FEELING GOOD: THE ROLE OF FEELINGS IN THE MAKING OF MORAL JUDGMENT

KOH JIAN MIN, JEREMIAS

(B.A. (Hons.), NUS)

A THESIS SUBMITTED FOR THE DEGREE OF MASTER OF ARTS IN PHILOSOPHY

DEPARTMENT OF PHILOSOPHY

NATIONAL UNIVERSITY OF SINGAPORE

2019

SUPERVISOR

ASSOC. PROF. NEILADRI SINHABABU

EXAMINERS

ASSISTANT. PROF. ZACHARY BARNETT

ASSISTANT. PROF. ABELARD PODGORSKI

DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all sources of information which have been used in this thesis.

This thesis has also not been submitted for any degree in any university previously.

_____

Koh Jian Min, Jeremias

14 August 2019

ACKNOWLEDGEMENTS

I would like to thank my supervisor, **Neil Sinhababu**, whose work is the inspiration for practically everything expressed in this thesis, **Abelard Podgorski** and **Zachary Barnett**, my examiners, for their instructive comments, **Prema Alexander**, my favorite person, who actually read this whole thing and encouraged me all the way, **Andrew**, **John**, **Gabriele**, for hanging around in school and playing videogames with me, and last but not least, my **family**, for their love and support.

Table of Contents

What needs to be said, finally, to assuage the embarrassment of the emotionally aroused intellectual, is that there is no necessary connection between emotionalism and irrationality. A lie may be calmly uttered, and a truth may be charged with emotion. Emotion can be used to make more rational decisions, if by that we mean decisions based on greater knowledge, for greater knowledge involves not only extension but intensity. Who "knows" more about slavery—the man who has in his head all the available information (how many Negroes are enslaved, how much money is spent by the plantation for their upkeep, how many run away, how many revolt, how many are whipped and how many are given special privileges) and calmly goes about his business, or the man who has less data, but is moved by the book (Harriet Beecher Stowe's) or by an orator (Wendell Phillips) to feel the reality of slavery so intensely that he will set up a station on the underground railroad? Rationality is limited by time, space, and status, which intervene between the individual and the truth. Emotion can liberate it.

(Howard Zinn, “Abolitionists, Freedom Riders and the Tactics of Agitation”)

Introduction

Theories of moral judgement and cognition can generally be divided into two categories; those belonging to the rationalist tradition and those of the sentimentalist tradition. Rationalists hold that moral judgement is fundamentally derived from our rational capacities (Kennett 2006) and rationalism can be construed as the thesis that moral judgement is essentially "the culmination of a process of reasoning" (Maibom 2010: 999). According to the rationalist, while emotions may influence moral cognition, they are not essential for making a judgement distinctively moral (May 2018).

On the other hand, sentimentalists typically give moral emotions a constitutive role in moral judgement (Prinz 2007), therefore holding that emotions are essential to moral judgement and that moral judgement is grounded in affective response. Thus, while rationalists might agree with sentimentalists that emotions are usually involved in human moral cognition, and sentimentalists with rationalists that reason and reflection are important in moral judgement, what they disagree on is the role of moral feelings in the making of moral judgement.

Accordingly, this thesis focuses on the question of whether moral feelings are necessary to the making of moral judgments. This is an important question and the answer one gives has more interesting implications than one might initially expect. I will argue that an experientialist account of moral concepts, on which moral judgments are beliefs about objective facts represented by moral feelings, provides the best naturalistic answer to the question. To make my point, I anchor my arguments in a series of comparisons between the experientialist account and its rivals, and how they handle metaethical puzzles to do with

the moral status of psychopaths, moral twin earth, the nature of moral motivation, and issues regarding the possibility of moral knowledge and virtue.

There are four chapters in this thesis. In chapter 1, I will provide a sketch of the experientialist position as formulated by Neil Sinhababu (2017) and explain how it relates to the Humean theory of motivation he favors. In chapter 2, I will introduce and discuss several responses to psychopathy from both the rationalist and sentimentalist camp. Chapter 3 will discuss experientialism and its provision of a novel solution to the problem of moral twin Earth. In chapter 4, I argue that rationalism appears committed to a picture of moral motivation that seems implausible from the point of view of human cognition, and address rationalist worries regarding how adopting sentimentalism would result in pessimism about ordinary moral thought and action.

Chapter 1: Experientialism and the Humean Theory of Motivation

Experientialism is a view of moral concepts recently proposed by Neil Sinhababu (2017) in his book 'Humean Nature'. Its central thesis is that "moral concepts apply to whatever accurate moral feelings objectively represent" (70). For example, an experientialist analysis of the moral concept 'good' would be something along the lines of the "states of affairs objectively represented by accurate hope and delight", and the moral concept 'bad' the "states of affairs objectively represented by accurate horror and sorrow" (ibid).

While it is possible to formulate a more ecumenical version of experientialism by omitting notions of accuracy (to accommodate non-cognitivism) and objectivity (to accommodate relativism) from its thesis, I will follow Sinhababu in articulating and defending an experientialist view that is both cognitivist and realist[1]. This commitment to cognitivism and moral realism distinguishes experientialism from sentimentalists like Prinz (2007), who supports moral relativism, and Kumar (2015) who thinks that moral judgments are hybrid states of belief and emotion. These differences between experientialism and other sentimentalist views will be discussed in greater detail later on in this thesis. At this point, it will suffice to note that experientialism takes moral judgement to be belief alone (rather than belief plus emotion), and holds that moral judgements are beliefs about objective facts.

[1] It should be noted that experientialism, strictly speaking, does not commit us to either realism or anti-realism. An error theorist might find experientialism persuasive, and regard our moral feelings as representing things that are not actually there. (Sinhababu 2017: 71)

Since Sinhababu's experientialism is suggested by[2] and closely related to his emotional perception model, it is worth spelling this model out at some length. According to the emotional perception model, moral judgements are beliefs "typically caused by feelings about actions, people, and states of affairs" (Sinhababu 2017: 57). The emotional dispositions that cause these feelings also contain desires which cause and motivate actions in accordance with the respective moral judgement. For example, guilt about lying to a friend causes belief that lying to a friend is wrong, which in turn involves aversion to having lied to a friend and lying to friends in the future, while moral admiration causes belief that a person is virtuous, which in turn involves a desire to emulate or help that person achieve their goals (ibid).

This view of moral judgment provides an externalist solution to a metaethical trilemma put forward by Michael Smith (1994). These are the propositions that Smith's trilemma consists of: (1) Moral judgements are beliefs (i.e. cognitivism), (2) Moral judgements can produce their own motivational force (i.e. motivational internalism), and (3) Beliefs alone can't motivate action (i.e. the Humean Theory). Many philosophers find each of these propositions attractive, but the truth of all three propositions means the impossibility of moral judgement, so at least one of them has to be false[3]. With the emotional perception model, we can solve the trilemma by accepting cognitivism and the Humean theory, and by

[2] That said, one might be an experientialist without subscribing to the emotional perception model. As will be discussed, the emotional perception model is committed to a Humeanism about moral motivation. Experientialism does not share this commitment. One might think that beliefs alone can motivate without the help of desires, and still hold that moral concepts apply to whatever moral feelings accurately represent.

[3] Smith's own solution to the trilemma is to accept "the anti-Humean descriptive psychological claim that believing its rational to do something can, by reasoning, generate a desire to do it" (Sinhababu 2017: 11). This is an anti-Humean claim because it allows for beliefs to motivate action (albeit in a roundabout away) by allowing beliefs to generate desires. By solving the trilemma in this way, Smith (who is ostensibly a Humean), betrays Hume by allowing for reason to create passions, and takes away from what make Humean views metaethically interesting.

rejecting internalism, while at the same time providing an explanation as to why moral judgment is so strongly correlated with motivation even though these judgements themselves cannot produce their own motivational force (Sinhababu 2017: 60).

The emotional perception model explains the tight correlation between moral judgement and motivation by placing desire at the source of both. The model is an externalist one as it does not allow for moral belief to produce motivational force by itself. Rather, it is the same emotional disposition that causes both the judgement as well as the motivation. This also explains why moral motivation so often automatically accompanies our moral judgements. If our moral judgements and motivations are rooted in the same emotional dispositions, then it's no surprise that they typically come together.

The model is also cognitivist, treating moral judgements as beliefs about objective moral facts. The model does not commit to the existence of these objective facts, but it requires that if moral facts exist, that they must be objective. An error theorist like John Mackie (1977) can accept Sinhababu's model, while holding that moral facts do not exist, and that our beliefs about them are largely mistaken. Treating moral facts as objective provides an explanation for how people from different cultures can engage in genuine disagreements about morality, instead of simply talk past each other. It also allows us to avoid the conceptual, semantic and logical problems faced by subjectivism and non-cognitivism (Sinhababu 2017: 60). While I will not give a detailed account[4] of these problems here, some of them will be discussed in the chapter on moral twin earth.

[4] For more on arguments against subjectivism, see Shafer-Landau (2003). For arguments against non-cognitivism, see Schroeder (2008).

The Humean Theory of Motivation

The emotional perception model supports Sinhababu's Humean theory of motivation, which comprises a conjunction of two principles describing human action and reasoning- The Desire-Belief Theory of Action and The Desire-Belief Theory of Reasoning. Using A, E, and M to represent "Action", "Ends" and "Means" respectively, Sinhababu (2017) presents the two principles as follows:

"Desire-Belief Theory of Action: One is motivated to A if and only if desire that E is combined with belief that one can raise E's probability by A-ing.

Desire-Belief Theory of Reasoning: Desire that M is created as the conclusion of reasoning if and only if the reasoning combines desire that E with belief that M would raise E's probability. It is eliminated as the conclusion of reasoning if and only if the reasoning eliminates such a combination." (3)

As Sinhababu goes on to note, both these principles are meant as true psychological claims about human beings, not as conceptual claims about what we should analyze 'action' or 'reasoning' as being, nor as normative claims that entail right/wrong, rational/irrational ways of acting or forming desires.

It is also important to note here that Sinhababu's defense of HT rests on an account of desire that is richer than most. Under this account, desire has the following five aspects:

(1)"The Motivational Aspect: Desire that E combined with belief that one could increase E's probability by A-ing motivates one to A, proportional to the desire's strength times the increase in subjective probability of E. (With belief that A-ing would reduce E's probability, it likewise motivates one not to A.)" (21)

(2)"The Hedonic Aspect: Desire that E combined with increasing subjective probability of E or vivid sensory or imaginative representation of E causes pleasure roughly proportional to the desire's strength times the increase in probability or the vividness of the representation. (With decreasing subjective probability of E or vivid sensory or imaginative representation of not-E, it likewise causes displeasure.)" (26)

(3)"The Attentional Aspect: Desire that E disposes one to attend to things one associates with E, increasing with the desire's strength and the strength of the association." (30)

(4)"Amplification by Vividness: The effects of desire that E increase proportionally

with the vividness of sensory or imaginative representations of things we associate with E." (33)

(5)"The Desire-Belief Theory of Reasoning: Desire is affected as the conclusion of reasoning if and only if desire that E is combined with belief that M would raise E's probability, constituting desire that M." (35)

In summary, according to the account of desire which HT rests on, desire (1) motivates action when combined with beliefs about how to achieve its object, (2) causes pleasure or displeasure when thinking about its object depending on the subjective probability of obtaining said object, (3) disposes one to attend to things associated with its object, (4) amplifies the vividness of our representations of things associated with its object, and (5) is never affected by reasoning alone.

The Emotional Perception Model upholds the Humean theory of motivation as spelled out above, by stating that "emotions motivate us because they have desires as components" (Sinhababu 2017: 61). Delight at a possible future state of affairs increases our motivation to make this state of affairs happen, pleasure at the thought of the situation, attention to it, and increases in these phenomena when the features of this state of affairs are represented more vividly. Treating the emotion 'delight' as containing a desire to make the delightful situation obtain explains how delight causes us to behave, feel, and think (ibid).

The Color Analogy

Both moral judgments and color judgments are beliefs about objective properties typically caused by experiences of the things judged. As Colin McGinn (1983) puts it, "No-one seriously denies that color judgments have cognitive content. Everyone thinks that color judgments express beliefs" (104). A visual experience of a surface as green causes belief that the surface is objectively green[5]. The emotional perception model thus likens moral judgement to color judgement. Sinhababu (2017) calls the parallels between these judgements the "color analogy" (62). The color analogy is useful to consider here because it gives us a helpful way to distinguish Sinhababu's cognitivist, externalist, and Humean view from views that are non-cognitivist, internalist, or anti-Humean.

The color analogy distinguishes the emotional perception model from cognitivist, internalist, and anti-Humean views which see moral judgements as beliefs that can motivate action independent of desire. These views either treat moral judgement as disconnected from feeling, or as causing feeling. The analogous view on color would either ignore color experiences entirely, or treat color beliefs as causing color experiences. Neither of these analogous views on color are viable since "color experience is enormously important for color judgement, and it's a cause rather than an effect of belief" (ibid, 66).

The color analogy also distinguishes the emotional perception model from Humean views that are non-cognitivist and internalist. Such views typically treat moral judgements as

[5] While our initial beliefs about the color properties of objects are that they're objective, this might change upon subsequent reflection. As Sinhababu (2017) notes, relativism is more plausible for color than for morality; "Consider a species of aliens whose color experiences are the reverse of us, who admire those who torture juveniles until they die, and who form beliefs about color and morality on the basis of these experiences. It's more plausible that their color beliefs are free from error than their moral beliefs are" (65).

something like emotions or desires. The corresponding view for color might identify color judgments with color experiences rather than color beliefs. This is not plausible as people who are blind or otherwise impaired in terms of color perception are capable of, and frequently make, color judgements that do not arise from their own color experiences. They might do this based on what others say or by means which do not directly rely on having accurate color experience. This shows that even though color judgements are typically caused by color experiences, they should be viewed as beliefs that can be had independently of these experiences. The color analog of this non-cognitivist view thus does not work for color judgment. Also, "since many non-cognitivists [and internalists] treat moral judgment as being a motivational state like desire, their view isn't even on the map as far as the color analogy is concerned" (ibid).

The color analogy thus provides an easy way to keep the commitments of the emotional perception model in mind. If the analogy holds, it also provides further support for experientialism. Knowing what it's like to have a certain color experience seems to be a necessary condition to fully grasping a certain color concept (Campbell 2006); it seems like I wouldn't really know what green is without ever having seen green[6]. If moral concepts are like color concepts, then we'd likewise expect the possession of, or ability to have, a certain moral feeling to be a necessary condition[7] to fully grasping the relevant moral concept.

[6] I might know all kinds of things about the color green (for example, that the wavelength of green light is about 550 nanometers), but it seems that not having had the experience of seeing green leaves a significant hole in my knowledge of what green is.

[7] Here we are talking about a full experiential grasp of moral concepts. As will be discussed in the next chapter, psychopaths who do not possess moral feelings might still gain moral knowledge by deferring to those who do in fact possess the capacity for moral feeling.

Having provided a summary of experientialism and its relation to the emotional perception model as well as Sinhababu's Humean theory of motivation, we can now proceed to evaluate it in comparison with rival views in the context of some puzzles in metaethics. In the next three chapters, I argue that experientialism provides the best solutions as regards these issues.

Chapter 2: Experientialism and Psychopaths

Psychopaths, with their seeming lack of morality, have long been of interest to philosophers interested in issues of moral motivation and the nature of moral judgment. However, as Thomas Schramme (2014) notes, "almost nothing in relation to [the] phenomenon [of psychopathy] can be taken for granted" (1). This is so as the term 'psychopath' has been used in a myriad of distinct ways across a variety of disciplines. Moreover, the terminological landscape as it relates to the psychopath has also changed considerably over the last few decades (ibid). Thus, before we proceed with our consideration of what experientialism and its rivals have to say about psychopathy, it would be wise to first define exactly what we mean when we use the term 'psychopath'.

Numerous different characterizations of psychopathy have been given throughout history. Some of these characterizations overlap and some are mutually contradictory (Skeem et al, 2011). Well into the twentieth century, conceptions of psychopathy developed more or less independently in the psychiatric traditions of France, Germany and Anglo-America (Sass and Felthous, 2014). It was only in the middle of the twentieth century, when the American Psychiatric Association (APA) introduced the first edition of its Diagnostic and Statistical Manual of Mental Disorders (DSM) that the nomenclature of mental disorders (psychopathy being one of them) begin to see standardization. Even today, psychopathy is not accepted as an official clinical diagnosis, though the APA has recently recognized psychopathy as a "specifier" of clinical antisocial personality disorder in the DSM-5 (Bonn, 2016). This modern notion of psychopathy referred to by the APA, as well as most researchers of psychopathy, stems primarily from the pioneering research efforts of Robert Hare who, along with his associates, developed the Psychopathy Check List Revised (PCL-R), which aims to provide a clinical assessment of the degree of psychopathy possessed by an individual.

The PCL-R scores an individual based on personality traits and behaviors which fall into four categories: interpersonal, affective, lifestyle and antisocial; “The interpersonal traits include glibness, superficial charm, grandiosity, pathological lying and manipulation of others. The affective traits include a lack of remorse and/or guilt, shallow affect, lack of empathy and failure to accept responsibility. The lifestyle behaviors include stimulation-seeking behavior, impulsivity, irresponsibility, parasitic orientation and a lack of realistic life goals. Antisocial behaviors include poor behavioral controls, early childhood behavior problems, juvenile delinquency, revocation of conditional release and committing a variety of crimes” (ibid).

An individual who possesses and exhibits all the above traits and behavior is considered a psychopath. However, it should be noted that “the results to date suggest that psychopathy is a continuum ranging from those who possess all of the traits and score highly on them to those who also have the traits but score lower on them” (ibid) This continuum is made more complicated by the fact that most instances of psychopathy seem to be distinguishable from normal personalities only by their outward behavioral manifestations (Smith, 1984). That is to say, in the majority of cases, the brain of a psychopath looks very similar[8] to the brain of a non-psychopath (Gale, 1975). A mere possession of the traits listed in the PCL-R thus does not mean that a person has a brain that is different in kind from us ‘normals’. This is relevant insofar as philosophers are prone to conceive of psychopaths as “a case apart, as a qualitative different phenomenon from the non-psychopath” (Smith 1984: 183), when really, studies show that “psychopathic behavior is more [appropriately conceived] as an exaggerated extension of the normal personality, and is not discontinuous with it” (ibid).

[8] Though recent theories (e.g. see Koenigs (2012)) on the neurobiological basis of psychopathy typically and plausibly propose dysfunction involving the ventromedial prefrontal cortext (vmPFC), it should be noted that the psychopaths considered in these cases are mostly convicted criminals and thus probably examples of more extreme (and rarer) instances of psychopathy.

All this is not to say that no useful distinctions can be drawn between the psychopath and the non-psychopath. Rather, we should be conscious of the subtleties sketched out above, and cautious when using the empirical study of psychopaths to prove (or disprove) philosophical points. Caution having been urged, we are now in a better position to consider two of the main questions philosophers are interested in with regard to psychopathy; are psychopaths capable of genuine moral judgments? And does the empirical study of psychopathy present us with clear evidence either for or against an experientialist sentimentalism? Before we move on to consider these two questions, let us first take a quick look at an important distinction often cited by philosophers interested in psychopathy.

The Moral / Conventional Distinction

The moral/conventional distinction plays heavily in many philosophers' discussion of whether psychopaths are capable of genuine moral judgment. The distinction is a psychological paradigm developed by Elliot Turiel (1983) for the investigation of moral understanding, and rests on the 'moral/conventional distinction task', where participants are given brief accounts, with some featuring moral transgressions and some featuring conventional transgressions, and asked to respond accordingly. The original idea (Nucci and Turiel 1978) was to find out if young children could track the difference between moral and conventional transgressions. Generally, "an action is a moral transgression when it has consequences for the rights and welfare of other individuals such as hurting another individual or damaging his/her property". Conventional transgressions on the other hand are "defined by their consequences for the social order; these are actions such as talking in class [and] dressing in opposite-sex clothes" (Malatesti 2010: 3). The task was later refined to test

more specifically for four types of judgments: (1) whether the behavior in the account was permissible i.e. right or wrong, (2) the seriousness of the transgression, (3) the justification provided for finding the action permissible or impermissible, and (4) if the behavior described was impermissible, how modifiable this judgment of impermissibility is upon the removal of authority from the picture. The results of a number of empirical studies applying Turiel's paradigm to psychopaths, such as those conducted by Blair (1995) and Blair, Jones, Clark, and Smith (1997), indicate that psychopathic adult criminals, when compared to non-psychopathic adult criminals, demonstrate a seeming incompetence in distinguishing between moral and conventional transgressions. While psychopaths are generally capable of correctly judging when an action is impermissible or permissible, they seem to not be able to tell which actions are wrong in distinctively moral ways. The results thus appear to provide evidence for the conclusion that psychopaths lack moral knowledge.

However, Aharoni et al (2012) have argued that this apparent lack of moral knowledge is plausibly the result of psychopaths trying to manage impressions. Blair (1995) also explains these results as a product of social desirability factors (i.e. psychopaths saying what they think will come across as socially desirable), though he takes it to confirm the view that psychopaths lack moral knowledge. According to Aharoni et al's account, psychopaths might know the difference between moral wrongs and conventional wrongs, but they mark all wrong acts as moral wrongs (i.e. wrong in a non-modifiable authority-independent way) in a misguided attempt to impress their surveyors. Thus, in order to eliminate the possibility of impression management, Aharoni et al modified the task to use a force-choice method; "In this method, [they] informed participants with varying degrees of psychopathy that exactly half of the listed acts were prerated by members of society to be morally wrong, and instructed them to determine which half met that criterion" (ibid, 486). With this modified

task, Aharoni et al found that their psychopathic subjects were capable of distinguishing between moral and conventional wrongs, thus providing evidence that psychopaths might not lack moral knowledge after all.

I will say more about these apparently conflicting results later on in this chapter. For now, it suffices to note that experiments have been conducted to provide empirical evidence for whether or not psychopaths are capable of moral knowledge. These results are of obvious interest to philosophers interested in issues of moral motivation and the constitution of moral judgment. If psychopaths are capable of genuine moral judgment, then it seems that they provide a real-life counterexample to internalism, which is the thesis that moral judgments have a strong and necessary connection to moral motivation. This is so as psychopaths typically are not motivated to act according to their ostensibly moral judgments. The psychopath's moral deficiencies are also of interest to the debate between sentimentalists and rationalists on the constitution of moral judgment; sentimentalists have often pointed to the psychopath's lack of empathy and moral feeling as evidence of the central role of emotion to the making of moral judgment, while rationalists have in turn explained the psychopath's moral deficits by pointing more broadly at the ways in which psychopaths appear irrational and unreasonable. The rest of this chapter will focus on these two issues respectively, and how experientialism might help illuminate both of them.

Psychopathy and Internalism

As mentioned, internalism is the view that if a person makes a moral judgment, then they are necessarily motivated to act according to said judgment. For example, if I judge that lying is morally wrong, then according to internalism I am motivated, at least to some degree, to not lie. The question of whether internalism is true is an interesting question in itself, as its

truth gives us a necessary condition for something to qualify as a moral judgment. However, its truth also has broader significance. As elaborated on in the previous chapter, internalism, Humeanism, and cognitivism form a jointly inconsistent triad (Smith 1994). If it turns out that internalism is true, then one would have to give up either Humeanism or cognitivism, both of which are closely tied to the experientialist position.

Traditionally, internalism has often been conceived as a conceptual thesis, where "the concept of moral judgment is the concept of a mental state that entails the presence of corresponding motivation" (Kumar 2016a: 319). Externalists commonly argue against this conceptual thesis by citing the conceivability of 'amoralists'- agents who make moral judgments without having the corresponding motivation. In turn, the internalists reply that these amoralists are not making genuine moral judgments (Smith 1994: 68-171). The debate over internalism as a conceptual thesis thus appears to reach a stalemate; "Externalists are able to conceive of amoralists because their theory of moral judgment is externalist. Internalists are unable to conceive of amoralists because their theory is internalist" (Kumar 2016a: 319).

Research on psychopaths seems to provide the empirical test cases required to overcome this stalemate of intuitions. Whether or not one can conceive of an amoralist is heavily dependent on one's theoretical commitments, but empirical studies on psychopaths seem to provide a more objective approach to deciding if, despite lacking motivation, psychopath moral judgments are indeed genuine. However, the problem of circularity remains, so long as the debate is framed as disagreement over a conceptual thesis. This circularity motivates the arguments in Victor Kumar's (2016a) paper 'Psychopathy and Internalism'. As he puts it:

the "problem with the naturalistic approach is that it cannot avoid begging the question against internalism. Internalists insist that moral judgment entails motivation in virtue of our concept of moral judgment, in which case anyone discovered to lack moral motivation simply does not count as making a moral judgment. To leave open the possibility that unmotivated psychopaths do make moral judgments is to deny the conceptual link, and thus to assume that internalism is false" (320).

How then might we overcome this stalemate? Kumar's suggestion is that we first characterize moral judgment as a natural kind. Once we've understood empirically the extension of moral judgment as a natural kind, then we might be able to settle the debate between internalists and externalists in a non-question begging way. This approach to moral judgment requires an alternative interpretation of internalism as a metaphysical, rather than conceptual, thesis. Under this interpretation, the internalist thesis that 'moral judgments are necessarily motivating' would be on par with statements like 'water is H2O'. This solution also assumes that moral judgment can be accurately characterized as a natural kind, which is a claim that Kumar (2015) argues for in an earlier paper. I will now give a quick summary of these arguments. Following this, I will explain how Kumar's arguments complement Sinhababu's experientialist position, and how Kumar and Sinhababu might be seen as endorsing largely similar (though not identical[9]) sentimentalist views that are opposed to some other popular forms of sentimentalism.

Moral Judgment as a Natural Kind

[9] Most importantly, Kumar sees moral emotion as being a part of moral judgment whereas Sinhababu sees moral emotion as being distinct from moral judgment. I will touch on this in greater detail towards the end of this chapter. Implications of this difference will also be discussed in the next chapter.

Generally, something is appropriately classified as a natural kind if it explains a wide range of phenomena by participating in generalizations like those found in scientific laws (Bird and Tobin 2018). There are two main approaches to defining a natural kind. The first approach is reflected in the natural definitions of chemical kinds by molecular formulas; "water is H20" is now the standard example made famous by Putnam (1975) and Kripke (1971). Natural definitions of this sort provide necessary and sufficient conditions for belonging to a certain natural kind. The second, perhaps lesser known, approach is one elaborated on by Richard Boyd (1995) in his paper 'How to be a Moral Realist'. Under this approach, "some terms have definitions which are provided by a collection of properties such that the possession of an adequate number of these properties is sufficient for falling within the extension of the term" (322). Biological species are paradigm examples of natural kinds defined in this manner, where "the appropriateness of any particular biological species for induction and explanation in biology depends upon the imperfectly shared and homeostatically related morphological, physiological, and behavioral features which characterize its members" (324).

According to Kumar (2015), moral judgments are like the natural kinds defined using the second approach. Moral judgment is appropriately classified as a natural kind since it explains reasoning in a number of different domains, and also explains[10] a range of different behaviors (2889). Drawing heavily on studies involving the moral/conventional task, Kumar concludes that a wrong is typically judged to be a moral wrong if it is (1.) serious (2.) general

[10] For evidence that moral judgment explains reasoning in other domains see Pettit and Knobe (2009), Beebe and Buckwalter (2010), and Hitchcock and Knobe (2009). For evidence that moral judgment explains a range of different behaviors see; Fischbacher et al. (2001), Keser and van Winden (2000), Brandts and Schram (2001), Fehr and Gachter (2000), and Turillo et al. (2002).

(3.) authority-independent and (4.) objective (2896)[11]. Kumar also claims, according to core research on the moral/conventional distinction, that these four features have a nomological tendency to cluster together, and typically[12] co-occur in moral judgment. In other words, Kumar argues that these four features form a homeostatic property cluster that provides a natural kind definition of moral judgment.

To support this definition of moral judgment, Kumar cites two sources of evidence. First, his definition of moral judgment as a homeostatic property cluster entails that there is a nomological tendency for them to co-occur. This view predicts that if people are told that a transgression has some of the four features, they will likely believe that the other features are present as well. A study by Judy Smetana (1985) tests and confirms this prediction. In the study, children are told that an unspecified action, signified by a nonsense word (e.g. piggling), is wrong at school and also at home, so the children understood the action as generally wrong. Subsequently, when given a variant of the moral/conventional task, most of the children said that the action's wrongness is serious and authority-independent. These responses support the conclusion that, even from a young age, people are likely to infer from the presence of some features that the other features are also present. The second source of evidence Kumar cites is one we've already discussed at some length in this chapter, which is studies conducted by Blair (1995) and Blair et al (1997) investigating psychopathic moral understanding. As discussed, these studies show that psychopaths demonstrate an impaired ability to distinguish between moral and conventional violations. Kumar's account

[11] A moral wrongdoing is typically considered more serious than a non-moral wrongdoing. This does not mean that all moral wrongs are equally serious (I.e. there might be, and probably are, such things as minor moral transgressions), but simply that moral wrongs are generally considered more serious than non-moral wrongs. Judgments of moral wrongdoings are also general, in that particular actions and action types that are considered morally wrong are typically considered as such independent of cultural or social context.

[12] But importantly, not necessarily.

provides a simple explanation for these results- psychopaths are morally abnormal because they lack a full grasp of the concept of moral judgment. They are less likely than non-psychopathic subjects to distinguish moral and conventional violations with respect to all three[13] of the features tested for (i.e. seriousness, generality, and authority-independence).

If Kumar is right, then moral judgment can be defined as a natural kind. This natural kind definition can in turn be used to settle the debate between externalists and internalists in a non-question begging fashion. Accordingly, we can conclude from the studies conducted by Blair and Blair et al that psychopaths "do not have a full grasp of moral concepts, but nor do they completely fail to grasp them either. Rather, psychopaths have a more limited or tenuous grasp than non-psychopaths" (333). An empirical argument against internalism, as a metaphysical thesis, can be formulated by asking which position, internalism or externalism, best explains the psychopath's moral deficits. Kumar argues that "the best explanation for psychopath's impaired capacity for moral judgment does not entail any sort of necessary link between moral judgment and motivation. [Instead,] the explanation relies on a theory of moral judgment that accords emotion an important role, not in online moral judgment, but in development of the capacity for moral judgment" (337). In short, it is the psychopath's affective deficits, and not any sort of necessary connection[14] between moral judgment and motivation, that best explains why they do not fully acquire moral concepts.

[13] Standard variations of the moral/conventional distinction task do not include the feature of objectivity, so we do not yet know whether or how psychopathic subjects distinguish moral violations from conventional violations with respect to objectivity.

[14] Kumar and Sinhababu both think that motivation tends to occur with non-pathological instances of moral judgment. What they both deny is that this connection is a necessary one, and that a judgment made without the accompanying motivation is immediately counted as non-moral.

At this point, it should be clear that Kumar's theory rests heavily on the moral/conventional distinction. It is thus relevant to Kumar that conclusions from studies on psychopathic subjects using moral/conventional distinction tasks like those conducted by Blair (1995) have been disputed by studies conducted by Aharoni et al (2012). While the Blair studies have often been used to support the conclusion that psychopaths lack moral knowledge, Aharoni et al have argued that their study shows that these conclusions are ill-founded, and that psychopaths perform as well as non-psychopaths on a variant of the moral/conventional distinction task that has been modified to eliminate the possibility of impression management. If Aharoni et al are right, then the explanation provided by Kumar's theory falls flat; psychopaths do not do worse at the moral/conventional distinction task because they lack moral knowledge or a full grasp of the moral concept. Rather, their poor results are due to misguided attempts at impression management.

To this, Kumar (2016a) responds that Aharoni et al's modified version of the moral/conventional task is easier relative to standard versions of the task used by Blair and others; "Because psychopaths were told that half of the violations are moral and the other half conventional, they could use their answers to some questions to figure out answers to other questions" (334). If, as Kumar theorizes, psychopaths have a partial grasp of moral concepts, it would then not be very surprising that they do well with the easier task, and worse with the more difficult one. Furthermore, the results of Aharoni et al's (2012) study also show that "reduced moral categorization accuracy was significantly predicted by affective and anti-social traits' (493). In other words, psychopaths who tested higher on these two traits generally performed worse at categorizing moral and conventional wrongs. This lends support to Kumar's theory that a lack of affect is what explains the psychopath's

moral deficits. Emotion is important to moral judgment because it is vital in the development of the capacity for moral judgment.

Prinz V. Kumar

Kumar is a sentimentalist insofar as he explains the psychopath's moral abnormalities as a result of their affective deficiency. His arguments as outlined above put him at odds with another popular sentimentalist, Jesse Prinz. While Kumar (2016a) argues on abductive grounds against internalism, Prinz (2007, 2015) argues that it is internalism that provides the best explanation of psychopaths' moral deficits. Kumar's position on these issues is friendly to experientialism, so it is worth spelling them out before we consider how experientialism fits into all of this. Doing so will also show how sentimentalism can avoid internalism, as experientialism does.

Prinz (2007) argues that psychopaths do not undermine internalism. Instead, he takes psychopaths as providing internalists with "a useful piece of supporting evidence" (44). According to Prinz, internalism best explains the psychopath's moral deficits:

> "The moral blindness of psychopaths issues from an emotional blindness. If this is right, psychopathy provides positive evidence for internalism … If moral judgments are intrinsically motivating, it may be due to the fact that standard moral concepts are essentially emotion-laden. That is precisely what research on psychopathy seems to confirm." (46)

Like Kumar, Prinz explains the psychopath's moral abnormalities as caused by a lack of emotion. However, and as Kumar (2016a) points out, "Prinz's argument rests on a false premise: psychopaths make proto moral judgments, rather than, as [Prinz] thinks, failing to

make moral judgments at all[15]" (336). This difference explains why Prinz draws the opposite conclusion than Kumar though both cite the same research; Prinz thinks psychopaths are morally blind, whereas Kumar thinks that psychopaths are not completely incapable of grasping moral concepts (i.e. psychopaths are capable of making proto moral judgments).

Since both Prinz and Kumar are making empirical arguments for and against internalism. One may weigh further the merits of their arguments by how well they explain the empirical data. There are two formulations of internalism that Prinz may turn to in order to explain the finding that psychopaths make proto moral judgments. First, he may say internalism is the view that full-fledged moral judgment necessitates motivation. Since proto moral judgments are not full-fledged, they do not necessitate motivation. If this is the case, then the lack of moral motivation in psychopaths does not present a counter-example to internalism. Unfortunately, this view does not do well in explaining psychopaths' moral deficits. If motivation always accompanies full-fledged moral judgment, then why does it not accompany proto moral judgment? It is not clear how this view can explain "that once a proto moral judgment becomes a full-fledged moral judgment, motivation suddenly and as a matter of necessity joins it" (Kumar 2016a: 336).

Accordingly, Prinz might prefer to turn to an alternative formulation of internalism that says that if one makes a moral judgment to some degree, then one is motivated to the same degree. According to this formulation, psychopaths have some motivation to act according to their moral judgments. If we ignore the fact that many psychopaths seem to lack moral

[15] Here it's important to remember that an integral detail of Kumar's natural kind definition of moral judgment is that a judgment doesn't cease to be moral if it's missing one of the properties in the cluster. Even if motivation is considered part of the property cluster that constitutes moral judgment, it being missing doesn't immediately disqualify a judgment from counting as some sort of moral judgment.

motivation altogether, Prinz's argument can be reformulated in this way to provide a plausible explanation for the finding that psychopaths make proto moral judgments. A psychopath's diminished motivation might be explained by their diminished emotional capacities. If their diminished emotional capacities also undermine their capacity for theory of mind, then this formulation of internalism would be able to explain why psychopaths do not reliably make inferences regarding how others distinguish between the moral and the conventional.

This second reformulation of Prinz's argument is able to provide an explanation of psychopath deficits, as well as the finding that psychopaths make proto moral judgments. However, it is unable to account for a key piece of relevant evidence relating to VM patients. While psychopathy is a disorder that appears in early childhood, 'acquired sociopathy' is a condition that shares a similar psychological profile with psychopathy, but is caused by damage in adulthood to the ventromedial (VM) cortex (Roskies, 2003). These subjects (VM patients) have the same affective traits as psychopaths, but acquire them relatively late in life. As it turns out, VM patients are able to reliably distinguish between moral wrongs and conventional wrongs on the standard moral/conventional distinction task (Saver and Damasio, 1991). This evidence strongly suggests that it is not occurrent emotional deficits that explain the psychopath's poor scores on the standard moral/conventional distinction task. If it were occurrent emotional deficits that explain this, as the second reformulation of Prinz's argument would have it, then one would expect the VM patients to fare as badly as their psychopathic counterparts. Instead, this evidence supports Kumar's view that the best explanation of psychopath deficits is a theory of moral judgment that gives emotions an important developmental role, but does not posit any sort of necessary connection between moral judgment and motivation.

Experientialism and Internalism

The above shows us how internalism, variously construed, falters as a metaphysical thesis. Internalism is not needed as an explanation of the empirical data, and is unable to account for certain pieces of key evidence. On the other hand, Kumar's theory that emotions serve an important role in the development of the capacities required for moral judgment explains all the relevant empirical evidence, and does so with a simpler ontology. This theory and ontology coheres perfectly with experientialism. Experientialism, as formulated in the previous chapter, sees our emotional dispositions as being activated by representations of objects in our beliefs, sensations, or imagination. These emotional dispositions contain feelings and desires. Moral feelings cause moral judgments in a way analogous to how perceptions of color cause color judgments, while the desires contained within the emotional dispositions motivate us to act according to these judgments. This explains why we are typically motivated to act according to our moral judgments, while also allowing for the two to come apart in deviant cases like the psychopath's.

Experientialism is able to explain the empirical evidence discussed above in ways complemented by Kumar's arguments. According to experientialism, moral judgment is analogous to color judgment. In cases where a person has impaired color vision (or no vision at all), he is still able to form color beliefs by testimony. For example, if a reliable source truthfully tells a blind man that he has blond hair, it seems fair to say that the blind man is able to form a true color belief about his hair despite not being able to see color. There are also varying degrees of color impairment. Some color blind people have no trouble identifying most colors, while some are totally color blind and can only see things as black

and white or in shades of gray. It seems like the empirical evidence confirms the color analogy with regard to these aspects of moral judgment; psychopaths are able to form moral beliefs by testimony, as evidenced by high-functioning nonviolent psychopaths like James Fallon (2013), and as reflected in their ability to perform well on easier versions of the moral/conventional distinction task. There are also varying degrees of affective impairment that seem to directly result in varying degrees of moral competence, as evidenced by Aharoni et al's (2012) finding that a reduced ability to perform in the moral/conventional distinction task moral was significantly predicted by the extent to which the subject possessed affective deficiencies and anti-social traits.

Moreover, the color analogy is complemented by Kumar's developmental account of moral emotions. Congenitally blind people might be able to grasp color concepts in a certain sense, but never having seen color, this grasp is severely limited. This limitation is one that does not usually apply to those who become blind later on in life. Having already seen color, these people possess color concepts comparable to the sighted. Similarly, and as discussed above, studies (Saver and Damasio, 1991) show that psychopaths typically demonstrate a tenuous grasp of moral concepts relative to VM patients who are able to reliably distinguish between moral and conventional wrongs. VM patients, having had moral emotions during their developmental years, possess a more secure grasp of moral concepts.

Sinhababu's experientialism and Kumar's account of moral judgment as a natural kind both do well in reflecting and explaining the relevant empirical findings. However, they differ significantly on at least two counts.

Firstly, since Kumar builds his theory on findings from research using the moral/conventional distinction task, it is entirely possible that future research on the topic finds his theory's predictions wrong. Kumar (2015) himself admits as much, saying his theory makes "predictions that can be falsified…in particular, it makes the empirical prediction that the four features are in homeostasis…[it] also entails that judgments that lack several of the four features (instead of just one) are not moral judgments, that is, are neither typical nor atypical moral judgments. So, if [clear] counterexamples could be produced, in which judgments that are intuitively classified as moral lack several of the four features, that would count against the view" (2903). Likewise, experientialism is built on a foundation that might be found false; the emotional perception model assumes the Humean theory of motivation (Sinhababu, 2017). Any finding that proves the Humean theory wrong would thus count against experientialism. Additionally, by saying that "moral concepts apply to whatever accurate moral feelings objectively represent" (70), experientialism makes the capacity for moral feeling in some sense necessary for moral judgment. While those without moral feeling might gain some facility with moral concepts through the testimony of those with moral feeling, an isolated community of beings without moral feelings seems, according to experientialism, incapable of genuine moral judgment. Any reason to accept that such a community should be considered capable of moral judgment despite their collective lack of moral feeling would thus count against experientialism. This shows that experientialism and Kumar's account can each be argued against and proven wrong quite independently of each other. They have different falsifiability conditions, so to speak. On a tangential note, a related difference between experientialism and Kumar's account is that experientialism gives us what seems like a more specific chemical kind type definition of moral judgment whereas Kumar's account gives us a vaguer property cluster definition of moral judgment.

Secondly, Kumar (2015) sees moral judgment as being "a hybrid state of moral belief and moral emotion" (2889), thus arguing that moral judgment is both a cognitive and a non-cognitive state[16]. On the other hand, Sinhababu (2017) sees moral judgment as "belief alone, rather than belief plus emotion" (67). This means that while Kumar is against Prinz and with Sinhababu on the issue of what psychopathy tells us about internalism, Kumar is with Prinz and against Sinhababu on the issue of what constitutes moral judgment, with Prinz also favoring a constitution model that takes moral judgment to be a hybrid of belief and emotion. The next chapter will touch on the benefits of a purely cognitive theory of moral judgment. For now, it suffices to point out that this is an important difference between the two theories.

Rationalism and Psychopathy

This chapter so far has been largely concerned with differentiating between sentimentalist treatments of psychopathy. We first considered how research on psychopathy might be used to settle the debate between internalism and externalism. In doing so, we have seen both that sentimentalism need not commit itself to internalism, and that a sentimentalism which resists internalism does better in explaining the relevant empirical evidence. We then looked at how experientialism is complemented by but distinct from Kumar's theory of moral judgment.

We now turn our attention to a rationalist treatment of psychopathy. Heidi Maibom (2005, 2010) argues that psychopaths' impaired ability to make moral judgments is a consequence

[16] This is consistent with Kumar's claim that psychopaths make proto moral judgments. Psychopath moral judgments are missing what Kumar considers to be a key aspect of ordinary moral cognition.

of both cognitive deficits as well as emotional ones. She (2010) says that "new sentimentalists [like Prinz, Kumar, and Sinhababu] tend to ignore the numerous cognitive deficits exhibited by psychopaths in favour of their emotional deficits" (1004). She (2005) has two main arguments regarding rationalism and psychopathy, one defensive and the other offensive. I will briefly sketch out these two arguments, and demonstrate that experientialism provides replies to both. Doing so will not disprove rationalism, but it will show that psychopathy presents a case against rationalism, and that sentimentalism is better suited to account for the psychopath's moral failings.

Maibom's (2005) defensive argument can be summarized with the following premises:

(1)If rationalism about morality is true, "we should expect [the] immorality demonstrated [by psychopaths] to be connected with a high degree of practical irrationality" (238).

(2)The immorality demonstrated by psychopaths is connected with a high degree of practical irrationality.

(3)Therefore, the truth of rationalism is consistent with psychopathy.

The flipside to premise (1) is that if sentimentalism is true, "we should expect amorality to correlate significantly with disturbed emotions" (238). Shaun Nichols (2002) argues that psychopaths have disturbed emotions but intact reason, and thus that psychopaths present a case for sentimentalism and against rationalism. Maibom (2005) acknowledges that psychopaths have disturbed emotions, but argues that they also have deficits in their practical reason. This means that while psychopathy might provide support for sentimentalism, it does not speak against rationalism. She writes:

"There is experimental and anecdotal evidence for a number of cognitive shortcomings in psychopathic individuals. They frequently act in their own worst interest (Hare, 1993; Blair et al., 2001b), exhibit cognitive-perceptual shortcomings in the recognition of certain emotions in others' faces and voices (Blair et al., 2001a; Blair and Coles, 2000; Blair et al., 2002), have attention deficits, a grossly inflated view of their abilities, and are intransigent to certain forms of conditioning (Hare, 1978)." (242)

With the above evidence, rationalism is able to claim quite plausibly that the psychopath's moral deficit viewed in terms of practical irrationality is explained by their more general rational deficits. While sentimentalists might point to global emotional deficits in support of sentimentalism, rationalists can also point to global rational deficits in defense of rationalism.

This then leads us to Maibom's offensive argument, which can be summarized with the following premises:

(1)If psychopaths' practical reasoning deficits cannot be explained in terms of an underlying emotional deficit, then sentimentalism cannot account for psychopaths' moral deficits.

(2)Psychopaths' practical reasoning deficits cannot be explained in terms of an underlying emotional deficit.

(3)Sentimentalism cannot account for psychopaths' moral deficits.

Premise (1) of this argument assumes that in order to explain psychopaths' moral deficits, we'll have to be able to explain their practical reasoning deficits. This seems a fair enough assumption to make, seeing that their moral deficits do indeed seem related in some way to

more general rational deficits. Some sentimentalists might reject this assumption, saying that psychopaths' moral deficits can be explained wholly in terms of their emotional deficits without reference to their practical irrationality. However, it is not necessary for sentimentalism to reject this assumption in order to challenge the argument; experientialism is able to grant the rationalist premise (1) and challenge the argument by rejecting premise (2).

The general rational deficits Maibom cites are (A) psychopaths frequently act in their own worst interest, (B) demonstrate shortcomings in the recognition of certain emotions in others' faces and voices, (C) have attention deficits, (D) a grossly inflated view of their abilities, and (E) are intransigent to certain forms of conditioning. (A) and (D) might be grouped together as stemming ultimately from (C), in that a psychopath's attention deficits can explain why they frequently act in their own worst interests and why they have a grossly inflated view of their abilities; psychopaths display attention only to certain things in the short term often (though not always) at the cost of their long term interests and they pay a disproportionate amount of attention to themselves which quite naturally (though not necessarily) results in an inflated view of their abilities[17]. The Humean aspect of the emotional perception model predicts these failures of attention in terms of whether desire is present. Recall that one of the aspects of Sinhababu's (2017) account of desire is the attentional aspect, which says "desire that E disposes one to attend to things one associates with E, increasing with the desire's strength and the strength of the association." (30) Since desire combines with moral feelings to form moral emotions, experientialism is able to explain a psychopath's attention deficits in terms of their emotional deficits. (B) is also to be expected given the color analogy; a psychopath's inability (or impaired ability) to recognize

[17] More can probably be said about how the Humean theory accounts for the psychopath's cognitive/rational deficits (and more should probably be said defending this account) but the point being made here is simply that the Humean theory predicts and explains these deficits in terms of emotional deficits.

certain emotions in others' faces and voices is analogous to a color-impaired or blind person's inability to recognize colors. Having never experienced these emotions (or having had only deficient versions of these emotions), psychopaths naturally have trouble recognizing these emotions in others. (E) is likewise explained by psychopaths' emotional deficits. If the psychopath's problem was mainly rational, it would seem that education would be sufficient for conditioning, especially given many psychopaths otherwise normal intelligence. The fact that despite their elsewhere intact rationality, psychopaths demonstrate an intransigence with respect to certain forms of conditioning, suggests that these are conditions that cannot be reasoned away. Going back to the color analogy, conditioning or reasoning with a psychopath regarding their peculiar practical irrationalities would be like trying to get a color-blind person to differentiate between two colors by explaining their spectral properties, or by punishing them when they fail to differentiate and rewarding them when they are successful. The problem with the color-blind person/ psychopath is that they are color/ emotionally deficient, and perceptual deficiencies are conditioning resistant and generally independent of reason.

In short, it seems that a psychopath's non-moral rational deficits are predicted by an experientialist sentimentalism. It also seems that since psychopaths demonstrate an intact rationality in areas unaffected by emotion, it is their emotional deficiencies that explain their rational deficiencies, and not the other way around. Accordingly, even if psychopaths' moral deficiencies are related to their more general rational deficiencies, they present a case in support of sentimentalism over rationalism insofar as it is possible to explain their rational deficiencies in terms of their emotional ones.

Chapter 3: Experientialism and Moral Twin Earth

A prominent variety of naturalistic moral realism endorsed by Richard Boyd (199) and Peter Railton (1986) views our first-order ethical theories as providing *a posteriori* definitions of moral predicates, and as specifying those natural properties that moral properties are constituted by or identical with (Rubin, 2014a). I will use 'NMR' to refer to this variety of naturalistic moral realism. According to NMR, a first-order ethical theory like hedonistic act-utilitarianism is seen as making the claim that the moral property 'rightness' is identical with the natural property 'maximizing pleasure' (ibid). The *a posteriori* nature of these moral-natural definitions is typically defended by appeal to the causal theory of reference, which says that the moral predicate 'right' refers to the property that stands in the appropriate causal relation to the speaker's use of 'right' (Boyd, 1995). Since there seems to be no *a priori* way for the speaker to determine the natural property that stands in this position, advocates of NMR argue that it is a matter for empirical investigation.

Insofar as it is committed to the causal theory of reference, NMR faces a serious challenge from moral twin earth (hereafter MTE) arguments pioneered by Terence Horgan and Mark Timmon's (1991). In this chapter, I will first explain the problem posed to NMR by MTE. I will then go through three responses to MTE and argue that none of them are completely satisfactory. Following this, I will present an experientialist solution to MTE, argue that it deals satisfactorily with the threat posed by MTE to NMR, and consider how this solution allows us to further distinguish experientialism from other positions considered in the previous chapter. Ultimately, the aim of this chapter will be to show that experientialism provides NMR with the best response to MTE.

Moral Twin Earth

Here is a simplified version of Horgan and Timmon's (1991) MTE thought experiment:

Suppose that at some point humans reach a moral consensus that aligns with that of hedonistic act-utilitarianism; actions are right insofar as they maximize pleasure and wrong insofar as they fail to do so. This hedonistic act-utilitarianism accords with and regulates humanity's usage of moral terms. Suppose further that humanity finds in a far off galaxy some planet they call Moral Twin Earth (MTE). MTE is populated by beings that are very similar to humans and is as close to being identical to Earth as possible, save for one important difference. While MTEarthians speak a language that is orthographically and phonologically identical with English, humans find that the MTEearthians have reached a moral consensus at odds with their own. The MTEarthian moral consensus aligns with that of a non-consequentialist, deontological theory of moral action; actions are right insofar as they accord with the maxim by which a person can also will that it would become a universal law. This deontological theory accords with and regulates MTEarthian use of moral terms.

Jim, a delegate from MTE, visits Earth, and says that "lying is always wrong because it can never accord with the maxim by which a person can also will that it would become a universal law". Joe, his Earthian counterpart, replies that "lying is not always wrong because a person would be morally obliged to lie were it the case that telling the truth would cause more pain than pleasure". On the face of it, it seems that the right way to characterize this exchange between Jim and Joe is to say that the two are expressing a substantive moral disagreement about whether lying is always wrong. This becomes more obvious when we add the further stipulation that moral terms like 'wrong' have the same practical role on Earth as it does on MTE. That is to say, moral terms on both Earth and MTE are (1) "used to evaluate actions, persons, and institutions"; (2) "used to reason about considerations…bearing on well-being"; (3) users of moral terms are typically disposed to

avoid actions that they deem 'bad' or 'wrong' and (4) users of moral terms take an action's being 'right' or 'wrong' to be "especially important in deciding what to do" (Horgan and Timmons, 1991). Unfortunately for NMR, the intuition that Jim and Joe are disagreeing is at odds with the causal theory of reference. If the causal theory of reference is correct, then the content of 'wrong' for Jim (let's call it wrong-M) is "not in accordance with the maxim by which a person can also will that it would become a universal law" whereas the content of 'wrong' for Joe (let's call it wrong-E) is "failing to maximize pleasure". This means that both Jim and Joe are making true claims[18]: Jim's claim that lying is always wrong is true just in case it accords with wrong-M and Joe's claim that lying is not always wrong is true just in case it accords with wrong-E. The causal theory of reference as applied to moral terms thus seems to entail that the disagreement between Jim and Joe is merely apparent, rather than genuine. As such, the MTE argument can be stated with the following premises (Rubin, 2014a):

(1) If it is appropriate to apply the causal theory of reference to moral terms, then 'wrong-M' expresses an entirely distinct meaning from 'wrong-E', and the two predicates are not intertranslatable.

(2) If 'wrong-M' expresses an entirely distinct meaning from 'wrong-E' and the two predicates are not intertranslatable, then there cannot be genuine moral disagreement between humans and MTEarthians.

(3) It is not the case that there cannot be genuine moral disagreement between humans and MTEarthians.

[18] If one assumes, plausibly, that lying can never accord with the maxim by which a person can also will that it would become a universal law.

(4) Therefore, it is not the case that it is appropriate to apply the causal theory of reference to moral terms.

Challenging Premise (1)

David Brink (2001) challenges the MTE argument by rejecting premise (1). He claims that the worries generated by MTE are most plausible when one thinks of causal regulation in a narrowly extensional way. As applied to NMR, this narrowly extensional way of thinking understands causal regulation "in terms of the features of the world that causally regulate people's *actual* use of moral terms" (168). In opposition to this, Brink argues that we should understand causal regulation in counterfactual terms; "on this view, terms refer to properties that regulate not just actual usage, but also counterfactual or hypothetical usage- in particular, the way speakers would apply terms upon due reflection in imagined situations and thought experiments" (ibid).

This counterfactual way of thinking about causal regulation coupled with an account of error that acknowledges our imperfect use of moral predicates leads us to a picture of causal regulation that is dialectical. This means that when deciding what property causally regulates our use of moral terms, "we make trade-offs among our…considered moral judgments…in response to conflicts, making adjustments here at one point and there at another, as coherence seems to require, until our ethical views are in dialectical equilibrium" (169). Brink argues that if this is right, then 'wrong-M' and 'wrong-E' are not distinct. With the counterfactual account of content-fixing for moral predicates, all Jim's exchange with Joe represents is a case of dialectical inequilibrium, and the issue here becomes the familiar one about whether "extant moral disagreement undermines prospects for dialectical

convergence" (ibid). Brink's argument against MTE can be summarized with the following premises:

(1) We should understand causal regulation in counterfactual terms (as opposed to a narrowly extensional understanding); on this counterfactual view we would get a picture of causal regulation that is dialectical.

(2) Under a dialectical picture of causal regulation, all Jim's exchange with Joe represents is a case of dialectical inequilibrium, where 'wrong-M' and 'wrong-E' are not distinct.

(3) Premise (1) of the MTE argument is thus false and MTE poses no special challenge to NMR apart from the familiar one about moral disagreement and dialectical convergence.

The problem with Brink's argument is that even on the counterfactual view of causal regulation, it still seems entirely plausible that two isolated communities reach opposing but self-consistent positions regarding what property regulates their use of moral terms. The consensus within the individual communities of Earth and MTE is stipulated to be perfect. Therefore, while Brink is probably right to say that our "actual usage does not track any one set of morally relevant properties consistently" (168), he is too hasty to conclude that this too must be the case in our thought experiment and that under a counterfactual view we would *necessarily* get a picture of causal regulation that is dialectical. Proponents of the MTE argument need only assume that the preferred ethical theories of Earth and MTE contain no internal contradictions. If this assumption is granted, then it is entirely possible for both Jim's statement "lying is always wrong" and Joe's statement "lying is not always wrong" to be true at the same time (even while assuming a counterfactual understanding of causal regulation), just in case Jim's use of 'wrong' coheres perfectly with wrong-M and Joe's with

wrong-E. Jim and Joe's exchange might then be more appropriately characterized as a dialectical stalemate rather than a case of dialectical inequilibrium. Importantly, this stalemate also seems like a case of genuine moral disagreement. In order for his objection to be convincing, Brink has to do more than point to the fact that our actual use of moral terms is imperfect and that we have to make adjustments here and there for the sake of coherence. He has to give us reason to think that is *impossible* for a community to have a perfectly coherent consensus regarding the use of moral terms, if it is also possible that there exists some separate community that has an equally perfect consensus aligned with some opposing theory of right and wrong. Since Brink does not do this, his argument fails to properly engage with MTE.

Challenging Premise (2)

In "Return to Moral Twin Earth", David Merli (2002) spells out three distinct challenges to the MTE argument. First, Merli challenges premise (1) by introducing an idealized account of content-fixing for moral predicates. Next, he questions the truth of premise (3), arguing that our intuition that Jim and Joe are in fact expressing genuine disagreement might not be reliable given our epistemic limitations. Finally, he attacks premise (2), arguing that even if wrong-M and wrong-E express distinct meanings, it does not follow that there cannot be substantive disagreement between humans and MTEarthians.

Merli's first argument can be replied to along more or less the same line as Brink's (2001)[19]. We will look at challenges to premise (3) in the next section of this chapter. For now, we will focus on Merli's third argument which is his attack on premise (2).

Merli's third argument proposes that there is genuine disagreement in MTE cases, but that the disagreement is more appropriately classified as regarding the practical "all-things-considered" 'ought' rather than the more narrowly construed moral 'ought'. In other words, "the speakers' disagreement should not be understood as a disagreement over whether a given action is morally wrong; instead, it is a disagreement over which act to do" (Rubin 2014a, 35). As Merli (2002) notes, "the strategy of this third argument urges us to revise our initial views of the disagreement's location" (233). If this response works, then we have an alternative explanation regarding our intuition that there is genuine disagreement between Jim and Joe. While there might not be moral disagreement, Merli maintains that "some disagreement is better than none at all" and that his "reply at least preserves an important part of the phenomena, namely our sense that there's something to talk about when [human] moralists and [MTEarthian] moralists get together" (ibid).

[19] Merli (2002) argues that "connecting issues of reference with questions about an idealized moral theory looks to be a promising direction for the realist… [For] if the correct account of our moral properties…were given by the end-of-the-day moral theory, then the issue of shared meaning comes down to…the question of convergence" (223). This is similar to Brink's (2001) characterization of Jim and Joe's exchange as being a case of dialectical inequilibrium. Under this view, Jim and Joe would make adjustments in the face of disagreement until their ethical views converge. As I've argued, this challenge to premise 1 is only persuasive if we also think it's impossible for two communities to each have a perfect consensus regarding the use of moral terms while embracing opposing theories of right and wrong. While Merli goes further than Brink in providing reasons to be optimistic about convergence, he admits that all this is tentative and reliant on the appeal of future well-worked-out moral views. However, even if we follow Merli in being optimistic about convergence, the worry remains that NMR allows for the possibility of a dialectical stalemate, as touched on earlier.

Accordingly, Merli's argument against premise (2) combines "realism about moral discourse with expressivism about all-in endorsement [regarding what to do]. According to this view, moral rightness is a matter of natural fact. [On the other hand,] an answer to the question of what to do…is not a factual judgment but an endorsement of one course of action or one set of reasons for action" (236). The problem with this combination view, as Rubin (2014a) points out, is that "it deprives NMR of its greatest theoretical advantage over its main metaethical rivals…[While] expressivism (a) requires us to view the declarative surface grammar of moral sentences as misleading, (b) cannot make good sense of moral sentences embedded in conditional statements…(c) cannot make good sense of the apparent logical validity of arguments involving moral predicates, and (d) cannot make good sense of our practice of predicating truth of some moral sentences..., realist treatments of moral discourse face none of these challenges….[Since] precisely the same problems confront expressivism about the normative [all-things-considered] 'ought'…, a moral realist who adopts expressivism about the normative 'ought' immediately takes on the burden of solving all of these difficulties" (39).

Furthermore, disagreement about what to do is closely connected to, and often follows from, moral disagreement. It seems that the most natural reading of Jim and Joe's disagreement is that it is primarily about right and wrong, and that this moral disagreement might lead to further disagreement about how to act. In the case of MTE then, we might very plausibly take Jim and Joe's moral disagreement to explain some further practical disagreement. While Merli (2002) says that "the question of whether [Jim and Joe's] dispute is centered on 'right' or somewhere else is a fairly sophisticated one, and it's not clear that our ordinary linguistic intuitions have much to say about the matter" (233), he does not offer any good reason to reject the stipulation that Jim and Joe are having a moral disagreement. Merli is

probably correct in pointing out that *if* we reject this stipulation we would still be able to formulate alternative explanations for our intuition that there is disagreement, but since we can accept both that there is moral disagreement and that it is likely that this moral disagreement leads to practical disagreement, his rejection of premise (2) is unpersuasive.

Challenging Premise (3)

As discussed, the MTE argument problematizes NMR's reliance on the causal theory of reference by showing that it leads to a conceptual relativism that moral realists are generally uncomfortable with. However, some philosophers argue that the relativism generated by MTE is, upon closer examination, nothing to worry about. This line of reply targets premise (3) of the MTE argument, which says that it is not the case that there cannot be genuine moral disagreement between humans and MTEarthians. By arguing that the truth of NMR is consistent with the impossibility of moral disagreement between Earth and MTE, proponents of this reply preserve the appropriateness of applying the causal theory of reference to moral terms.

I will follow Michael Rubin (2014b) in calling responses of this sort "bullet-biting replies to MTE" (286). Proponents of the bullet-biting response include David Merli (2002), Andrea Viggiano (2008), and Neil Levy (2011). Their challenge to premise (3) centers on a rejection of what Rubin (2014b) calls the "univocity judgment", which is "the judgment that [wrong-M and wrong-E] are univocal: they have a common meaning that makes substantive disagreement possible" (290). In this section, I will first summarize Rubin's arguments for why bullet-biting replies are generally unpersuasive. In the next section, I will consider an additional argument from Neil Sinhababu (2019) for thinking that even if we grant

bullet-biters the claim that standard MTE cases do not demonstrate moral disagreement, we still have reason to be wary of applying the causal theory of reference to moral terms.

The univocity judgment is based on the semantic intuition that wrong-M and wrong-E are univocal (i.e. that they are about the same thing). Since most proponents of the bullet-biting response agree that this semantic intuition seems prima facie correct, they have to provide an explanation as to why it's ultimately mistaken. Broadly speaking, there are two ways that bullet-biters go about doing this. The first is exemplified in Neil Levy's (2011) argument that the univocal judgment becomes less plausible when we reflect on the inevitable divergence in Earth's and MTE's futures. The second is represented by Andrea Viggiano's (2008) argument that the seeming correctness of the univocity judgment is a result of our current imperfect epistemic position and that when we eventually arrive at our best possible moral theory we will conclude that wrong-M and wrong-E are not univocal.

Levy (2011) claims that MTEarthians lack both a genuine moral vocabulary and genuine moral thought. If this claim is correct, then it follows that wrong-M and wrong-E are not univocal. In support of this claim, Levy offers the following explanation as to why we might have been misled by the MTE thought experiment:

> "[T]he histories of these two planets ought to diverge over time. The psychological differences alone between us and the inhabitants of MTE would be sufficient to force a divergence, which would set the planets off on quite different trajectories; when we add to that the fact that these psychological differences entail differences in moral institutions and practices, the divergence becomes quite radical. So were we to encounter MTE later in its (and our) history, the differences between us and them would be striking. These differences would be so striking, I think it is fair to claim, that there would be little temptation to think that our moral terms had the same reference" (143).

According to Levy then, we were misled by the thought experiment because it was vague about which time in the histories of the two planets the supposed disagreement took place. If the time were specified, we would be more alert to the possibility of radical divergence, given our different psychological and institutional trajectories.

Rubin (2014b) argues, and I agree, that Levy's objection is unpersuasive. First of all, the MTE thought experiment is able to concede significant divergences in moral institutions and practices. MTEarthians will probably exhibit a greater tendency than humans to avoid those actions that do not accord with their deontological maxim, even when those actions maximize pleasure. This tendency alone will most likely result in the moral institutions of MTE differing quite dramatically from those of Earth. However, we should be careful not to overstate the extent of these divergences. While there will obviously be differences in the moral practices of humans and MTEarthians, we should also predict a considerable amount of overlap. For example, torturing an innocent person for no reason is an action that most probably falls within the extensions of both wrong-E and wrong-M (ibid).

Secondly, there appear to be many concrete examples of historical divergences in Earth's own history that show that even in the face of radical difference, univocity judgments still hold. For example, a large portion of Christian fundamentalists believe that promiscuity is inherently morally wrong (i.e. promiscuity is wrong-C). Their condemnation of promiscuity is based in a religious world view not shared by secularists who think that there is nothing inherently morally wrong with being promiscuous (i.e. promiscuity is not wrong-S). Despite the radical difference between the world view of the Christian fundamentalist and that of the secularist, it seems that most of us would agree that wrong-C and wrong-S are univocal,

and that when a Christian fundamentalist says “promiscuity is inherently wrong” and a secularist replies that “promiscuity is not inherently wrong”, they are expressing a genuine moral disagreement.

Furthermore, Levy’s objection seems to ignore (or at least pays insufficient attention to) a key stipulation in the thought experiment. It is stipulated that moral terms like ‘wrong’ play the same practical role on MTE as they do on Earth. This similarity in the practical role played by ‘wrong-M’ and ‘wrong-E’, coupled with the fact that “from our present point of view, both planets in the MTE thought experiment represent epistemic possibilities for us” (Rubin 2014b: 298), makes it likely that even granting significant divergences, we would still judge ‘wrong-M’ and ‘wrong-E’ to be univocal. As Rubin says: “If the similarity in practical roles is not sufficient to make this likely, then bullet-biters need to explain why it is not. It is not enough simply to point out that the terms will be applied to different extensions as the histories of the planets diverge” (ibid, 296).

This then brings us to Viggiano’s (2008) objection to MTE. If we grant significant divergences in moral practices and still consider both planets in the thought experiment as representing epistemic possibilities for us, then maybe the fault lies with our current epistemic position. Accordingly, Viggiano claims that the univocity judgment is a product of our current ignorance and argues that our epistemically advanced descendants, having concluded that hedonistic act-utilitarianism is the theory that regulates human use of ‘right’ and ‘wrong’, will be aware of certain semantic constraints that remain unknown to humans now. Our epistemically advanced ancestors will thus be able to see that MTEarthian “moral” terms like ‘wrong-M’ are not genuine pieces of moral vocabulary since they violate certain substantive

constraints as yet unknown to us in the present: “they will neither judge [‘wrong-M’] to be translatable as [‘wrong-E’], nor will they see themselves as expressing substantive moral disagreement with [MTEarthians]” (Rubin 2014b: 302). Viggiano’s argument can thus be summarized with the following premises:

(1) When our human descendants conclude that hedonistic act-utilitarianism is the moral theory that regulates our use of ‘right’ and ‘wrong’, they will have uncovered semantic constraints that exclude all other moral theories from regulating the use of ‘right’ and ‘wrong’ i.e. they will conclude that no predicate counts as a moral predicate if it is regulated by some other set of natural properties different from that of our final moral theory.

(2) MTEarthian use of ‘right’ and ‘wrong’ is regulated by a different set of natural properties than that of our final moral theory.

(3) From the perspective of our epistemically advanced descendants, the MTEarthian predicates ‘right’ and ‘wrong’ do not count as moral predicates and the univocity judgment is wrong.

Since premise (2) of this argument is a stipulation of the MTE thought experiment, proponents of MTE will have to find fault with the first premise. Fortunately for them, this seems easy enough since the truth of premise (1) is far from obvious: “Even if we grant that speakers in our community eventually will converge on [a] single moral standard, this does not by itself entail that those speakers will treat having the same extension picked out by that standard as itself partly definitive of what it is to be a moral predicate” (Rubin 2014b: 303). In other words, it does not seem that concluding that hedonistic act-utilitarianism is the moral theory that regulates our use of ‘right’ and ‘wrong’ entails the discovery of semantic constraints that necessarily exclude rival moral theories by definition.

Furthermore, let's say that MTEarthians subscribe to a radical anti-consequentialist deontology according to which it is not wrong to let a child suffer, even when you can easily stop the child's suffering at little personal expense (ibid, 305). It seems that this theory is one that we presently exclude based on our best moral reasoning. Let's then ask ourselves if there is moral disagreement when a human says to this that it is wrong to let a child suffer when you can easily stop their suffering at little personal expense. It would help to note that our disagreement with MTEarthians here plausibly concerns normative ethical issues where we disagree with actually existing humans; one need only turn on the news to see that there are unfortunately some extreme libertarian types who would say it's okay to let the child suffer (especially if the child is a foreigner) even if they could stop it at little cost to themselves. If we would take ourselves to be in moral disagreement with these actually existing humans (as I think we should), then we should do the same in the case of MTE. It thus appears that premise (1) of Viggiano's argument is false for here "we have a case in which we have ruled out a first-order moral theory, but no corresponding substantive constraint is revealed to constrain our use of moral terms" (ibid). Our epistemically advanced descendants can thus conclude on some final moral theory without discovering any substantive moral constraints that speak against the univocity judgment.

If I'm right then none of the challenges discussed thus far satisfactorily defuse the problem of MTE for NMR. Brink's (2001) challenge to premise (1) of the MTE argument fails because a dialectical stalemate remains possible under his proposed counterfactual account of causal regulation. Merli's (2002) argument against premise (2) is problematic because it deprives NMR of its greatest theoretical advantage over its main metaethical rivals, and his move to view the disagreement in question as practical rather than moral appears undermotivated,

especially since it seems most natural to accept both that there is moral disagreement and that it is likely that this moral disagreement leads to practical disagreement. Levy's (2011) and Viggiano's (2008) objection to premise (3) both falter in the face of concrete real-life examples that speak against their attempts to explain away the univocity judgment.

A Further Problem for NMR

In 'One-Person Moral Twin Earth Cases', Neil Sinhababu (2019) presents two cases which demonstrate that the causal theory of reference[20] generates incorrect truth-conditions for moral terms. Unlike typical MTE cases which try to show that the causal theory cannot account for moral disagreement, Sinhababu's one-person MTE cases do not essentially involve interpersonal disagreement. This allows opponents of the causal theory to grant that even if standard MTE cases do not demonstrate moral disagreement[21], the causal theory nonetheless provides unrealistic truth-conditions for moral claims. In this section, I will summarize the first case presented by Sinhababu in his paper and explain the problem it poses to NMR. In the next section, I will present an experientialist solution to MTE that is also able to handle the one-person MTE case summarized here.

In Sinhababu's first one-person MTE case, *Alien Nurse*, you are an astronaut who wakes with a headache while an alien nurse greets you and tells they've rescued you from a crashed

[20] Sinhababu's (2019) target is more generally theories that allow the environment to determine the content of moral concepts. This includes the causal theory of regulation, the stabilizing function account which "treats the content of a concept as what it covaries with when serving the function that explains its persistence" (19), and the connectedness model of Schroeter and Schroeter (2014) which determines the content of concepts by the "whole set of attitudes, dispositions, social practices, and environmental feedback loops associated with the historically and socially extended representational tradition" (14). Since the causal theory is by far the most popular of the three and the object of Horgan and Timmon's (1991) original thought experiment, I will restrict my attention to it and say nothing else about the latter two.

[21] Though, as discussed, it seems that standard MTE cases really *do* demonstrate moral disagreement.

spacecraft and that this is their first meeting with someone from another planet. The nurse tells you that they don't yet know whether you're from C-Earth or D-Earth, both of which they've been observing using high-tech apparatus that allows them an intimate knowledge of both planets. C-Earth and D-Earth are nearly identical and neither planet is aware of the other's existence. The main difference between the two planets has to do with the moral judgments of their inhabitants so a test involving trolley problems has been prepared to determine which planet you're from. The nurse tells you that the crash might have scrambled your brain's moral judgment centers and so your answers to the test might be misleading about which planet you're from but that they have been asked to administer the test anyway. The nurse then reads from a script the following trolley problem: "The only way a bystander can save the lives of five people from being run over by a trolley is by pushing a large man off the bridge to block it. Pushing the large man off the bridge will kill him. What is your moral judgment of pushing the large man so as to save the lives of five people?" You think to yourself that it is always wrong to push an innocent bystander to his death, even if doing so would save the lives of the five people, so you tell the nurse that pushing the large man off the bridge is morally wrong. The nurse replies that your answer concords with the moral judgments of D-Earthians, whose moral concepts are causally regulated by a deontological theory which says that you should always treat others as ends in themselves. On the other hand, C-Earthians have their moral concepts regulated by aggregate happiness and have agreed on the consequentialist answer that you should push the large man. The nurse then reminds you that as things stand, they still do not have enough information to determine which planet you come from. You might be a D-Earthian with intact moral judgment centers or a C-Earthian with moral judgments scrambled by injury. An alien doctor then enters the room and tells you that they've just gotten back the results from their forensics team which has determined that you're actually from C-Earth. The doctor says this means that the crash scrambled your moral judgments but that they'll care for you until you

recover. As the nurse and doctor leave you to attend to other patients, your thoughts return to the trolley problem. You ask yourself: "Does knowing that aggregate happiness causally regulates moral concepts in my linguistic community really settle the issue of whether it's wrong to push the large man?"

Like Sinhababu, I take the natural answer here to be: "no, the facts about causal regulation don't settle the moral issue" (Ibid, 18). The argument from *Alien Nurse* can be summarized as follows:

(1) If it is appropriate to apply the causal theory of reference to moral concepts, then knowing that aggregate happiness causally regulates moral concepts in my linguistic community settles the issue of whether it's wrong to push the large man.
(2) It is not the case that knowing that aggregate happiness causally regulates moral concepts in my linguistic community settles the issue of whether it's wrong to push the large man.
(3) Therefore, it is not appropriate to apply the causal theory of reference to moral concepts.

Alien Nurse is superior to standard MTE cases for several reasons: Firstly, the univocity judgment in this one-person case is more obvious than in the standard cases; the alien nurse uses 'moral' here in a broad sense that encompasses concepts on both planets, and her question about whether pushing the large man is morally permissible is quite clearly neutral between the moral concepts of C-Earth and D-Earth, leaving open responses from either. Secondly, the fact that C-Earth and D-Earth don't actually come into contact in this one-person case guards against distracting discussions of dialectical convergence. Finally, *Alien Nurse* doesn't require assessing disagreement with rival linguistic communities instead

relying purely on first-order moral judgments thus blocking bullet-biting objections which question our metasemantic competence, and objections like Merli's (2002) which question our ability to make sophisticated linguistic judgments about whether Jim and Joe's disagreement is centered on 'morally wrong' or somewhere else. This means that even if you found the replies discussed in the previous sections indecisive, NMR remains in trouble insofar as it continues to rely on the causal theory of regulation. The argument from *Alien Nurse* sidesteps objections to standard MTE cases and argues more directly against the causal theory by showing that it generates implausible truth conditions.

Empathic Representation Instead of Causal Regulation

At this point, it seems that the problem might be NMR's refusal to admit a certain asymmetry between natural terms and moral terms. The causal theory of reference works fine for natural kinds like water as demonstrated in Hilary Putnam's (1973) original Twin Earth thought experiment. In the original thought experiment, we find out that while 'water' on Earth is causally regulated by the substance H20, 'water' on Twin Earth is causally regulated by the substance XYZ. This means that in Putnam's case, Earthian and Twin Earthian theories about water do not disagree. The problem for NMR is that this non-disagreement is fine for 'water' but not for moral terms like 'good'. The defenses of NMR discussed so far all try to preserve the appropriateness of applying the causal theory to moral properties, and if I'm right, none of them do so satisfactorily.

Fortunately for NMR, experientialism offers an alternative moral semantics that provides a direct answer to MTE cases (both standard and one-person). Recall that experientialism understands moral concepts in terms of the objective states of affairs accurately

represented by the moral feelings that they cause, with 'feelings' here referring to the experienced phenomenological part of emotion. With this theory of moral concepts, the concept 'good' might be said to apply to those states of affairs objectively represented by accurate horror or sorrow, the concept 'bad' to those states of affairs objectively represented by accurate guilt and indignation, and the concept 'wrong' to those states of affairs objective represented by accurate guilt and indignation.

Under the experientialist framework, Jim and Joe's disagreement about whether lying is always wrong does not concern what natural properties causally regulate their use of moral terms. Rather, their disagreement concerns whether lying is always accurately represented by feelings of guilt and indignation. Experientialism is thus able to provide a way out of the dialectical stalemate that the causal theory leads us to. If Jim and Joe share the ability to have moral feelings and if moral concepts apply to things that are accurately represented by these feelings, then their moral terms can share meaning and reference allowing for genuine disagreement about moral issues.

However, our discussion so far leaves open the question of the correct metaethical position to take as regards the nature of moral properties. For example, one might be an experientialist and an error theorist by regarding our moral feelings as representing things that do not actually exist, or one might be an experientialist and also think that moral properties are non-natural by holding that there is some sort of non-natural relation between our feelings and their objects. To solve MTE as a challenge to NMR, a naturalist friendly account of how moral feelings represent moral facts is required. Neil Sinhababu

(2016) provides just such an account in 'Edenic Representation of Pleasure Solves Moral Twin Earth', which I will now summarize.

'Edenic representation' references an idea about visual perception from David Chalmers' (2006) 'Perception and the Fall from Eden'. In this paper, Chalmers characterizes the Edenic world as one "populated by perfect colours and shapes, with objects and properties that are revealed to us directly" (51); "In the Garden of Eden, we had unmediated contact with the world. We were directly acquainted with objects in the world and with their properties. Objects were presented to us without causal mediation, and properties were revealed to us in their true intrinsic glory" (50). We fell out of Eden after eating from the Tree of Illusion (i.e. our realization that "objects sometimes seemed to have different colors and shapes at different times, even though there was reason to believe that the object itself had not changed" (50)) and the Tree of Science (i.e. our realization "that when we see an object, there is always a causal chain involving the transmission of light from the object to the retina, and the transmission of electrical activity from the retina to the brain" (50-51)). Chalmers goes on to argue that even though we no longer live in Eden, Eden still acts as a regulative ideal for the content of our perceptual experience.

While we might no longer be in an Edenic world with regard to visual perception, Sinhababu argues that this need not be the case for moral properties. This is his account of representation that solves MTE:

> "[M]oral feelings represent moral properties partly by sharing aspects of their phenomenal character. The pleasant feeling of hope represents pleasure, while the unpleasant feelings of horror and sadness represent displeasure….moral feelings represent hedonic elements of

their own phenomenal character as being objectively instantiated in reality. They don't rigidly designate whatever causally regulates them, apply to whatever disposes perceivers like us to have them, or represent non-natural moral properties. Instead, moral feelings represent reality as having hedonic character isomorphic to their own -- one might say, as being isohedonic to themselves" (12).

Sinhababu calls this account 'empathic representation'. Empathic representation is Edenic for it gives us the intrinsic nature of moral facts by having moral feelings. When we hope for a good outcome our experience represents and is isohedonic with the way a good outcome would be. When we feel guilty about a wrong action our experience represents and is isohedonic with the unhappiness we think our action would cause. If moral feelings empathically represent hedonic tone, then it follows that some form of ethical hedonism is true. While moral feelings definitely consist of more than hedonic tone, nothing else in the phenomenal character of moral feelings supports empathic representation in the unified way that hedonic tone represents pleasure's moral value. As Sinhababu puts it: "Ethical hedonism allows hope to represent pleasure, and allows the other pleasant moral feelings to represent relations to increases in pleasure. It allows horror to represent decreases in pleasure and the other unpleasant moral feelings to represent relations to decreases in pleasure. So experientialism, empathic representation, and ethical hedonism combine into a systematic theory of moral concepts, moral semantics, and normative ethics" (14).

Experientialism gives Jim and Joe a way to share moral concepts, empathic representation gives us a theory of representation that is friendly to NMR, and ethical hedonism gives us a way to decide who is correct. Accurate representations match the world. Since it is not the

case that lying always produces an unpleasurable state of affairs, it is not the case that lying is always accurately represented by feelings of guilt and indignation. Thus, Joe is right and Jim is wrong, and lying is morally permitted in cases where it would give rise to more pleasure than displeasure. The same goes for one-person cases like *Alien Nurse*. While knowing that some moral theory causally regulates the use of moral language in your linguistic community does not settle the issue of whether it's wrong to push the large man, empathic representation does so. Experientialism and empathic representation save NMR from the problem of non-disagreement by providing an alternative moral semantics and theory of representation.

One challenge to Sinhababu's view is that some moral feelings like pride and hatred aren't always so straightforwardly related to pleasure. Even hedonists will agree that right actions, which we correctly take pride in, are not always pleasant. Increasing the pleasure of others might require suffering on one's own part. Likewise, villains might enjoy causing suffering, and we might hate them for taking pleasure in the pain they produce. In both these cases, the immediate objects of our attitudes (i.e. the painful right action and the happy villain) may not share the moral feelings' hedonic character (i.e. the positive hedonic tone of pride and the negative hedonic tone of hatred), even if the states of affairs they regard do. This kind of complexity is to be expected if moral representation is analogous with visual representation. I can see images of objects reflected in water, or after having been transmitted by electromagnetic waves on my television. Such representations are usually accurate if one is aware of their provenance. I can have a visually accurate representation of the moon by seeing it reflected on water or broadcasted on TV. Similarly, if the rightness of an action is reflected by the value of its consequences, then the hedonic tone of pride and

hatred can accurately represent the pleasure caused by right action and the suffering caused by villainy.

Another challenge arises from the fact that feelings like hope, pride, and hatred don't always represent moral properties. Hoping that a paper gets a good grade may not involve seeing good grades as morally valuable, and might feel the same as hoping for world peace which represents world peace as morally valuable. Even though both instances of hoping have the same phenomenal character, their representational content differs in an important way. What this shows is that phenomenal character doesn't fully determine representational content. In response to this challenge, Sinhababu says that he is "content to allow for impure combinations of empathic representation with descriptive factors" (16). This allowance is not detrimental to his theory, since all reasonable theories of reference are similarly impure. Take causal regulation for instance, even if 'centaur' is causally regulated by human upper bodies and horsey lower bodies, the existence of these things do not entail that centaurs actually exist as detached aggregates of humans and horses. Further descriptive content stipulates that a centaur is one animal that has both a human upper body and a horse's lower body. Empathic representation also uses this strategy; one might say that descriptive content in hoping for world peace represents its value as objective, while descriptive content in hoping for a good grade represents its value as subjective (ibid), and that it is this difference in descriptive content that explains the difference in representational content.

Rival Theories on MTE

This section will discuss MTE as it relates to rival theories brought up in the previous chapter, and show that experientialism is uniquely positioned to handle MTE. I will start the discussion with Jesse Prinz's (2015) sentimentalism, followed by Victor Kumar's (2015) hybrid theory, and end with Heidi Maibom's (2005) rationalism.

Jesse Prinz (2015) is a relativist about moral judgment. According to Prinz, there are two ways sentimentalists might approach the question of whether moral properties are objective. The first is to say that our moral sentiments track objective moral truths. This he calls the 'moral sense theory'. The second is to say that moral judgments refer to response-dependent properties. This he calls the 'sensibility theory'. Prinz then argues for the sensibility theory over the moral sense theory:

> "I think there is some reason to favor sensibility over moral sense. For the moral sense theory to be true, there would have to be a candidate objective property to which our moral concepts could refer. Unfortunately, I cannot undertake a review of modern moral sense theories here, but I will offer, instead, a more general line of empirically-informed resistance. Moral rules are emotionally conditioned, and communities condition people to avoid a wide range of different behaviors. Within a given society, the range of things that we learn to condemn is remarkably varied. Examples include physical harm, theft, unfair distributions, neglect, disrespect, selfishness, self-destruction, insults, harassment, privacy invasions, indecent exposure, and sex with the wrong partners (children, animals, relatives, people who are married to other people). One might think that all of these wrongs have a common underlying essence. For example, one might propose that each involves a form of harm. But this is simply not true. Empirical evidence shows that people condemn actions that have no victims, such as consensual sex between adult siblings and eating the bodies of people who die in accidents (Murphy et al. 2000). Furthermore, harm itself is a subjective construct. It cannot be reduced to something like physical injury. Privacy violations are regarded as a kind of harm, even though they don't hurt or threaten health, whereas manual labor is not considered a harm, but it threatens the body more than, say, theft. Similar problems arise if we try to define moral wrongs in terms of autonomy violations. Mandatory education violates autonomy, but it is considered good, and consensual incest is an expression of autonomy, but is considered bad" (21-22).

In other words, Prinz thinks that because the range of things people express moral condemnation of is incredibly varied, it's difficult for moral sense theory to provide a candidate objective property to which our moral concepts could refer, as there doesn't seem to be a shared underlying essence across the range of things that we are emotionally conditioned to condemned. He then argues by inference to the best explanation that sensibility theory is true, and that "the property of being wrong is the property of causing negative sentiments, not a response-independent property that those sentiments are designed to detect" (22).

The MTE thought experiment generates intuitions that are opposed to relativism; MTE is a problem for NMR because we think that Jim and Joe are in disagreement with each other, and that this disagreement is one that concerns objective properties. If Prinz is correct and "[the truth of] moral judgments depends on our sentiments" (22), then divergent responses like Jim and Joe's can both be true at the same time. This is exactly the problem discussed for the causal theory of regulation; just like the causal theory, relativism leads us either to the problem of non-disagreement, or the problem of a dialectical stalemate where dissenting sides have equal claim to the truth. Prinz's solution to MTE would be to abandon NMR in favor of his sentimental relativism, which is a solution moral realists would find unappealing.

It's fortunate then that experientialism provides the realist both with an answer to MTE and an answer to Prinz. While Prinz comes close to empathic representation by saying that "the property of being wrong is the property of causing negative sentiments", he fails to see how this might be made consistent with moral realism. Prinz is certainly right in saying that there

is a lot of variation in what people find morally wrong, but with the account of empathic representation presented in the previous section, we can see how this variation need not lead us to relativism. With a criterion for accuracy, empathic representation is able to account for variation while giving us objectivity. For example, the claim that privacy violation is morally wrong is one that represents privacy violation with a negative hedonic tone, but this representation is only accurate if privacy violation in fact causes displeasure out there in the world. Whether privacy violation in fact causes displeasure is dependent on context and culture, but the displeasure it might cause is an objective feature of the world that is independent of an individual's response. The experientialist would thus take Prinz's claim that "the property of being wrong is the property of causing negative sentiments" and restate it as "the property of *seeming* wrong is the property of causing negative sentiments", adding that what *seems* wrong need not necessarily *be* wrong.

Victor Kumar (2015) thinks research done on the moral/ conventional distinction shows that we should treat moral judgment as a homeostatic property cluster with a paradigmatic moral wrong being a wrong that is (1.) serious, (2.) general, (3.) authority-independent, and (4.) objective (2896). Kumar also thinks that this theory of moral judgment offers a unique account for the possibility of genuine moral disagreements in MTE type cases:

> "In moral disagreement two people must have attitudes that oppose one another. But to disagree, or even simply agree, they must also conceive of the issue at play as moral. Otherwise accord and discord are not distinctively moral…Genuinely moral agreement and disagreement require shared moral concepts. As I explained [earlier], the moral/conventional distinction is universal or at least close to universal. Individuals from a wide range of groups exhibit the standard pattern of responses in the moral/conventional task, including adults from many different cultures, as well as children from normal and abnormal populations. Thus, MCT explains why people not just from the same local group but from many different demographic groups genuinely agree and disagree when they voice moral opinions, and thus are not merely talking past one another. They have a shared concept of morality—as (prototypically) serious, general, authority-independent, and objective. Thus, when

business owners in one country and workers in another country disagree about, say, the morality of factory conditions, they disagree— genuinely—about whether correct social standards that have the four features permit or forbid the conditions" (2904).

In other words, Kumar's definition of moral judgment besides being empirically supported is also able to account for genuine moral disagreement. Since MTEarthians are stipulated to be largely similar to humans with the exception of their moral consensus, it's highly plausible that our standard pattern of responses in the moral/ conventional task should also extend to them. If that is the case, then Kumar's definition of moral judgment might account for genuine moral disagreement between humans and MTEarthians by having them share a concept of moral as serious, general, authority-independent, and objective. So far, this is consistent with the truth of experientialism and empathic representation; an experientialist might say that moral feelings represent moral wrongs with a negative hedonic tone, and also as wrongs with the four features Kumar identifies. Where experientialism disagrees with Kumar lies in his answer to another metaethical question. Namely, the question of what type of attitude is constitutive of moral judgment.

Kumar (2016b) argues in 'The Empirical Identity of Moral Judgment' that moral judgment is "a hybrid psychological state of moral belief and moral emotion- that is, a belief with moral content and a moral emotion like resentment, guilt, sympathy, outrage, repugnance, etc" (789). Kumar's argument that moral judgment is a hybrid state is distinct from his argument that moral judgment is a homeostatic property cluster, though both are rooted in his treatment of moral judgment as a natural kind. According to Kumar, moral judgment qualifies as a natural kind because it plays a causal/explanatory role in a range of different behaviors and domains of reasoning. Both his conceptual account of moral judgment and his

theory of moral judgment as a hybrid state is meant to be supportive of moral judgment's causal/explanatory role in reasoning and behavior.

Kumar argues for his hybrid theory by considering 'conflict cases' where moral reasoning and moral action diverge:

> "For example, Jonathan Bennett (1974) finds a moral conflict case in The Adventures of Huckleberry Finn. Huck has helped Jim escape his life as a slave, and the two are paddling a raft down the Mississippi river. But then Huck reasons (speciously) that Jim is the property of Miss Watson, that in helping Jim escape he is acting as a thief, and therefore that he ought to hand him over to the authorities. This conclusion leads Huck to doubt his course of action and deliberate with himself about what to do. Later on, however, when the chance arrives to give Jim up, Huck cannot bring himself to do so. He feels too much sympathy for Jim and his plight, and thus, he acts to save Jim from slave catchers" (790)

In this conflict case, Huck's reasoning leads to him believing that he should not save Jim, but Huck acts as if he judges that he should save Jim. From such cases, Kumar surmises that "when moral reasoning and moral action diverge, it seems to be belief and emotion that are at the root of the conflict" (ibid). Huck believes that it is wrong to help Jim, but his sympathy for Jim motivates Huck to help him. Kumar thus concludes that "if we want to explain how individuals in conflict cases reason, we should appeal to their moral beliefs. If we want to explain how they act, we should appeal to their moral emotions…moral beliefs support the role of moral judgement in reasoning, while moral emotions support the role of moral judgement in action" (791).

Kumar's hybrid theory thus provides a natural explanation for conflict cases. The beliefs and emotions that comprise our moral judgments are usually in agreement, and that is why our

moral actions are typically indicative of our moral judgments. In conflict cases, the two come apart revealing the attitudes that constitute moral judgment. Kumar goes on to explain how moral emotions and moral beliefs dissociate by appealing to a 'minimalist' dual process model of the mind, which says that moral judgments are generated by two types of processes: type 1 processes, which are fast, unconscious, and involve emotional processing, and type 2 processes, which are slow, conscious, and involve reasoned processing. An in-depth discussion of this model is not necessary for our purposes. Suffice it to say that Kumar thinks this model well-supported by empirical evidence and that "belief and emotion support the causal/explanatory role of moral judgement in a dual process model of moral cognition, including conflict cases" (792). The hybrid theory conflicts with experientialism, which sees moral judgment "as belief alone, rather than belief plus emotion" (Sinhababu 2017, 67). I will argue in favor of experientialism's purely cognitive theory of moral judgment by first showing how it is able to provide a better account of conflict cases. I will then consider the problems faced by the moral semantics implied by hybrid theories like Kumar's.

Experientialism says that moral judgments are beliefs typically caused by moral feelings. Huck's belief that it is wrong to help Jim is caused by his moral feelings representing the act of helping Jim with a negative emotion like guilt or regret. The fact that he nonetheless helps Jim might be attributed to another belief that you should do your best to help those in need. According to experientialism, this latter belief is also caused by Huck's moral feelings, which represents helping those in need with positive emotions like pride or admiration. The emotional perception model has moral emotions consisting of moral feelings and desires. In conflict cases, we have two sets of moral feelings and desires pointing in different directions. For example, with Huck we have one set representing the act of helping Jim with a negative moral feeling accompanied with a desire to do the right thing by not helping Jim, while the

other set represents helping Jim with a positive moral feeling accompanied with the desire to do the right thing by helping Jim. When Huck decides to help Jim, the latter set triumphs over the former.

Experientialism might thus agree with Kumar in saying that beliefs and emotions lie at the heart of conflict cases, but add that the conflict is between different sets of beliefs and emotions. Instead of saying that "moral beliefs support the role of moral judgement in reasoning, while moral emotions support the role of moral judgement in action", experientialism shows how moral emotions can play a role in our moral reasoning. Furthermore, by talking about conflict cases as being between sets of beliefs and emotions, experientialism does better than hybrid theory in preserving the complex and multi-layered nature of moral conflict; in conflict cases, there are often different beliefs and desires in play, and it is hardly ever so simple as being a case of one set of beliefs versus one set of emotions.

Kumar's hybrid theory also implies a hybrid semantics that is problematic for NMR. It follows naturally from the hybrid theory that a speaker's utterance of "X is wrong" expresses both the cognitive attitude that 'X is wrong' and some non-cognitive attitude such as 'moral subscription to a standard that morally forbids X'. Let's say that the case of MTE is one of 'fundamental moral disagreement' where moral disagreement remains despite agreement on all the non-moral facts. The hybrid semantics implied by Kumar's theory allows for a deeper kind of moral disagreement where disagreement persists despite agreement on all the moral and non-moral facts. Michael Rubin (2015) calls disagreement of this latter sort 'radical moral disagreement', which is realized "when the interlocutors' moral predicates are

R-related to different properties, yet they would still disagree in attitude, even if they recognized this fact" (701). What this means is that moral disagreement at its deepest level is disagreement in attitude, not belief, and that where moral disagreement is deepest, "the norms appropriate for a resolution are precisely the kinds of non-epistemic norms that anti-realist expressivists recommend for resolving all fundamental moral disagreements" (702).

Kumar's hybrid theory thus paves the way for anti-realism. An easy response would be for him to abandon the hybrid theory in favor of experientialism. Experientialism is able to better account for conflict cases and explain the tight connection between motivation and moral judgment while maintaining a purely cognitivist semantics. It is also consistent with Kumar's preferred dual process model of moral cognition, so long as he is willing to accept emotion's place in belief formation. With all these benefits and virtually no costs, experientialism seems to offer Kumar (and naturalistic moral realists in general) a deal that is too good to refuse.

Heidi Maibom (2005,2010) is a rationalist about moral judgment. While sentimentalists claim that moral emotions are central to the making of moral judgments, rationalists like Maibom claim that moral judgment is a matter of practical reason. As discussed in the previous chapter, it seems that a sentimentalist experientialism is better able to account for the moral failings of the psychopath than rationalism. I will end this chapter by considering the options a rationalist like Maibom has in the face of MTE.

If moral rationalism is true, then moral truths can be discovered through a purely rational procedure. Like facts about the chemical structure of water, we can know that things are right or wrong without feeling any way towards them[22]. This obviously runs counter to empathic representation and the moral semantics of experientialism, which are thus not available to the rationalist as a solution to MTE. The rationalist could fall back on the causal theory of regulation and maintain the symmetry between moral facts and scientific facts in this regard. They might do this by either challenging premises (1) or (2) of the MTE argument, or by challenging premise (3) and biting the bullet. However, as we've discussed at some length, neither of these strategies seem promising, especially in light of one-person MTE cases. Perhaps rationalists could fall back on Kumar's definition of moral judgment as a homeostatic property cluster. This would allow them to account for genuine moral disagreement between humans and MTEarthians by having them share a concept of moral judgment. I think this is the most promising strategy for naturalistic moral rationalists, but only when considered in isolation from the rest of Kumar's arguments. Recall here that the same research that informs Kumar's definition of moral judgment also leads him to settle on a species of sentimentalism. Furthermore, while Kumar's definition might account for genuine moral disagreement, it doesn't rule out a situation where dissenting sides maintain an equal claim to the truth. Without empathic representation, it seems that NMR is consigned either to a fate of interplanetary non-disagreement, or to the possibility of a dialectical stalemate. Just as it did with the problem of psychopathy, experientialism coupled with empathic representation provides the best answer to the problem of MTE.

[22] Note here that I'm not saying Maibom thinks that something like chemistry will reveal the moral truth, just that we can discover what things are right or wrong without feeling any way towards them.

Chapter 4: Experientialism and Rationalism

The previous chapters have argued that experientialism does better than both its sentimentalist and rationalist rivals in responding to issues related to psychopathic moral judgment and MTE arguments. This chapter will restrict its attention more or less exclusively to a comparative evaluation of experientialism, as articulated by Neil Sinhababu (2017), versus rationalism, as defended by Joshua May (2018). The first half of this chapter will argue that rationalism appears committed to a picture of moral motivation that is implausible from the point of view of human cognition. The second half will address rationalist worries that adopting sentimentalism would result in pessimism about ordinary moral thought and action.

Cognitivist Internalism

Under the Humean theory of motivation, action requires desire, and belief alone cannot generate desire. This means that belief alone cannot motivate action. Cognitivism says that moral judgments are beliefs, and internalism says that moral judgments motivate action. The cognitivist internalist view thus says that moral judgments are beliefs that have motivational force, and this is incompatible with Humeanism about moral motivation. This relates to a trilemma touched on briefly in the previous chapters; taken individually, cognitivism, internalism, and the Humean theory of motivation are attractive to many philosophers, but if all are true, then human beings are incapable of making moral judgments. A reasonable person would thus choose at most two of the three.

Sentimentalism is typically non-cognitivist and internalist. Non-cognitivism, simply put, is the denial of the cognitivist claim that moral judgments express propositions that can therefore

be true or false. The reason for sentimentalism's affinity with non-cognitivism is that if one thinks that our moral judgments are grounded in our moral emotions, it follows quite easily (but not necessarily) that our moral judgments express attitudes similar to feelings of approval and disapproval. Likewise for internalism- if our moral judgments are emotionally grounded and emotions motivate by virtue of having desire as a component, then it's unsurprising that moral judgments necessarily motivate action. Thus, of the three positions that comprise the trilemma, sentimentalism usually rejects cognitivism in favor of Humeanism and internalism.

Rationalism, on the other hand, is traditionally allied with cognitivism and opposed to the Humean theory. If our moral judgments are grounded in our moral beliefs as rationalism says they are, it follows naturally that they are truth-apt just like our non-moral beliefs. Since desires lie in the rationalist-shunned domain of emotion, rationalists are inclined to adopt an anti-Humean theory of motivation by which our moral beliefs can motivate action either directly with their own motivational force or indirectly by producing the relevant desire. This also naturally translates itself to an internalism distinct from that of sentimentalism's; while sentimentalist internalism typically holds that our moral judgments are non-cognitive and motivate by virtue of being grounded in emotion, rationalist internalism holds that our moral judgments are cognitive and are able to motivate without help from our emotions or pre-existing desires.

Probably the most attractive feature of internalism to both rationalism and sentimentalism is that it gives us a unified account of moral judgment (Sinhababu, 2017: 8). As G.E. Moore (1903) points out, even obviously false moral claims usually seem substantively false rather

than conceptually confused. This raises the question that if moral concepts are so conceptually open such that many false moral claims and theories still count as moral, what does 'being considered a moral claim' involve? Internalists like Allan Gibbard (1990) have an answer: motivation. According to this internalist answer, people might think that a vast range of different things qualify as morally right without conceptual confusion or contradiction, but a proper understanding of rightness involves being motivated to act rightly. In other words, if you think that a certain action is morally right but you lack any motivation to do it, you simply haven't grasped the concept of rightness (Sinhababu, 2017: 8).

Experientialism presents a unique sentimentalist way forward with its commitment to cognitivism and its rejection of internalism. As has been discussed, it is also able to give just as tidy a story of moral concepts as internalism does, without positing any sort of necessary connection between moral judgment and motivation. The merits of experientialism's uniqueness will be discussed at greater length in the conclusion of this thesis. The next section will discuss rationalism's commitment to an implausible theory of human motivation.

Motivating Beliefs

Human beings are typically motivated to act according to their moral judgments. This does not mean that we always actually do what we think is morally right. Rather, when we judge that some action is morally right, there is usually some motivation to act accordingly. For example, if I am a decent human being and I judge that lying is always morally wrong, I would be motivated to always tell the truth. I might tell lies some (or even much) of the time,

perhaps in cases where competing motivations trump my moral motivation to tell the truth (e.g. my motivation to please my mother by telling her that her cooking is good versus my motivation to do the right thing), but for the most part, thinking that something is morally right motivates me to act accordingly.

To this fact about human behavior the rationalist has two broad ways of responding. Either affirm a strong version of the internalist view that no person could sincerely judge an action to be morally right while remaining completely unmoved by said judgment, or a weakened formulation of internalism that says human beings can sometimes be motivated by their moral beliefs alone without an antecedent desire to do the right thing. As discussed, both formulations of internalism are at odds with the Humean theory of motivation. Having already argued against the first formulation in the second chapter, I will focus my attention on the second[23]. The Humean response is unattractive to rationalists because it claims that "reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them" (2.3.3). Rationalist theories of motivation are thus typically anti-Humean attempts to free reason from the passions. I will now go through three such attempts, and argue that all three fail at demonstrating the plausibility of motivational beliefs in human beings.

General Desires

One way of freeing reason from the passions is by showing that the Humean theory is false. This might be done by providing examples in which someone is motivated to act through

[23] Also, since the first formulation is committed to the claim that belief alone can sometimes motivate action (in addition to the claim that moral judgments necessarily motivate action), any argument against the second would likewise count against the first.

moral reasoning that doesn't begin with some antecedent desire. Stephen Darwall (1983) for example claims that someone can be "moved by awareness of some consideration, without that [motivation] being explained by a prior desire" (39). This contradicts the Humean claim that belief alone cannot generate a new desire. In support of this anti-Humean claim, Darwall presents the following case:

> "Roberta grows up comfortably in a small town. The newspapers she reads, what she sees on television, what she learns in school, and what she hears in conversation with family and friends present her with a congenial view of the world and her place in it. She is aware in a vague way that there is poverty and suffering somewhere, but sees no relation between it and her own life. On going to a university she sees a film that vividly presents the plight of textile workers in the southern United States: the high incidence of brown lung, low wages, and long history of employers undermining attempts of workers to organize a union, both violently and through other extralegal means. Roberta is shocked and dismayed by the suffering she sees. After the film there is a discussion of what the students might do to help alleviate the situation. It is suggested that they might actively work in promoting a boycott of the goods of one company that has been particularly flagrant in its illegal attempts to destroy the union. She decides to donate a few hours a week to distributing leaflets at local stores." (39)

From this, Darwall argues that Roberta's newfound knowledge of the textile workers' plight can motivate her to act even without "some such general desire as the desire to relieve suffering prior to seeing the film" (40). While the Humean theory treats Roberta's new desire to help the workers as instrumentally derived from a more general desire, Darwall thinks that that need not be the case, and that sometimes our normative beliefs can generate new desires without the help of any antecedent desire. He also provides a positive argument against the Humean account of Roberta's new desire, saying that since a desire "includes dispositions to think about its object [and] to inquire into whether there are conditions that enable its realization [i.e. desire's attentional aspect]" (40), then Roberta's lack of attention to the suffering of others is evidence that she lacked a general desire to relieve suffering prior to watching the film; if Roberta had such a general desire, then the

attentional aspect of this desire would direct her to think and inquire about how to relieve the suffering of others.

However, the Humean theory is able to handle the case of Roberta quite easily. As Sinhababu (2017) responds, “Roberta’s upbringing provides the answer” as to why she’s only moved to relieve suffering after watching the film (48). According to Darwall (1983), Roberta’s comfortable life leads to her having a “congenial view of the world and her place in it” (39)- “Stimuli that would activate her preexisting aversion to others' suffering are largely absent from her early environment” (Sinhababu 2017, 48). Her desire is not brought to the forefront of her mind since prior to watching the film, she is only vaguely aware of poverty and suffering. No vivid representations of suffering or plans about how she might act to ameliorate it have been presented to her. Her desire lies dormant, not yet activated by “neither a changing probability of satisfaction nor a vivid representation of its object to stir [her] thoughts (ibid). The case of Roberta is not an unusual one. Many of us have desires that though strong do not usually occupy a position of priority in our thoughts and drive us to inquire into means to promote their realization. For example, most of us strongly desire that our loved ones be safe from physical harm. However, we are usually quite satisfied to go about our lives without constantly thinking of ways to protect them from danger. So long as there is no changing probability of our loved ones getting hurt, or no vivid representation of harm befalling them, nothing activates this desire for their safety. The strength of this desire is only made evident when it is activated (e.g. by a vivid nightmare or by news that they are in danger) and by the intensity in which it would then drive our thoughts.

Ironically, the case of Roberta ends up supporting the Humean theory. Her shock and dismay upon seeing the film is predicted and explained by the Hedonic Aspect of desire. Unpleasant feelings usually accompany the realization that our desires will not be satisfied, especially if in our ignorance we did not worry about their satisfaction. As Al Mele (2003) argues, if Roberta never had a general desire to relieve suffering prior to watching the film, "what could explain her being shocked and dismayed by the suffering she sees?" (98). Note here a subtle distinction between 'a desire to relieve suffering in general' and 'a general desire to relieve suffering'. With the former, 'general' might be understood as 'being actually concerned with most instances of suffering', and this sort of 'general desire' might still be missing from Roberta even after she has watched the film (i.e. she might still not actually care about most instances of suffering and desire to lessen all of them). With the latter, 'general' might be understood as 'inexact' or 'vague', and this sort of 'general desire' is the one that would explain Roberta's shock and dismay. The information and vivid representations gleamed from the film would fill in the details of her vague desire to alleviate suffering, and motivate her to action she would otherwise not take if she had not watched the film. I think some confusion between these two senses of 'general' motivates Darwall's thought that Roberta might be motivated on her newfound knowledge without the help of an antecedent desire. Roberta obviously doesn't need a general desire in the first sense but if she lacks one in the second sense then it seems to me that she would not be shocked and dismayed by the suffering portrayed in the film. Darwall's counterexample thus fails to undermine the Humean theory, which is better placed to explain a key aspect of the case (i.e. Roberta's shock and dismay) and is able to do so without invoking the existence of special beliefs capable of generating desire.

Antecedent Desires V. Mere Disposition

Accordingly, one common criticism Humeans make of anti-Humeanism is that anti-Humeanism is at odds with a naturalistic account of human motivation. Al Mele (2003) for example argues for Humeanism by saying that unlike anti-Humeanism's need to invoke the existence of special beliefs, it is "not at all mysterious how a desire to A would derive some of its force from a relevant antecedent desire" (94). Likewise, Bernard Williams (1979) criticizes anti-Humeanism for having to "conceive in a special way the connexion between acquiring a motivation and coming to believe [some statement] of reason" (108). A second way of freeing reason from the passions is thus to find a way to reconcile the motivational power of moral beliefs with a scientific account of human action.

Joshua May (2018) proposes just such a way, saying that "the best anti-Humean theory merely posits a disposition for normative beliefs to generate the corresponding desires" and does not "require positing special mechanisms in human motivation": "while a strong-willed person, for example, may lack the antecedent desire to throw her pack of cigarettes away, she may simply have a disposition to do so if she believes it's best to trash them" (189). May thinks that while Humeans might be "tempted to count such dispositions as desires…we shouldn't broaden [our conception of desires so that it applies] to any mere disposition of a person to do something" (ibid). Adapting an example from Darwall (1983: 40), May points out that a person might be disposed to eat a piece of pie without this disposition constituting a desire for some pie. Moreover, May argues that Humeans should not want to count mere dispositions as desires, for doing so renders them "unable to provide their characteristic explanation of a person's moral belief as promoting an antecedent desire" (May 2018, 189). He provides the following example to illustrate this point:

> "Simone believes she ought to hold her tongue and refrain from insulting her foolish coworker. Furthermore, she doesn't have, or isn't in this case, motivated by an antecedent

goal to do whatever is right. She merely has a disposition to desire to do something after coming to believe it's right. Humeans could call this disposition a "desire" but then the explanation is not a Humean one, since a mere disposition lacks the specification of anything like a goal that can then be served or furthered by the subsequent desire. Such explanations are part and parcel of the Humean theory" (ibid).

May is certainly right that we should not count just any disposition as a desire, but he doesn't go further in characterizing 'mere disposition' beyond saying that Humeans should not want to count them as desires because they do not specify goals. It seems to me that in the example provided by May, Simone's "disposition to desire to do something after coming to believe it's right" can be straightforwardly reduced to or explained by some general desire to do the right thing. Just as with Darwall's Roberta case, some confusion here might stem from the subtle distinction between 'a desire to do what is right generally' and 'a general desire to do what is right'. Simone might not possess or be motivated by the former desire to do all (or most of) the right things, but it's likely that she possesses the latter general desire to do right things (though this general desire is vague and inexact). If she did not possess the latter desire, her disposition to be motivated to do something after coming to believe it's right is unexplained and mysterious, unless one posits, as May does, the existence of yet another mysterious disposition. Desires, beliefs, and so on are psychological dispositions. We can explain Simone's motivation to hold her tongue by positing an antecedent general desire to do what is right, and it is likely that Simone herself would affirm that she possesses this general desire. In positing "a disposition for normative beliefs to generate the corresponding desires" (ibid), May is taking on an additional psychological commitment to a disposition that he fails to elaborate on. The obvious question here is: Why should we posit that type of disposition when we already have desires to work with? Instead

of positing a special mechanism in human motivation, May posits a special psychological disposition. Unfortunately for May, Occam's razor cuts against both.

Parsimony

At this point, someone sympathetic to rationalism might ask whether it's appropriate to apply Occam's razor to psychological theory, and whether doing so incontrovertibly declares the Humean theory the winner. A third way of freeing reason from the passions is to show either (1) that parsimony isn't necessarily a virtue of empirical theories (especially in the case of human psychology), or (2) that the Humean theory is not in fact any more parsimonious than the best rationalist theory. May (2018) tries to do both, and I will reply to each respectively.

Regarding (1), May writes:

> "Parsimony isn't an uncontroversial virtue of empirical theories, for it alone only increases the probability of a hypothesis in rather specific conditions (Sober 2015). In the particular case of motivation, there is special reason to worry about staking one's account solely on simplicity. The history of psychological theory has shown a trend in the proliferation of moving parts, such as types of mental states, processes, or modules…Consider memory as an example (cf. Holton 2009: xii-xiii). Rather than develop a unified conception of memory, psychologists have posited quite distinct kinds with rather different functions…Similarly, Humeans do not shy away from distinguishing different types of desires…we should in advance expect the architecture of our evolved minds to be rather disjointed and modulated rather than simple and elegant. While the Razor might still be of some value, we might at least bet that its role in psychological theorizing will be limited" (196).

As May mentions, Humeans do not shy away from distinguishing different types of desires. Also, and as discussed in the first chapter of this thesis, Sinhababu's account of desire is one

that is richer than most. This goes to show that judicious users of Occam's razor do not value simplicity for its own sake, at the cost of good explanations that respect the empirical data. We might indeed expect the architecture of human minds to be disjointed and modulated, but this expectation does not entail that we should postulate the existence of dispositions which we have no evidence for, and which don't do any further explanatory work. The Humean account of motivation does not stake its claim solely on simplicity. If it did, it would not be a very interesting or useful theory. Rather, the Humean claim is that we only need to appeal to desire and belief in providing good explanations regarding human motivation and practical reasoning. In other words, the Humean theory stakes its claim both on simplicity, *and* on the grounds that it's able to explain all the relevant phenomena by appealing to well-defined and familiar psychological dispositions.

Regarding (2), May writes:

> "It's unclear whether the Humean theory is in fact any more parsimonious…On the Humean theory, motivational relationships only arise between one mental state and a desire, and the latter must initiate the process…Anti-Humeans just allow more than desires to initiate a motivational relationship between two states…Sometimes we do seem to distinguish processes based on what's related…[but] many processes are counted the same while relating different things so long as they aren't importantly different. For example, we don't posit two kinds of baking or two kinds of corrosion just because the relationship can hold between different entities. A human or a robot can bake a cake (or a quiche); water or acid can corrode a pipe (or a rock)… We needn't posit two kinds of motivational process just because one is initiated by a desire while the other is initiated by a belief" (197)

May is correct in saying that the Humean theory differentiates instrumental desire generation from anti-Humean desire generation based on their relata; with instrumental reasoning, the relata are an antecedent desire for something, a belief about a means to get it, and a new desire for that means, while with anti-Humean desire generation, the relata

are a normative belief and a new desire generated by this normative belief. Individuating these processes by their relata allows the Humean to argue that their theory is simpler because it needn't posit a mental process that anti-Humeans do. In the passage quoted above, May argues that we don't always individuate processes by their relata, and that we needn't do so in the case of instrumental reasoning versus anti-Humean desire generation. The upshot of this argument is that May's anti-Humean theory isn't less parsimonious than the Humean theory, since anti-Humean desire generation and instrumental reasoning can be treated as the same process initiated by different entities. Giving up the assumption that processes should be individuated by their relata allows May the claim that the Humean theory isn't any more parsimonious than opposing views.

On a related note, May writes that

> "[The above point] is amplified when we consider the fact that whatever anti-Humeans say here, Humeans must say something quite similar. The only difference is that, instead of a disposition, Humeans posit a full-blown desire. While the motivation attached to such a desire is perfectly explicable (since desires are by hypothesis motivational states), the fact that it appears in the individuals it does would be mysterious unless the Humean holds…that it's partly constitutive of their rationality or good character to possess such antecedent desires. But this isn't importantly different from the anti-Humean claim that it's partly constitutive of being a rational, virtuous, or strong-willed person that one possesses the disposition to desire in accordance with one's normative beliefs." (190)

One might respond that May's examples involving baking and corrosion do not provide good analogies for the psychological processes being discussed, because the relata May mentions that fail to distinguish processes (i.e. human vs robot baking and water vs acid corrosion) aren't essential to these processes. As Sinhababu (forthcoming) writes in a commentary on

May's book, "The reason why we might not divide up baking into separate processes depending on whether the baker is a human or a robot…is that [this isn't] essential to characterizing baking. What makes something an instance of baking…are a general way of applying heat and general sorts of effects on the food, not the identity of the baker or precise nature of the dish…once we're sufficiently precise about the nature of the processes, we see that we do individuate them by their relata" (4).

Also, while May might be correct in saying that the only difference between him and the Humean in these cases is that "instead of a disposition, Humeans posit a full-blown desire" (190), this is of far bigger importance than May accords, and it is what the debate between May and Humeans like Sinhababu ultimately boils down to. Firstly, and as discussed earlier, while Humeans are working with a well-defined and familiar mental state, May fails to adequately elaborate on the characteristics of his "mere disposition". Secondly, and perhaps more importantly, Humeans can hold both "that it's partly constitutive of [an individual's] rationality or good character to possess [the relevant] antecedent desires" and also "that it's partly constitutive of being a rational, virtuous, or strong-willed person that one possesses the disposition to desire in accordance with one's normative beliefs". It is misleading to label the former a Humean claim and the latter an anti-Humean one. A more accurate characterization of the difference between Humeans and anti-Humeans is that while Humeans insist that the disposition to desire in accordance with one's normative beliefs is reducible to possessing the relevant antecedent desires, anti-Humeans think that the disposition is sui generis of normative beliefs. The debate between the Humean and the anti-Humean centers around whether we should believe in the existence of this disposition.

To illustrate with another example, consider two zoologists arguing about whether unicorns exist. The first zoologist says “it seems unlikely that unicorns exist because if they did, we’d most likely have seen one by now. The most parsimonious explanation for our never having seen unicorns is that they don’t exist”. The second zoologist says “you would agree with me that horses exist. If we carve up the horse category more broadly to include unicorns, then my theory that unicorns exist is as parsimonious as yours. What we’re saying is hardly any different!” What May is saying is akin to the second zoologist. It might be entirely possible to talk at a general level of explanation about instrumental reasoning and anti-Humean desire generation as cases of the same general process of reasoning (just as it would be entirely possible to talk at a general horsey level about regular horses and unicorns as cases of the same species), but doing so takes May away from the level on which he is debating Sinhababu and other Humeans (as it does the second zoologist from his debate with the first zoologist about the existence of unicorns), and provides no further reason to think that humans are capable of anti-Humean desire generation.

I will conclude this section by considering a question analogous to the question of what makes a mental state motivational. That is, the question of what makes an animal a koala. A bad answer to this question is that a koala is a mouse that grew very large by eating eucalyptus leaves. The problems with this answer are analogous to the problems with the anti-Humean claim that normative beliefs are beliefs that gained motivational force because of their evaluative content. Just as animals in general do not grow unusually large because they eat special foods, mental states do not in general acquire new functional properties because they have special content. Rather, what mental states (like beliefs and desires) do can be generalized across many possible kinds of content, just as the effects of eating on a species can be characterized in a general way across many possible types of food. If one

wants to argue that koalas are special types of mice that grow very large because they consumed unusual foods, one should (at minimum) be able to point to other instances of animals that grow very large from eating unusual foods. Similarly, anti-Humeans who claim that normative beliefs are motivational (or able to create new desires without the help of already existing ones) should be able to point to other cases where mental states gain new functional abilities by virtue of having unusual contents. Doing so will make it more empirically plausible that special contents can bestow special functional properties on mental states. So far, anti-Humeans have not done this, most likely because there are no such cases to begin with. At the end of the day, it seems anti-Humeans are no closer to demonstrating the plausibility of motivational beliefs in human beings than the second zoologist is at demonstrating the existence of unicorns.

Moral Pessimism

If my arguments so far are correct, then rationalism fails to free reason from the passions and is committed to a picture of moral motivation that is implausible from the point of view of human cognition. What then is the attraction of rationalism that intelligent people try so hard to defend it against the tide of naturalistic viability? In what follows, I will first run through an instance of a popular answer to this question from Joshua May (2018): that without rationalism we are doomed to a pessimism about moral cognition. Following which, I will argue that the truth of rationalism would carry with it its own reasons for pessimism. Finally, I will explain how experientialism is able to keep many of the perceived benefits of rationalism, while at the same time giving emotion a central role in the making of moral judgment.

Rationalist Worries

May (2018) identifies the empirical trend towards sentimentalism as being a prime source of pessimism about moral cognition. He sees those who contend that rational processes do not ultimately drive moral thought and action as pessimists who view the pursuit of moral truth as an enterprise doomed to failure. May writes:

> "Many philosophers and scientists argue that our moral minds are grounded primarily in mere feelings, not rational principles. Emotions, such as disgust, appear to play a significant role in our propensities toward racism, sexism, homophobia, and other discriminatory actions and attitudes. Scientists have been increasingly suggesting that much, if not all, of our ordinary moral thinking is different only in degree, not in kind. Even rather reflective people are fundamentally driven by emotional reactions, using reasoning only to concoct illusory justifications after the fact. As Jonathan Haidt has put it, "the emotions are in fact in charge of the temple of morality" while "moral reasoning is really just a servant masquerading as the high priest" (2003: 852)." (3)

> "This is the challenge from a brand of *sentimentalism* which contends that moral cognition is fundamentally driven by emotion, passion, or sentiment that is distinct from reason (e.g., Nichols 2004; Prinz 2007). Many now take the science to vindicate sentimentalism and Hume's famous derogation of reason. Frans de Waal, for example, urges us to "anchor morality in the so-called sentiments, a view that fits well with evolutionary theory, modern neuroscience, and the behavior of our primate relatives" (2009: 9). Even if reasoning plays some role in ordinary moral judgment, the idea is that sentiment runs the show (Haidt 2012: 77; Prinz 2016: 65)." (7)

Though May acknowledges that "emotions aren't necessarily illicit influences" (7), he worries that the sentimentalist claim that genuine moral cognition requires having certain moral feelings and desires makes morality a fundamentally "arational enterprise in which reason is a slave to the passions" (ibid); May thinks that if our ordinary moral cognition is not a fundamentally rational enterprise, then we won't be able to rely on our basic modes of moral thought and motivation to know right from wrong and to act virtuously. In a bid to save moral cognition from the spectre of sentimentalism, May tries to argue for "an empirically informed rationalism" where "moral judgment is fundamentally an inferential

enterprise that is not ultimately dependent on non-rational emotions, sentiments, or passions" (ibid). I've argued that May's attempts to free reason from the passions are unconvincing. I will now give further reason to think that May's brand of rationalism carries with it its own set of worries about ordinary moral cognition.

Rationalist Pessimism

It seems likely that May's rationalism would result in an (1) unseemly elitism about moral virtue and a (2) mistaken diagnosis of widespread irrationality.

Regarding (1), if ordinary moral cognition is a fundamentally rational enterprise, then someone who has the wrong moral beliefs would never be able to do the right thing intentionally. Take the case of Huckleberry Finn as discussed in the previous chapter for example. Huck believes wrongly that Jim is the property of Miss Watson and that in helping Jim escape he would be acting as a thief. Yet, despite his wrong beliefs, Huck, moved by sympathy for Jim's plight, does the right thing by helping Jim escape. On the face of it, Huck seems to do the virtuous thing intentionally. However, given the rationalist schema, Huck is acting irrationally by helping Jim to escape. While he might have done the right thing, the source of his right action is mere feeling, not rational principle. If ordinary moral cognition is a fundamentally rational enterprise, then someone like Huck would not be able to act virtuously, insofar as they possess the wrong moral beliefs.

But what if we viewed Huck's case as one where distinct sets of beliefs and emotions compete, instead of one where emotions trump reason? After all, that is how I argued the

experientialist would interpret most cases of moral conflict. Unfortunately for rationalism, this interpretation of cases like Huck's only pushes back the problem. Let's say that Huck has two competing beliefs in this situation. The belief that helping Jim is wrong because it would be like stealing, and the belief that helping Jim is right because one should generally help those in need. Each belief is accompanied by the relevant moral emotion. Huck plausibly has an equal credence in both beliefs. How then is he moved in favor of one and not the other? The obvious answer would be that he felt more strongly regarding one of his moral beliefs. One might also say that even if Huck has a higher credence in the former belief (and lacks a strong desire to act rationally), he might still be moved by his emotions to act in accordance with the latter. If something like this is right, then this interpretation of moral conflict tells against the rationalist claim that moral cognition is a fundamentally rational enterprise. Emotion and not credence explains Huck's ostensibly virtuous actions.

A rationalist might then argue that even though Huck seems to do the virtuous thing, he in fact fails to act rightly because he fails to reason properly. Taking their cue from Kant, many contemporary rationalists (e.g. Nagel 1978, Korsgaard 2008) think that an action qualifies as moral only if it was motivated by reflective deliberation or considerations of duty (rather than by love or sympathy as in Huck's case). I think there is reason for us to resist this view, as it makes virtuous action overly dependent on the possession of moral knowledge, and disqualifies the morally mistaken from doing the right thing intentionally. Just as with other kinds of knowledge, it is likely that those who possess moral knowledge would only come to it after a substantial amount of education. This puts rationalism at a high risk of falling into the elitist view that uneducated people or people with less time for reflective deliberation are less capable (or even incapable) of acting virtuously; moral philosophy professors would

be more capable of virtuous action than impoverished farmers who can't afford school, based solely on the fact that they spend more time deliberating moral principles.

All this is not to say that moral inquiry is unimportant or doomed to failure. More will be said about this in the next section when I explain how experientialism is able to afford reason an important role in moral cognition, without allowing it to dominate emotion. For now, let's turn our attention to (2): May's rationalism will likely lead to a mistaken diagnosis of widespread irrationality. If some version of anti-Humean rationalism is true and normative beliefs were capable of directly changing our desires or creating new desires, then moral argument (on its own) would be able to change our aesthetic preferences, motivate us to give up hated addictions, or overcome tendencies to procrastinate on the internet[24]. Accordingly, rationalism either has it that normative beliefs necessarily motivate (i.e. motivational internalism), or that normative beliefs motivate in cases where the individual is strong-willed and rational. I will refer to the latter position as 'disjunctive internalism'. Once again, having already offered arguments against the former formulation of internalism in chapter 2, I will restrict my attention here to disjunctive internalism.

Disjunctive internalism, as formulated by Michael Smith (1994), is the view that "If an agent judges that it is right for her to ϕ in circumstances C, then either she is motivated to ϕ in C or she is practically irrational" (61). In other words, if an individual judges some action to be morally right and is not motivated to act accordingly, then they are behaving irrationally.

[24] An examiner of this thesis notes that "even anti-humean rationalism of a fairly strong sort need not include this. Even if one thinks that normative beliefs always generate desires, one needn't hold that those desires always outweigh other desires". The second sentence in the quoted remark is true; one needn't hold that those desires *always* outweigh other desires. But it would be strange if they never did. So long as rationalists are committed to the ability (*not* inevitability) of beliefs to motivate, my point holds.

This formulation doesn't entail that not being motivated by one's normative beliefs makes one irrational; it leaves open the possibility for one to be rational by being motivated by some other mental state, so long as that motivation is in line with one's normative judgments. Recall now that the truth of motivational internalism (even when weakly formulated as "moral judgments can produced their own motivational force") combined with the truth of the Humean theory and cognitivism jointly entail human incapabilism about moral judgment (i.e. the truth of all three means that humans are incapable of making moral judgments). The disjunctive formulation of internalism manages to avoid incapabilism in two ways. First, it allows for an individual to be motivated to act in accordance with their moral judgments by some other non-belief mental state. Second, an individual not motivated to act according to their moral judgments is irrational (Sinhababu 2017: 170). I'll now examine both options and show, with arguments adapted from Sinhababu (ibid), that internalists should not like the first option, while the truth of the second option, when combined with broader internalist commitments, means the irrationality of all humans who make moral judgments.

To see why internalist should not like the first option, consider the following example: Norman is an escaped convict who is on a date with Marion at a restaurant. He recognizes a fellow diner, Sam, as someone who works at the prison he's recently escaped. Norman knows that murdering Sam would be wrong, but is indifferent to this fact. He's also motivated not to murder Sam, but only because he's certain that Sam has not spotted him, and he does not want to ruin his date with Marion. As formulated, disjunctive internalism allows Norman to be making a genuinely moral judgment. He judges it wrong to murder Sam, and he's motivated not to murder Sam. I don't think internalists would be comfortable saying that Norman is a rational agent making a moral judgment simply because he has

some other motivation that is in line with it. As Steven Swartzer (2015) argues, formulations of internalism should have moral judgments explaining action, rather than simply saying that everyone who makes a moral judgment has to be motivated accordingly (Sinhababu 2017: 170).

An easy fix is available to the above problem. Internalists might modify their formulation to require the moral judgment to be the source of the motivational force: if an individual judges some action to be morally right and they are *not motivated by this judgment* to act accordingly, then they are behaving irrationally. Since Norman is indifferent to his belief that murdering Sam is morally wrong, he's either not making a genuine moral judgment or is irrational according to this reformulation. The problem now is that given the truth of the Humean theory and cognitivism, the truth of this version of internalism means that all humans who make moral judgments are irrational; "Their moral beliefs aren't the motivational states driving them, because humans can't be motivated that way, so they're irrational. They're driven by other mental states –perhaps the desires contained within the emotions that caused their moral judgments, or desires with de dicto moral content, which isn't sufficient for [this view of] rationality" (ibid, 171).

But let's say that the Humean theory is false and normative beliefs can sometimes generate their own motivational force. The above formulation of internalism would still mean that large numbers of people we would consider virtuous are behaving irrationally, so long as they are motivated by something other than their normative beliefs, or the new desires that their normative beliefs supposedly create. Rationalists like Darwall and May say that all they're trying to do is allow for other ways in which moral motivation can be generated, and

that their theories don't disqualify those motivated by pre-existing desires from moral action, but the least problematic formulation of rationalist internalism says otherwise.

Conclusion: Sentimentalist Optimism

In the last chapter, I did some defending of the Humean theory against anti-Humean rationalism. In the process, I've hopefully given some reason to think that the Humean theory provides the best picture of human motivation. In the previous chapters, I've also argued against various formulations of internalism. If my arguments so far work, then the best metaethical theory would be one that embraces the Humean theory and rejects internalism. Experientialism is just such a theory. Our emotions allow us to experience the world in distinctly moral ways. Internalism and rationalism can't do this because motivation and reasons are common to all sorts of non-moral modes of thought and action. On the other hand, the patterns of moral feeling as suggested by experientialism- the pride we feel upon acting rightly, the happiness we feel at the consequences of right action, the disgust we feel at those who betray what seem to be unquestionable principles of human decency- are distinctive of morality. In this final section I show how a unique commitment to cognitivism allows experientialism to assuage the rationalist worries expressed by May (2018) toward sentimentalism.

.

The thought that reason is a slave to the passions drives rationalists like Joshua May (2018) to view sentimentalism as a source of pessimism about ordinary moral cognition. May thinks that our general regard for reason complicates the emotion/reason dichotomy, and makes us capable of moral knowledge and virtue (228). While sentimentalists like Jonathan Haidt (2012) liken the relationship between the passions and reason to that of one between a powerful elephant and its rider, with the emotional elephant going where it wants and the rational rider just along for the ride, May (2018) argues that "a better analogy in light of the science is to a ruler (reason) and her trusted advisor (mere feelings/passions)" (229). I think the analogy of one mental state as being in some form of indentured servitude to another

has become increasingly unhelpful and obfuscatory. Hume's original expression was about the different functional properties of desires and beliefs as they relate to human motivational psychology. The debate between Humean sentimentalists and anti-Humean rationalists should be centered around which theory provides the best and most defensible psychological picture regarding what these mental states can do, and not about which is subservient to which. Unfortunately, despite all the talk about 'science' and 'empirical data', it often appears that that is what much of the debate amounts to.

Accordingly, I suggest that we look at reason and the passions as distinct offices not caught in a power struggle. The office of reason is concerned with truth, while the office of the passions is concerned with motivation. Moral feelings might represent a certain state of affairs as good or bad, but whether said state of affairs is in fact good or bad is an issue for the office of reason to resolve. Reason might decide that a certain state of affairs is in fact good, but whether this translates into motivation to try to bring about this state of affairs depends on the passions. May thinks that our general regard for reason complicates this dichotomy between reason and the passions. I agree that that might seem to be the case regarding analogies that characterize one state as the master and the other as slave; if both states have to defer to each other regarding matters not of their own domain, then the master/slave dichotomy is indeed made complicated. However, under my suggested picture of office 1/office 2, all a general regard for reason means is a tendency for the two states to cooperate. Both sentimentalists like Haidt and rationalists like May would do well to remind themselves that we're talking about mental states, and take care not to project unhelpful power relations that don't explain anything.

This picture of office 1/office 2 runs counter to sentimentalist non-cognitivisms that say our moral judgments do not express propositions and thus cannot be true or false. In that case, the office of reason is given a diminished role. The picture also runs counter to rationalist cognitivisms like May's which say that the office of the passions "[is] still the slave if anything is" (229). What the picture accords with is a sentimentalist cognitivism as given by experientialism, which respects the domains of both offices.

According to experientialism, moral feelings cause moral beliefs, but don't tell us whether these beliefs are objectively true. For example, Mitch McConnell's hope that the republicans continue to control the senate doesn't make republican control of the senate morally good for him. Instead, the question of whether republican control of the senate is morally good is "the question of whether it's an objective fact that hope represents it accurately" (Sinhababu 2017: 71). Just as with perceptual experiences, moral feelings represent the world as being a certain way. When the world is how you experience it, your experiences are accurate. If republican control of the senate leads to more suffering than there would have been if they had lost control, then that objectively makes republican control something misrepresented by McConnell's hope. In that case, McConnell's hope is inaccurate, and leads him to a false moral belief. The question of whether republican control would in fact lead to more suffering is something that we have to investigate with reason, and not something that our passions can reliably decide. Experientialism's commitment to cognitivism thus makes moral inquiry and knowledge possible, and cooperation between the offices of reason and the passions allows for moral virtue. While experientialism places emotions at the heart of moral judgment, it too has a regard for reason.

Bibliography

Aharoni, Eyal., Sinnott-Armstrong, Walter., Kiehl, Kent (2012) Can Psychopathic Offenders Discern Moral Wrongs? A New Look at the Moral/Conventional Distinction. Journal of Abnormal Psychology. 121:2, 484-497.

Beebe, James & Buckwalter, Wesley (2010). The Epistemic Side-Effect Effect. Mind and Language. 25:4, 474-498.

Bird, Alexander and Tobin, Emma, "Natural Kinds", The Stanford Encyclopedia of Philosophy (Spring 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2018/entries/natural-kinds/>.

Blair, Rachel (1995). A Cognitive Developmental Approach to Morality: Investigating the Psychopath. Cognition 57 (1):1-29.

Blair, Richard., Jones, Lawrence., Clark, Fiona., & Smith, Margaret. (1997). The psychopathic individual: A lack of responsiveness to distress cues? Psychophysiology*, 34*:2, 192-198.

Brink, David (2001) Realism, Naturalism, and Moral Semantics. Social Philosophy and Policy 18:2, 154– 176.

Bonn, Scott (2016). Diagnosing Psychopathy. Psychology Today. URL= <https://www.psychologytoday.com/us/blog/wicked-deeds/201610/diagnosing-psychopathy>

Boyd, Richard (1995). 'How to be a Moral Realist' in Contemporary Materialism: A Reader. Paul K. Moser & J. D. Trout (eds.). Routledge.

Campbell, John (2006). Manipulating colour: pounding an almond. In Tamar Gendler & John Hawthorne (eds.), Perceptual Experience. Oxford University Press.

Chalmers, David (2006). 'Perception and the Fall from Eden' in Perceptual Experience. Tamar Gendler & John Hawthorne (eds.). Oxford Scholarship Online.

Darwall, Stephen (1983). Impartial Reason. Cornell University Press.

Fallon, James (2013). The Psychopath Inside: A Neuroscientist's Personal Journey Into the Dark Side of the Brain. Current.

Fischbacher, Urs., Gachter, Simon & Fehr, Ernst. (2001). Are People Conditionally Cooperative? Evidence from a Public Goods Experiment. Economic Letters. 71, 397-404.

Gale, A. (1975). “Can EEG Studies Make a Contribution to the Experimental Investigation of Psychopathy?” Paper presented at the Advanced Study Institute on Psychopathic Behavior, Les Arc, Bourg St. Maurice, France (September).

Hitchcock, Christopher & Knobe, Joshua (2009). Cause and Norm. Journal of Philosophy. 106:11, 587-612.

Horgan, Terence & Timmons, Mark (1991). New Wave Moral Realism Meets Moral Twin Earth. Journal of Philosophical Research. 16, 447– 465.

Kennett, Jeanette (2006). Do Psychopaths Really Threaten Moral Rationalism? Philosophical Explorations. 9:1. 69-82.

Keser, Claudia & Van Winden, Franz (2002). Conditional Cooperation and Voluntary Contributions to Public Goods. 102:1, 23-39.

Kripke, Saul (1971). ‘Identity and Necessity’ in Identity and Individuation, Milton Karl Munitz (ed.), New York University Press.

Korsgaard, Christine M (2008). “Realism and Constructivism in Twentieth Century Moral Philosophy” In The Constitution of Agency, 302-326. Oxford University Press.

Koenigs, Michael. (2012). The Role of Prefrontal Cortex in Psychopathy. Reviews in the Neurosciences, 23:3, 253-262.

Kumar, Victor (2015). Moral judgment as a Natural Kind. Philosophical Studies. 172:11, 2887-2910.

Kumar, Victor (2016a). Psychopathy and Internalism. Canadian Journal of Philosophy. 46:3, 318-345.

Kumar Victor (2016b). The Empirical Identity of Moral Judgment. Philosophy Quarterly. 66:265, 783-804.

Levy, Neil (2011). Moore on Twin Earth. Erkenntnis. 75, 137-146.

Mackie, John (1977). Ethics: Inventing Right and Wrong. Penguin Books.

Maibom, Heidi (2005). Moral Unreason: The Case of Psychopathy. Mind and Language. 20:2, 237-257.

Maibom, Heidi (2010). What Experimental Evidence Shows Us about the Role of Emotions in Moral Judgement. 5:11. 999-1012

Malatesti, Luca. (2010). Moral Understanding in the Psychopath. Synthesis Philosophica. *24*:2, 337–348.

May, Joshua (2018). Regard for Reason in the Moral Mind. Oxford University Press.

McGinn, Colin (1983). The Subjective View: Secondary Qualities and Indexical Thoughts. Clarendon Press.

Mele, Alfred (2003). Motivation and Agency. Oxford University Press

Merli, David (2002). Return to Moral Twin Earth. Canadian Journal of Philosophy 32:2, 207–240.

Moore, G.E (1903). Principia Ethica. Cambridge University Press.

Nagel, Thomas (1978). The Possibility of Altruism. Princeton University Press.

Nichols, Shaun (2002). On The Genealogy of Norms: A Case for the Role of Emotion in Cultural Evolution. Philosophy of Science. 69:2, 234-255.

Nucci, Larry, & Turiel, Elliot (1978). Social Interactions and the Development of Social Concepts in Preschool Children. Child Development, 49:2, 400-407.

Pettit, Dean & Knobe, Joshua (2009). The Pervasive Impact of Moral Judgment. Mind and Language. 24:5, 586-604.

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

Putnam, Hillary (1975). The Meaning of “Meaning”. Minnesota Studies in the Philosophy of Science. 7, 131-193.

Putnam, Hillary (1973). Meaning and Reference. Journal of Philosophy. 70:19, 699-711.

Railton, Peter (1986). Moral Realism. Philosophical Review. 95:2, 163-207.

Roskies, Adina. (2003). Are Ethical Judgments Intrinsically Motivational? Lessons from "Acquired Sociopathy". Philosophical Psychology, 16:1, 51-66.

Rubin, Michael (2014a). On Two Responses to Moral Twin Earth. Theoria. 80:1, 26-43.

Rubin, Michael (2014b). Biting the Bullet on Moral Twin Earth. Philosophical Papers. 43:2, 285-309.

Rubin, Michael (2015). The Promise and Perils of Hybrid Moral Semantics for Naturalistic Moral Realism. Philosophical Studies. 172:3, 691-710.

Sass, Hanna & Felthous, Alan. (2014). 'The Heterogeneous Construct of Psychopathy' in Being Amoral: Psychopathy and Moral Incapacity. Schramme, Thomas (ed.), MIT Press.

Saver, JL & Damasio, AR (1991). Preserved Access and Processing of Social Knowledge in a Patient with Acquired Sociopathy due to Ventromedial Frontal Damage. Neuropsychologia. 29:12, 1241-1249.

Schramme, Thomas (ed.) (2014). Being Amoral: Psychopathy and Moral Incapacity. MIT Press.

Schroeder, Mark (2008). Being For. Oxford University Press

Shafer-Landau, Russ (2003). Moral Realism: A Defence. Clarendon Press.

Sinhababu, Neil (2017). Humean Nature. Oxford University Press.

Sinhababu, Neil (2019). One-Person Moral Twin Earth Cases. Thought. 8:1, 16-22.

Sinhababu, Neil (forthcoming). Humean Replies to Regard for Reason. Behavioral and Brain Sciences.

Skeem, J. L., Polaschek, D. L. L., Patrick, C. J., & Lilienfeld, S. O. (2011). Psychopathic Personality: Bridging the Gap Between Scientific Evidence and Public Policy. Psychological Science in the Public Interest, 12:3, 95–162.

Smetana, Judy (1985). Children's Impressions of Moral and Conventional Transgressors. Developmental Psychology. 21:4, 715.

Smith, Robert. (1984). The Psychopath as Moral Agent. Philosophy & Phenomenological Research, 45:2, 177-193.

Smith, Michael (1994). The Moral Problem. Blackwell.

Swartzer, Steve (2015). Humean externalism and the argument from depression. Journal

of Ethics and Social Philosophy. 9:2, 1-16.

Turiel, Elliot (1983). The Development of Social Knowledge: Morality and Convention. Cambridge University Press.

Viggiano, Andrea (2008). Ethical Naturalism and Moral Twin Earth. Ethical Theory and Moral Practice 11, 213-224.

Williams, Bernard (1979). Internal and external reasons. In Ross Harrison (ed.), Rational Action. Cambridge University Press.