are compounded by the fact that emotion regulation strategies solidify into *unconscious habits*. Pascual-Leone et al., (2012) summarize this model in the following way:

> A surge of secondary emotion dispels vulnerability and seemingly empowers the individual. Because of the reduction in distress (a negative reinforcer) and the experience of power (a positive reinforcer) the strategy of experiencing and expressing anger can become a conditioned response cued by experiences of shame (Korman, 2005).

In turn, processes of reinforcement are associated with feelings of inappropriate epistemic certainty. Specifically, reward expectancy usually leads to *over-generalization* or the process of classifying distinct categories of situations as categorically similar (Redish et. al., 2007). For example, reward associated with gambling causes gamblers to feel especially epistemically secure in misclassifying "winning" situations as categorically distinct from "losing" situations, even though both situations are instances of "gambling" situations. Similarly reward expectancy from using instrumental emotions would cause us to misclassify all properties of the situation as relevant to the category of "anger-worthy" evaluative properties, even in the presence of both "anger-worthy" and "non-anger worthy" evaluative properties.

While instrumental emotions appear to be epistemically secure, they are also resistant to reappraisal. In particular, using instrumental emotions for emotion regulation compromises our cognitive flexibility (i.e., the ability to re-examine and inhibit impulsive thoughts and behavior) (e.g., Kashdan and Rottenberg, 2010). Arguably, cognitive flexibility allows us to interrogate evaluative properties for moral relevance.

Specifically, the use of instrumental emotions for emotion regulation qualifies as a *response focused* strategy of emotion regulation. Response focused strategies are built around the desire to avoid the phenomenological impact of a given emotion. Focusing on avoiding the experience of anxiety altogether by enacting another emotion is a type of response suppression. Response suppression strategies are in turn associated with a lack of cognitive flexibility (Kashdan and Rottenberg, 2010; Hollenstein et al. 2013; Bonanno and Burton, 2014). In particular, Szczygieł and Maruszewski (2015) found that expressive suppression compromises performance on cognitively demanding tasks. In other words, avoiding certain emotions by tacitly enacting instrumental emotions compromises cognitive flexibility and the ability to interrogate emotional responses for moral relevance.

5.5 Cultivating Self-Complexity and Cognitive Flexibility

Now that we have outlined the epistemic threats posed by instrumental emotions, I conclude by offering a way of partially guarding against these threats. One reason why instrumental emotions are resistant to reappraisal is their emotion regulating workload. For example, if anger regulates anxiety, doubting anger's epistemic significance during moral deliberation leaves my anxiety running rampant (e.g., Greenberg and Watson 2006; Paivio and Pascual-Leone 2010; Greenberg and Goldman 2008). In this section I propose a way to increase cognitive flexibility by limiting the emotion regulation workload of instrumental emotions.

Proponents of the expectancy value model of emotions theorize that patterns of instrumental emotions solidify into unconscious habits, which in turn solidify into instrumental values (Korman, 2005). Instrumental values provide easily accessible societal scripts for enacting instrumental emotions.[22] For example, if instrumental outgroup anger has served a reliable emotion regulating role in Jim's mental economy, the modern world provides a variety of readily accessible societal scripts that secure a regular flow of outgroup anger (Griffiths and Scarantino, 2005, p. 6).

How can we learn to limit the emotion regulation workload of a given set of instrumental emotions associated with particular values? The answer may be found in cultivating a diverse set of values. Social psychologists define self-diversity in terms of cognitive differentiation between one's values or self-concepts. One's self-conceptions are said to be more differentiated if they are less cognitively associated with one another. For example, my self-conception as a "good educator" is more cognitively associated with my self-conception as a "mentor" than with my self-perception as an "artist" (Linville, 1985; Renaud and McConnell, 2002; Niedenthal et al.,1992). Likewise, while my friend's self-conception as a "good mother" is greatly cognitively associated to her self-perception as a "good wife", this might not be the case for other women. Renaud and McConnell characterize the notion of differentiation in the following way:

> Differences in self-complexity are based on both the number of self-aspects and the degree of redundancy among the traits describing those self-aspects. Greater self-complexity is revealed by a greater number of self-aspects that are described by traits that are less redundant with, and thus are more independent of, one another. Lower self-complexity, on the other hand, is revealed by fewer self-aspects that are described by more redundant traits and thus are more interrelated with one another… this conceptualization of self-concept organization is concerned with the relative amount of association among the traits describing aspects of one's self (Renaud and McConnell, 2002, p. 80).

[22] Societal scripts are publicly accessible recipes for thinking, feeling, and behaving (e.g., Mesquita and Frijda 1992; Averill 1990).

How does cultivating self-diversity limit the emotion regulation workload of a given set of instrumental emotions? If my self-conception as a great athlete is under threat, my self-conception as an artist remains intact enough to offer temporary emotion regulation benefits (Renaud and McConnell, 2002). Social psychologists have shown that people with greatly differentiated self-perceptions have better cognitive flexibility than those with less differentiated self-perceptions (for reviews see Rafaeli-Mor and Steinberg, 2002; Showers and Zeigler-Hill, 2012). Researchers theorize that having greatly differentiated self-conceptions helps prevent "emotional spillover" during stressful events particular to one self-conception.

We can apply this model toward cultivating cognitive flexibility during moral deliberation. If outgroup anger regulates anxiety, doubting anger's epistemic significance during moral deliberation leaves my anxiety running rampant (e.g., Greenberg and Watson 2006; Paivio and Pascual-Leone 2010; Greenberg and Goldman 2008). However, if another, unconnected, set of instrumental emotions *also* work to regulate anxiety, doubting anger's epistemic significance during moral deliberation might prove to be more manageable.

For example, say a high-school athlete raised in the 1950's suburbia is asked in interrogate his outgroup disgust for moral relevance (e.g., someone asked him to reconsider the moral relevance of his disgust towards same sex marriage). Could he "step back" and re-examine his affective reactions toward same-sex marriage? After all, his emotional reactions could fail to be fitting. Unfortunately, if he simply tried to set his initial affective reactions aside, he would become vulnerable to emotions his instrumental emotions function to regulate. However, if his values portfolio contains other values unrelated to his masculine athleticism, he could temporarily use instrumental emotions related to the to the other set of values for emotions regulation. In this way, he could prevent the “emotional spillover” from being threatened in one set of instrumental values and find relief in another set of instrumental values.

In sum, while forming instrumental values might be inevitable, perhaps there is a way to limit their toll on cognitive flexibility in everyday life.

5.6 Conclusion

In this work I have offered a new argument for incongruence. I have sketched a novel way in which emotions prepare us for dealing with the world and its toll on their epistemic significance. I have outlined converging empirical findings suggesting that emotions tacitly prepare for dealing with the world by activating other instrumental emotions. Unfortunately, while instrumental emotions provide emotional relief, they also invite a false sense of epistemic security. As a result, while people take themselves to be responding to morally relevant reasons for moral castigation of the target, they are often driven by morally irrelevant emotion regulation needs. I concluded by offering a way of partially guarding against this type of incongruence. In particular, I outlined a way to limit the emotion regulation workload of instrumental emotions.

References

Averill, J. R. (1990). Inner feelings, works of the flesh, the beast within, diseases of the mind, driving force, and putting on a show: Six metaphors of emotion and their theoretical extensions. In D. E. Leary (Ed.), Metaphors in the history of psychology (pp. 104-132). New York: Cambridge University Press.

Baker, R., Holloway, J., Thomas, P. W., Thomas, S., & Owens, M. (2004). Emotional processing and panic. Behaviour Research and Therapy, 42, 1271–1287.

Butler, E. A., & Gross, J. J. (2009). Emotion and emotion regulation: Integrating individual and social levels of analysis. Emotion Review, 1, 86–87. doi:10.1177/175407390809913

Bonanno GA, Burton CL. Regulatory flexibility: An individual differences perspective on coping and emotion regulation. Perspectives on Psychological Science. 2013;8(6):591–612. doi: 10.1177/1745691613504116.

Cameron, C. D., Payne, B. K., & Doris, J. M. (2013). Morality in high definition: Emotion differentiation calibrates the influence of incidental disgust on moral judgments. Journal of Experimental Social Psychology, 49, 719–725.

Crone EA. Executive functions in adolescence: Inferences from brain and behavior. Developmental Science. 2009;12:825–830

Damasio, A. R. (1994). Descartes error: Emotion, rationality and the human brain. New York: Putnam (Grosset Books).

Deonna, J. A. (2006), 'Emotion, Perception and Perspective', Dialectica 60,1, pp. 29–46.

Deonna, J., & Teroni, F. (2015). Emotions as attitudes. Dialectica, 69(3), 293–311.

De Sousa, R. (1987), The Rationality of Emotions, Cambridge, MA: MIT Press.

Döring, S. (2003), 'Explaining Action by Emotion', The Philosophical Quarterly 211, pp. 214–30. (2007). Seeing what to do: Affective perception and rational motivation. Dialectica, 61, 363–394.

Doris, J. (2015) Talking to Ourselves: Reflection, Ignorance, and Agency. Oxford University Press.

Eskine, K., Cacinik, N. A., & Prinz, J. J. (2011). A bad taste in the mouth: Gustatory disgust influences moral judgments. Psychological Science, 22, 295-299.

Frijda, N. (2007). The laws of emotion. Mahwah, NJ: Lawrence Erlbaum.

Greenberg, L. S., & Goldman, R. N. (2008). Emotion-focused couples therapy: The dynamics of emotion, love, and power. Washington: APA.

Greenberg, L. S., & Watson, J. C. (2006). Emotion-focused therapy for depression. Washington, DC: American Psychological Association.

Greene, J.D. (2008). The secret joke of Kant's soul. In W. SinnottArmstrong (Ed.), Moral psychology: Vol. 2. The cognitive science of morality. Cambridge, MA: MIT Press

Griffiths, P., & Scarantino, A. (2009). Emotions in the wild: the situated perspective on emotion. In P. Robbins &

M. Aydede (Eds.), The Cambridge handbook of situated cognition (pp. 437–453). Cambridge: Cambridge University Press.

Haidt, J., & Hersh, M. (2001). Sexual morality: The cultures and emotions of conservatives and liberals. Journal of Applied Social Psychology, 31, 191-221.

Hollenstein, T., Lichtwarck-Aschoff, A., & Potworowski, G. (2013). A model of socioemotional flexibility at three time scales. Emotion Review. Advance online publication. doi:10.1177/1754073913484181.

Horberg, E. J., Oveis, C., Keltner, D., & Cohen, A. B. (2009). Disgust and the moralization of purity. Journal of Personality and Social Psychology, 97(6), 963–976.

Kashdan, T. B., and Rottenberg, J. Psychological flexibility as a fundamental aspect of health. Clinical Psychology Review. doi:10.1016/j.cpr.2010.03.001.

Korman, L. M. (2005). Treating anger and addictions concurrently. In W. J. Skinner (Ed.), Treating concurrent disorders: A guide for counselors (pp. 215–234). Toronto: Centre for Addiction and Mental Health.

Kriegel, Uriah (2017). "Reductive Representationalism and Emotional Phenomenology". en. In: Midwest Studies In Philosophy 41.1, pp. 41–59. issn: 1475-4975. doi: 10.1111/misp.12072.

Landy, J. F., & Goodwin, G. P. (2015). Does incidental disgust amplify moral judgment? A meta-analytic review of experimental evidence. Perspectives on Psychological Science, 10, 518-536.

LeDoux, J., (1996). Emotional networks and motor control: a fearful view. Prog. Brain Res. 107, 437 – 446.

Leung, A. K., Liou, S., Qiu, L., Kwan, L. Y., Chiu, C., & Yong, J. C. (2014). The role of instrumental emotion regulation in the emotions–creativity link: How worries render individuals with high neuroticism more creative. Emotion, 14(5), 846-856. doi:10.1037/a0036965.

Linehan, M. M. (1993b). Skills training manual for treating borderline personality disorder. New York: Guilford.

Linville PW. Self-complexity as a cognitive buffer against stress-related illness and depression. Journal of Personality and Social Psychology. 1987;52:663–676.

Ma X, Tamir M, & Miyamoto Y (2018). A socio-cultural instrumental approach to emotion regulation: Culture and the regulation of positive emotions. Emotion, 18, 138–152. doi: 10.1037/0000315.

May, J. 2018. Regard for Reason in the Moral Mind. Oxford University Press. ———. 2019. “Précis of ‘Regard for Reason in the Moral Mind.’” Behavioral and Brain Sciences 42 (e146): 1–60.

Mesquita, B., & Frijda, N. H. (2011). An emotion perspective on emotion regulation. Cognition and Emotion, 25, 782–784. doi:10.1080/02699931.2011.586824.

Milona, M. (2016). Taking the Perceptual Analogy Seriously. Ethical Theory and Moral Practice, 19(4), 897–915.

Mitchell, J. (2020). The attitudinal opacity of emotional experience. Philosophical Quarterly, 70(280), 524–546.

Mugg, J. (2016): “The dual-process turn: How recent defenses of dual-process theories of reasoning fail” Philosophical Psychology 29 (2), 300–309.

Naar, H. (2020). Emotion: More like action than perception. Erkenntnis.

Navarette, C. D., Kurzban, R., Fessler, D. M. T., & Kirkpatrick, L. A. (2004). Anxiety and intergroup bias: Terror management or coalitional psychology. Group Processes and Intergroup Relations, 7, 370–397.

Netzer L, Van Kleef GA, Tamir M. Interpersonal instrumental emotion regulation. Journal of Experimental Social Psychology. 2015;58:124–135.

Niedenthal, P. M., Setterlund, M. B., & Wherry, M. B. (1992). Possible self-complexity and affective reactions to goal-relevant evaluation. Journal of Personality and Social Psychology, 63, 5-16.

Paivio, S. C., & Pascual-Leone, A. (2010). Emotion focused therapy for complex trauma: An integrative approach. Washington, DC: American Psychological Association.

Pascual-Leone, A., & Paivio, S. C. (2013). Emotion-focused therapy for anger in complex trauma. In E. Fernandez (Ed.), Treatments for anger in specific populations. Theory, Application and Outcome (pp. 33-51). New York: Oxford University Press.

Prinz, J. (2004). Gut reactions. A perceptual theory of the emotions. New York, NY: Oxford University Press.

NA. (2009). ‘Is Consciousness Embodied?’, in P. Robbins and. M. Aydede (eds) Cambridge Handbook of Situated Cognition. Cambridge: Cambridge University Press, pp. 419-436.

Rafaeli-Mor, E., & Steinberg, J. (2002). Self-complexity and well-being: A review and research synthesis. Personality and Social Psychology Review, 6, 31–58.

Railton, P. (2017). At the core of our capacity to act for a reason: The affective system and dynamic model-based learning and control. Emotion Review.

Redish AD, et al. A unified framework for addiction: vulnerabilities in the decision process. Behav. Brain Sci. 2008;31:415–437.

Renaud, J. M., & McConnell, A. R. (2002). Organization of the self-concept and the suppression of self-relevant thoughts. Journal of Experimental Social Psychology, 38, 79-86.

Rozin, P., Haidt, J., McCauley, C. R., Dunlop, L., & Ashmore, M. (1999). Individual differences in disgust sensitivity: Journal of Research in Personality, 33, 330-351.

Szczygieł D, & Maruszewski T (2015). Why expressive suppression does not pay? Cognitive costs of negative emotion suppression: The mediating role of subjective tense-arousal. Polish Psychological Bulletin, 46(3), 336–349. 10.1515/ppb-2015-0041.

Tappolet, C. (2000). Truth pluralism and many-valued logics: a reply to Beall. Philosophical Quarterly, 50: 382–38.

NA. (2016) Emotions, Values, and Agency. Oxford: Oxford University Press.

Tamir, M. (2011). The maturing field of emotion regulation. Emotion Review, 3, 3–7.

Tangney, J. P., Wagner, P., Fletcher, C., & Gramzow, R. (1992). Shamed into anger? The relation of shame and guilt to anger and self-reported aggression. Journal of Personality and Social Psychology, 62, 669–675.

[i] Bayne (2010) and others write that "the split-brain procedure has surprisingly little impact on cognitive function in everyday life. Split-brain patients can drive, hold down jobs, and carry out routine day to day tasks. Early researchers remarked on their "social ordinariness," and were baffled by their inability to detect any cognitive impairments arising from the operation.
Dahlia W. Zaidel, "AView of the World from a Split-Brain Perspective," in E.M.R. Critchley, ed., The Neurological Boundaries of Reality (Northvale, NJ: Aronson, 1995), pp. 161–74; S.M. Fergusen et al., "Neuropsychiatric Observation on Behavioral Consequences of Corpus Callosum Section for Seizure Control," in A.G. Reeves, ed., Epilepsy and the Corpus Callosum (New York: Plenum, 1985), pp. 501–14. But see also Victor Mark, "Conflicting Communicative Behavior in a Split-Brain Patient: Support for Dual Consciousness," in S.R. Hameroff et al., eds., Towards a Science of Consciousness (Cambridge: MIT, 1996), pp. 189–96.

[ii] Sperry, Vogel, and Bogen (1967) The Syndorme of hemisphere deconnection. Dahlia W. Zaidel, "AView of the World from a Split-Brain Perspective," in E.M.R. Critchley, ed., The Neurological Boundaries of Reality (Northvale, NJ: Aronson, 1995).

[iii] Schechter (2018)

[iv] Left hemisphere is in control of speech, whereas the right hemisphere controls the left hand. Carlson, (2004). Physiology of Behavior, 6th Ed. The received view has been that visual information projected to the right visual field cannot be verbally reported, and visual information projected to the LVF is unavailable for behavior involving the right hand (Bayne, 2010). However, recent findings by Pinto et. al., (2017) show that split-brain patients can indeed respond accurately to stimuli appearing *anywhere* in the visual field using *any* response modality (e.g. using speech, right, and left hand). I save the discussion of these results for later in the paper.

[v] I use the terms "subjective perspective" and "point of view" interchangeably.

[vi] At the very low level, the limits of an action event are delineated by affect (Muhle-Karbe and Krebs, 2012). According to the theory of event coding (Hommel et al., 2001), various perceptual elements are part of the same event if they are represented in the same event code. This theory explains interactions between products of perceptual processes and the first steps of action planning (Eder and Klauer, 2007). Theory of event coding states that perceived features of objects and planned features of motor actions are cognitively represented through structurally identical "event codes". Because of this common code, stimulus and action features can prime one another (Becker's et. al., 2002; Elsner and Hommel, 2001). The main outcome of this binding is action-effect blindness. That is, codes that are already in use cannot be accessed for further use. The common code assumption or the fact that coding of percepts and actions relies on identical format of representations is supported by findings on selective impairments. Using the same model of code overlap, researchers proposed the affective coding hypothesis to demonstrate how event files are organized by affect. Specifically, affectively charged action plans (e.g. saying ''good'' or ''bad'') impair simultaneous evaluations of stimuli with the same valence (Eder and Klauer, 2007) Furthermore, the single feature valence code irrevocably activates the entire event file (Hommel & Musseler, 2006; Musseler & Hommel, 1997a, 1997b)". Similarly, Eder et al. (2012) demonstrated that preparing a button press that signals the affective value of a picture delayed the performance of an affectively congruent approach and avoidance movement simply because that concurrent processes was utilizing the same affective code.

[vii] Since some emotional states are cognitively penetrable, this seems like a reasonable hypothesis. Karl Lange and William James have theorized that bodily changes associated with emotional episodes precede the feeling of emotion itself. They explain examples of emotional episodes such as fear by pointing to the bodily accompaniments of such episodes (e.g. sweaty palms, increased heart rate, dry mouth). They hypothesized that emotions simply are neocortical "readouts" of bodily autonomic arousals (Cannon, 1927)

[viii] Here I replace "percepts" used in the original quote with "emotions".

[ix] In particular, while Bayne and Chalmers argue that enjoying a unified phenomenal perspective entails experiencing all phenomenal states as *merely co-conscious,* I argue that enjoying a unified phenomenal perspective entails experiencing aspects of your phenomenal field in orders of perceived importance and perceived changeability. Since affective states allow us to experience aspects of our phenomenal field via orders of perceived importance and orders of perceived changeability, affective phenomenology is not just another phenomenal state in the unstructured phenomenal field. Since affect plays a special role in ensuring a unified subjective perspective, unified affect in the split-brain is enough to ensure a unified subjective perspective on the world.

[x] Transcranial magnetic stimulation (TMS) is a type of targeted brain stimulation induced by a changing magnetic field (e.g. Killgore and Yurgelun-Todd, 2004)