2-1-2014

Acting Wide Awake: Attention and the Ethics of Emotion

Jacob Davis
Graduate Center, City University of New York

ACTING WIDE AWAKE:
ATTENTION AND THE ETHICS OF EMOTION

by

JACOB H. DAVIS

A dissertation submitted to the Graduate Faculty in Philosophy
in partial fulfillment of the requirements for the degree
of Doctor of Philosophy, The City University of New York

2014

This manuscript has been read and accepted for the Graduate Faculty in Philosophy in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

John Greenwood

_____ _____

Date Chair of Examining Committee

Iakovos Vasiliou

_____ _____

Date Executive Officer

Jesse Prinz _____

Steven Ross _____

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

ACTING WIDE AWAKE:
ATTENTION AND THE ETHICS OF EMOTION

by

Jacob H. Davis

Advisor: Professor Jesse J. Prinz

In cases where two human cultures disagree over fundamental ethical values, metaethical questions about what could make one or the other position correct arise with great force. Philosophers committed to naturalistically plausible accounts of ethics have offered little hope of adjudicating such conflicts, leading some to embrace moral relativism. In my dissertation, I develop an empirically grounded response to moral relativism by turning away from debates over which action types are right and wrong and focusing instead on shared features of human emotional motivation. On my account, being motivated by ill-will is ethically bad (if it is), just because human beings who are fully and accurately aware of how unpleasant it is to be motivated in this way will agree that we ought not to act out of ill-will. Conversely, good-will is ethically good (if it is) just because we ourselves would judge it to be so, if we were fully and accurately aware of how much more ease is present in being motivated in this way. More generally, by appealing to ethical judgments that all members of our human moral community would make if they were alert and unbiased, we can make sense of the idea that individuals and groups sometimes get the normative truth wrong, and that we sometimes get it right. In this way, the experiential ease and unease that is characteristic of various emotional motivations in virtue of our shared human neurobiology can ground a circumscribed set of universal

claims about which motivations we ought to act out of, while leaving many other aspects of how we ought to live open to cultural determination.

Dedication

To three elders in three lineages of mine, Ken, Jerry, and Sayadaw U Pandita; this is a conversation with them about what matters most. And to Sarah, for all she has given to that question of mine - blood, sweat, and tears.

Acknowledgements

In addition to the support of my supervisor Jesse Prinz and the faculty members who served on my Prospectus and Dissertation Committees, John Greenwood, Owen Flanagan, Angelica Nuzzo, Nicholas Pappas, Sharon Street, Steven Ross, and Mariann Weierich, I am also specially indebted to formative conversations along the way with Mark Alfano, Bhikkhu Anālayo, Judson Brewer, Willoughby Britton, Arindam Chakrabarti, Fiery Cushman, Richard Davidson, Angus Davis, Andrew Dreitcer, John Dunne, Joseph Goldstein, Daniel Goleman, Roshi Joan Halifax, Cathy Kerr, Uriah Kriegel, Sayadaw U Lakkhana, Jared Lindahl, Michele McDonald, U Hla Myint, Shaun Nichols, Sayadaw U Panditābhivamsa, Roy Perrett, Lynette Rummel, Bob Scharf, Dan Shargel, Steven Smith, James (J.E.T.) Thomas, Evan Thompson, Tsoknyi Rinpoche, Susanna Siegel, Bill Waldron, David Vago, Nicholas Van Dam, and Jan Westerhoff. In addition, opportunities to present this work during 2011-2013, and ensuing discussions with more smart, good people than I can thank by name, have led to important refinements. I want in particular to thank audiences at the Society for Asian and Comparative Philosophy, the East-West Philosopher's Conference, the Symposium of Cognition, Logic and Communication: Morality and the Cognitive Sciences held in Riga, the Mind and Life Summer Research Institute, the Society for Philosophy and Psychology, the conference on Contemporary Perspectives on Buddhist Ethics at Columbia University, the Eastern Division Meeting of the American Philosophical Association, the Contemplative Studies Initiative at Brown University, the Contemplative Development Mapping Project at the Barre Center for Buddhist Studies, the University of Massachusetts at Boston Philosophy Department, the NEH Summer Institute on Investigating Consciousness: Buddhist and Contemporary Philosoph-

ical Perspectives, the CUNY Cognitive Science Colloquium, the Metro-Area Research Group on Awareness and Meditation, the SIUCC at the University of the Basque Country, Marlboro College, the Moral Psychology Research Group, the Mellon Workshop on Ethics and Aesthetics at Brown University, and the Workshop on Mind and Attention in Indian Philosophy at Harvard University. In addition, I owe a deep debt of gratitude to the abbots, staff, residents, and supporters of the Panditārama Meditation Center, Kyaswa Monastery, the Upaya Zen Center, and the Prajña Mountain Forest Refuge, for providing the conducive conditions for quiet and concentration that resulted in four extraordinarily productive weeks of writing, two in Burma and two in New Mexico.

Contents

Chapter 1

Introduction: An Outline of the Argument

> Violence is immoral because it thrives on hatred rather than love.
>
> - Martin Luther King, Jr.[1]

In the epigraph above, Martin Luther King Jr. proposes answers to two questions: whether violence is immoral, and what makes it so. It is answers to the second sort of question that will be of particular interest to me here. King seems to be taking qualities such as love and hatred, what I will call Qualities of Heart, to determine whether a type of action (in this case violence) is moral or immoral. In this dissertation, I explore such a thesis as an approach to understanding morality and ethics very generally. My central proposal is that the relative ease or unease characteristic of various types of emotional motivations can ground a circumscribed set of ethical claims that apply to all human beings, on their own terms. On my approach, hatred or ill-will or some Quality of Heart in the vicinity of these is a bad thing (if it is) just because individuals who are fully and accurately aware of episodes of ill-will feel for themselves how unpleasant it is. Conversely, love or goodwill or some Quality of Heart in the vicinity of these is a good thing (if it is) just because individuals who are fully and accurately aware of episodes of goodwill feel for themselves how much more ease there is in that kind of an emotional state. More generally, our ability to converge on a thorough and unbiased awareness of the relative ease and dis-ease of various types of emotion

[1] Nobel Lecture, University of Oslo, December 11, 1964

can ground a circumscribed set of universal claims about which sort of emotional motivations human beings should act out of and which we should not.

Using this approach, my account attempts to answer two different sorts of questions that are taken up in moral philosophy. Suppose you think that unwavering but non-violent resistance is the only way anyone should respond to aggression from others, and suppose I think that violence is sometimes justified. Or take a more extreme example. Suppose that you think that the practice of "honor killings" of young women for being raped is monstrous, an unthinkable thing to advocate; and suppose that someone else feels strongly that such killings are required to save the honor of the family in question, that not to continue this traditional practice is what is unthinkable. Attempting to resolve such ethical questions, in the context of disputes over fundamental values, very quickly brings us to deeper philosophical questions. What would make it the case that one of these judgments was correct? My way of taking emotional motivations as the focus of ethical evaluation, if it is cogent, settles some ethical questions in a way that applies to all human beings, from their own point of view. Secondly, the approach provides a principled way of delineating which sorts of ethical questions can and which cannot be settled in this universal way. In particular, my proposal is that the only ethical questions to which we can give answers that apply to all human beings on their own terms are questions about which Qualities of Heart one ought to be motivated by, along with implications for action that derive directly from this.

To take a limiting case, it is rude in Burma to stick a fork in one's mouth (one uses it to push the rice and curry onto one's spoon, of course!), but not so in my culture of origin. In this case, my approach holds that there is nothing to settle the question of how one ought to use one's silverware in a way that applies to all human beings, from their own point of view. One might well think that if one knows that sticking one's fork in one's mouth is rude in Burma, one would have to be disrespectful to do so. And perhaps disrespectfulness is a Quality of Heart that no human being ought to act out of. But considering cases where this action could arise without being motivated by disrespect (no one told me about this aspect of Burmese etiquette, when I first went there!), it seems likely that there is nothing intrinsic to the action-type of sticking one's fork in one's mouth

that requires it be motivated by a bad intention. In contrast, there may be some types of action that are tied more directly to particular Qualities of Heart. I am sympathetic for instance to a suggestion we could draw from MLK's comment above, that violence even for a justified cause is immoral for any human being, and is so just because in any human psychology the force of hatred must outcompete that of love for violent action to be taken. I take this one step further by offering an account of what would make it true (if it is) that human beings should be motivated by love rather by hatred.

My normative account of what makes certain Qualities of Heart ethically better than others is premised on a psychological claim, a hypothesis that under conditions of ideal emotional awareness human beings will converge on a certain circumscribed set of judgments about how one ought to be. In Chapter 2 I make the case for this empirical claim of convergence. I draw on recent empirical research on mindfulness meditation to suggest an objectively measurable sense in which we can be fully and accurately aware of external and internal stimuli, what I call being Wide Awake. Secondly, I draw on recent research in moral psychology to suggest that we express in our ethical evaluations of others' intentions, and more specifically the emotional motivation that we perceive as giving rise to those intentions, whatever attitudes we hold toward that emotional motivation in ourselves. This does not suggest that we would, if pressed, cite our own preferences as *justification* for the ethical claims we make. Rather the claim is that whatever attitudes we hold toward a Qualities of Heart in ourselves, these are what we express in ethical judgments about the intentions we perceive behind others' actions.

Together these claims suggest that to the degree human beings are Wide Awake, regardless of their cultural socialization, they will come converge in judging certain Qualities of Heart to be ethically better than others. This is so, I suggest, not only in virtue of the project of avoiding unease that may be shared among all animals, but also in virtue of the particular sorts of ease and unease characteristic of various Qualities of Heart for human beings, given our shared human neurobiology. For instance, I suspect that any human being who feels fully the unpleasantness of being motivated by hatred and the relative ease of being motivated by friendliness will prefer

to be motivated to the later. My further prediction is that shifts toward an attitude of preferring friendliness to hatred in oneself will be expressed in shifts toward ethical judgments to the effect that friendliness is a better emotional motivation than hatred for any human being to have. If this is correct, then regardless of cultural background, increases in emotional awareness should lead to systematic convergence on a certain circumscribed set of ethical judgments, in particular judgments about which intentions are good ones, along with any implications for action that there may be from this. Because this hypothesis of convergence has yet to be tested, my argument for it is more suggestive than conclusive. But I take it as a strength of the account that this foundational premise is open to direct empirical confirmation or disconfirmation.

In Chapter 3 I argue from the empirical claim for convergence to a normative claim. In particular, I suggest that the ethical judgments converged on by those who are Wide Awake are ones that any human being has reason to defer to. To be Wide Awake, as I define it, implies feeling more fully which ways of being bring ease and unease. The ethical claims converged on by anyone to the degree they are Wide Awake have force for all human beings because, given our shared project of avoiding unease and the shared human neurobiology of emotional motivations, any internally consistent plan for a human life will be one that requires us to act as if we were Wide Awake. We can indeed choose to cultivate Qualities of Heart that are characterized by unease, and we often do. In order to do so, one can choose to distract oneself so as not to feel the unpleasantness of being motivated by greed, or by hatred, or by apathy to others suffering. What I want to suggest, though, is that we can only partially succeed at making ourselves insensitive in this way. If so, to the degree a human being to choose to cultivate Qualities of Heart that are characterized by negative affective valence, she is subject to experiencing that valence as unease. Being motivated as we all are to avoid such states, in choosing to cultivate and act out of Qualities of Heart she experiences as uneasy, she leaves herself with a certain level of cognitive dissonance. In contrast, there is a possibility for a human being to be wholehearted in choosing to cultivate and act out of Qualities of Heart that are characterized by ease. One sort of choice puts us in for the unease of cognitive dissonance while the other allows us the ease of wholeheartedness. Ultimately, it is this

hedonic asymmetry, between choosing to cultivate Qualities of Heart characterized by ease and those characterized by unease, that underwrites my claim that all of us have reason to be one way and not the other. I suggest that we all share the motivation not to be vulnerable to unease, and that we can succeed at this most effectively by making judgments about how to live from a place of more full and accurate awareness of what it is like to live in various ways, that is by being Wide Awake. If so, then all else being equal it is the judgments about how to live that we would make to the degree that we were Wide Awake that we ought, by our own lights, to defer to. In short, we can *act* Wide Awake, even when we are not. If for instance there is an answer about whether a human being ever ought to engage in violence, according to AWA the answer will turn on what kinds of emotional motivations are required to perform the actions in question to the degree, and whether we ourselves would want to be motivated in those ways if we were feeling fully what it is like to be motivated in that way.

Chapter 4 surveys four prominent approaches to ethical theorizing in Western philosophy, noting points of convergence with and divergence from AWA. I noted above that AWA appeals to shared features of our human neurobiology, in particular features that would make it the case that certain Qualities of Heart are characterized by greater ease than others, for any human being. On the one hand, the appeal to ease and un-ease has obvious parallels the Utilitarian approaches to ethical theorizing that are known historically from theorists such as Bentham and Mill; I focus here on the recent empirically-grounded work of Joshua Greene. On the other hand, AWA's appeal to human nature to ground ethical judgment has a clear affinity with Phillipa Foot's neo-Aristotelian approach of grounding ethical claims in objective, ultimately scientific, facts about what is necessary for the human form of life. Both neo-Aristotelian and Utilitarian approaches take ethical claims to be grounded from outside of the particular point of view of any human agent. This is their strength but also their weakness; the descriptive premises about objective natural properties that these approaches begin from lead to ethical conclusions only with the addition of a suppressed normative premise. In this sense, the Kantian and Humean moves to a more subjective ground are each (if very different) steps in the right general direction. Like Kant, AWA takes a human

being to be "subject only to laws given by himself yet universal" (GMM 4:432). Kant takes this universality of moral truths to be grounded in laws of reason that would be binding for all rational beings. But this means of grounding substantive moral truths in a way that is immune to empirical disconfirmation is implausible, and ironically, is also itself subject to empirical disconfirmation. In taking ethical truths to be grounded instead in emotional dispositions, AWA adopts instead a kind of Humean subjectivism. Many empirically-oriented theorists in the Humean tradition, noting the robust evidence for the dependence of moral judgment on socialized emotional responses, have concluded that ethical questions about how we ought to live can only be answered from the standpoint of a particular human culture or individual. AWA offers one avenue for opposing such a culturally relativist conclusion, from a naturalistically plausible standpoint.

The key move AWA makes in resisting cultural relativism is to distinguish the range of normative frameworks that are logically possible from the subset of those normative frameworks that are humanly possible. For it is a very particular sort of beings with whom we can actually make and debate ethical claims. The specific proposal made by AWA is that the relative ease and unease characteristic of various Qualities of Heart for human beings, given our shared human neurobiology, puts a constraint on which normative frameworks we can actually inhabit. Aside from whatever promise or failings this particular proposal may have, AWA can serve as an example of a more general strategy for opposing cultural relativism from a naturalistic perspective. The premise of this general approach is just that there is some shared psychological structure that puts enough of a constraint on human normative frameworks to provide a ground for adjudicating some of the ethical disputes that arise between cultures, individuals, and even parts of ourselves. Call this the claim for Shared Human Contingency. It is peculiar that in recent debates in moral philosophy the avenue of resisting cultural relativism from the empirical possibility of Shared Human Contingency seems not to have been emphasized. In part this may be due to overlooking the possibility that the Qualities of Heart motivating our actions can serve as a point of evaluative focus on which there could be such commonality. But more generally if there were to be certain shared psychological structures that constrain what starting commitments a human being can hold, then

modern metaethical accounts inspired by Hume can be reconciled with Hume's own conclusion that we do share "certain instincts originally implanted in our natures, such as benevolence and resentment, the love of life, and kindness to children" (Hume, 2000, II.3.ii). Chapter 5 shows how the strategy for resisting cultural relativism based on this premise is at least initially compatible with a variety of recent naturalistic metaethical theories. I focus on three examples in particular: Simon Blackburn's Quasi-Realist approach, Jesse Prinz's culturally relativist sentimentalism, and Sharon Street's Humean constructivism. All three share an emphasis on the radical contingency of the particular framework of values one happens to inhabit, that starting set of commitments from which one engages in practical reasoning. Precisely because modern Humean approaches take one's starting normative commitments to be radically contingent, on any of them it could just so happen for evolutionary and psychological reasons that all the beings with whom we can in fact make and debate ethical claims share enough of their normative frameworks in common with one another to adjudicate certain of these differences.

Chapter 2

The Empirical Claim for Ethical Convergence

The mental states behind one's actions are a central focus in Buddhist contemplative traditions. Echoing remarks from the early Buddhist dialogues (e.g. AN.6.63) the Dalai Lama (2001, 31) suggests that "the individual's overall state of heart and mind, or motivation, in the moment of action is, generally speaking, the key to determining its ethical content". A similar sentiment can be found in spiritual leaders from other traditions; I noted in the introduction Martin Luther King Jr.'s suggestion that "violence is immoral because it thrives on hatred rather than love." The shared suggestion here seems to be that qualities such as love and hatred, what we might more generally refer to as "Qualities of Heart", can determine whether an action is ethically good or not. In some Buddhist contemplative traditions, this focus gives rise to the proposal that developing mindful awareness of one's own bodily, perceptual, and emotional states endows one with the wisdom to know right from wrong.

These claims by religious leaders are thoroughly normative ones. And they seem to include an implicit claim to universality, that for instance an action motivated by hatred is an immoral one for all human beings. We could dismiss these suggestions as simply the result of the particular socializations of Martin Luther King and the Dalai Lama, and the claim for universality as so

much hubris. On the other hand, perhaps these two leaders have got something importantly right. My suggestion is that we can draw out from their claims testable empirical hypotheses. Take for instance the proposal that developing mindful awareness of one's own bodily, perceptual, and emotional states endows one with the wisdom to know right from wrong. Implicit in this normative claim is an empirical claim that mindful awareness of one's own bodily, perceptual, and emotional states will lead to convergence in ethical judgments. That is, whatever cultural socialization various human beings happen to have, to the degree they are able to develop mindfulness, they will come to agree on certain judgments about what is right and what is wrong. We can formulate this as the empirical claim for Convergence in Ethical Judgment (CEJ).

> Convergence in Ethical Judgment (CEJ): To the degree individuals are fully conscious and accurately aware of the interoceptive stimuli characteristic of various emotional motivations, they will converge on judgments about which emotional motivations other agents ought act on.

The currently available evidence is insufficient to confirm or disconfirm this hypothesis. Even if the empirical claim for Convergence in Ethical Judgment were to be robustly confirmed, moreover, this would not be sufficient to support the normative claim that we ought to defer to whatever judgments are so converged on. That normative project is the topic of subsequent chapters. My goal in this chapter is simply make the empirical claim for CEJ plausible, in light of recent research on ethical judgment and on mindfulness meditation.

I argue for the empirical plausibility of CEJ by appealing to two related empirical claims. First is an account the psychology of moral judgment that I call the Second-order Attitude Theory. The basic idea is that if I have an attitude against being motivated by hatred myself, when I perceive another's actions as motivated by hatred, I will express my attitude in judging that sort of intention to be a bad one. In Section 2.1 I review evidence in favor of SAT.

> Second-order Attitude Theory (SAT): Judgments that other agents ought or ought not to act out of a certain emotional motivation express the affective attitude that the person making the judgment holds toward that type emotional motivation in themselves.

On its own SAT implies nothing about convergence. Righteous anger towards injustice is valued in liberal Western culture, for instance, but discouraged by many Buddhist traditions. Various cultures socialize various and opposing attitudes towards particular Qualities of Heart. Nonetheless, a type of convergence is possible. In Sections 2.2 and 2.3 I review evidence in favor of this claim, Convergence in Second-order Attitudes.

> Convergence in Second-order Attitudes (CSA): To the degree individuals are fully conscious and accurately aware of the interoceptive stimuli characteristic of various emotional motivations, they will converge in which sorts of emotional motivations they will want for their own actions.

The idea here is that in virtue of the way human emotions are instantiated neurophysiologically, to the degree any human being really feels what it is like to be motivated by hatred and by care, she will prefer to be motivated by care rather than hatred. To bring out the plausibility of this claim it can help to consider the converse claim. I think it is wildly implausible that to the degree any human being really feels what it feels like to be motivated by hatred and by care, she will prefer to be motivated by hatred rather than by care.

The central challenge for an argument such as mine is the manifest diversity of moral values expressed in behavior and in speech: different human cultures, sub-cultures, and individuals give very different answers to the ethical question of how one ought to live. Even in regard to which sorts of emotional states we value, there is variation. My central answer to this challenge is detailed in Section 2.3. Human beings often do not feel the emotional motivations behind their actions, or do not feel them in a full and unbiased way. I argue that the resulting lack of awareness and biases of awareness explain some of the evident diversity in ethical judgments. Because of this, to the degree we can come to feel our own emotional motivations in a more full and unbiased way, this will in itself lead to convergence in the specific area of ethical judgment concerned with which sorts of emotional motivations we all ought to have. As I detail Chapter 3, however, this area of circumscribed on convergence ethical evaluation may have more widespread implications for the evaluation of action.

2.1 Moral Judgment of Intentions

Does how we feel about how we feel about things matter ethically? Drawing from recent empirical research in moral psychology, I argue in this section that (second-order) affective attitudes towards specific types of (first-order) affective attitudes play a primary role in shaping ethical judgment at the levels of psychological mechanism and social function. If I am right, in judging an agent's motivation as "right" or "wrong", "good" or "bad", "saintly" or "evil", we express our own attitudes towards specific types of attitudes, and thereby apply social pressure to higher-order affective attitudes that differ from our own.

This second-order approach is not intended as a comprehensive account of moral judgment. Recent evidence suggests that first-order emotions may drive many of our decisions about what is right and wrong. Anger seems to drive fairness judgments as well as retributive choices in economic games (Fehr and Gächter, 2002). Similarly, emotional reactions of disgust are at least partly responsible for the severity of certain moral judgments (Wheatley and Haidt, 2005). Recent empirically-oriented moral philosophers such as Prinz (2007) have used such evidence to argue for moral relativism, citing both the ubiquitous role of emotional dispositions in moral judgment, and also evidence that the actions towards which individuals are disposed to feel such first-order emotions varies widely between cultures. I respond by taking a divide-and-conquer approach. I grant for the sake of argument both that many of our moral judgments are driven by first-order attitudes, and also that these attitudes vary widely across cultures. Disgust is highly culturally malleable and other than its primary gustatory function, may have no social object in common across human cultures. Anger about harmful or unfair outcomes of action may be more universal, but the factors determining what counts as unfair and which harms are allowable are still largely culturally determined. However, these sorts of outcome-responsive judgements can be distinguished systematically from a more developmentally advanced psychological system that is also operative in adult human ethical judgment.

The crucial distinction can be drawn out with the classic example of moral luck. Imagine that two negligent drivers happen to swerve off the road, the one hitting and killing a small tree, the

other hitting and killing a small child. Judgments about blame and deserved punishment seem to respond primarily to the outcome: killing a child is much more blameworthy than killing a tree. What is interesting is that, according to recent evidence, judgments of how wrong the driver was or how bad a character he has seem to track a different variable: the mental-states that were responsible for his action, independent of the outcome. And while judgments about which sorts of emotional motivations are good may also diverge widely between cultures and sub-cultures, I propose that there is more hope for convergence here.

Responsibility and Intention

Recent evidence suggests that two distinct and interacting systems are operative in producing judgments about harm violations, the one focused on mental states, the other responding in a significant way to actual outcomes. In one experiment by Cushman (2008), respondents were given vignettes of an agent, Jenny, engaged in welding with a partner. In some scenarios, Jenny burns the partner's hand with heat traveling down hot metal. In others, she narrowly misses because the partner happens to let go at the last minute. The vignettes further varied as to whether or not Jenny wanted to burn her partner's hand, and as to whether or not she believed that her action would cause her partner's hand to be burnt. For judgments about whether the action was wrong or forbidden, the agent's mental states, such as her belief that the action would burn the partner's hand and her desire to do so, explained the 83% of the variance. In contrast, for judgments about whether blame and punishment were deserved, outcome was a major factor; whether the partner's hand was actually burned explained 22% of the variance in such punishment judgments, as opposed to 3% of the variance in wrongness judgments.

In the context of developmental psychology, Cushman et al. (2013) present evidence suggesting that the mechanism driving outcome-responsive punishment judgments emerges earlier than, and is later constrained by, mental-state based ethical judgments. Although even four-year-olds were sensitive to intent as well as outcome, older children judged accidental harms less punishable and much less indicative of a naughty character than did younger children. Mediation analysis

suggested that the development of intent-based naughtiness judgments explained this variance. Cushman et al. take this as evidence for a "constraint hypothesis", on which "intent-based moral judgment emerges in the form of a new concept of moral wrongness that subsequently constrains judgments of deserved punishment" (Cushman et al. 2013: 12).

This developmental evidence for competitive interaction between distinct outcome-responsive and mental-state-responsive systems of moral judgment accords with demonstrations that disrupting neural activity in late-maturing areas associated with mental-state attribution (i.e. right temporalparietal junction) decreased the weight of perceived intentions in adult moral judgments (Young et al., 2010). The proposal also lines up neatly with a phylogenetic suggestion, that the outcome-responsive system is a more ancient evolutionary adaptation. In animal (and human) contexts in which theory-of-mind abilities might be limited or non-existent, punishing negative behavioral outcomes may be the most efficient way to enforce cooperative behavior. In contrast, the mental-state based system of judgment may be limited to much later evolutionary developments, or may be transmitted culturally rather than genetically. Thus it might be suggested that both over historical and also developmental trajectories the outcome-responsive system of retribution comes to be constrained in adult humans to various degrees by a mental-state responsive system of moral judgment.

Recent work by Inbar et al. (2012) suggests a more precise account of what this constraint might amount to. Inbar et al. found that respondents judged actions such as a financial investor putting himself in a position to benefit from harm to be both "morally wrong" and "blameworthy", even when there was in fact no beneficial outcome for the agent, or when there was no outcome information available. The authors explain the results as reflecting judgments of ethical evaluation of underlying desires, in particular "perceptions of desires for a harmful outcome."

Although Inbar et al. offer compelling evidence that the mental-state based system of moral judgment does operate in some cases on an agent's desires alone, other studies question the relative importance of this factor. In Cushman's (2008) study, explicit information about desires was manipulated, and yet the belief variable explained a larger part of the variation across conditions.

But the situation is complicated. What are subjects judging when presented with a case in which it is stipulated that an agent, Jenny, believes that welding the metal will burn her partner, does not want to cause that harm, but nonetheless performs the action of welding the metal? To help elucidate this difficult case, we might consider other cases that share this structure, in which an agent doesn't want to do something, but nonetheless goes ahead and does it. Consider the classical example given by Frankfurt (1971), of a reluctant addict, one who desires narcotics but also wishes he didn't. Frankfurt describes this as a case in which the addict does not identify with his first-order desire. Notice that in describing the addict's actions, we can use the language of desire in two ways. On the one hand we can say that the addict clearly does want to continue to use drugs, as evidenced by his actions. On the other hand, we might say that the addict continues to use, despite the fact that he really doesn't want to. In the first case, the language of desire is used to describe a first-order affective attitude, the sort that we might take as necessary for an agent to act. In the second case, the language of desire is used instead in a higher-order sense; although the addict does have the first-order desire for drugs, the fact that he has that desire is distressing for him, because that is not the sort of desire that he wants to have. Put more generally, the addict has a higher-order affective attitude against the sort of lower-order affective attitude that actually motivates his action in this particular case.

Using this higher-order analysis can help explain Cushman's results. Presented with a vignette specifying that Jenny believes that her action will cause harm to another, does not want to cause that harm, yet does the action anyway, respondents are faced with a choice. They might take this as evidence that Jenny's is a case of abnormal psychology, in which beliefs and desires don't hook up in the right way. But the vignettes offer no reason to think that this extrapolation about Jenny is warranted. Alternatively, they could look for evidence that Jenny wants some greater good that requires burning her partner's hand: perhaps many other people will be saved a violent death by this small act of harm. But the vignettes offer no reason to think that such an extrapolation is warranted, either. A third option is to read the vignette in the way we do the description of Frankfurt's reluctant addict, who doesn't want to use, but does anyway. Given the difficulties of

the alternative explanations, this one seems a plausible option for respondents to take. The idea here is that presented with a vignette specifying that Jenny believes that her action will cause harm to another, does not want to cause that harm, yet does the action anyway, respondents infer that like the reluctant addict, Jenny has a first-order desire to harm her partner, but also a second-order desire not to have such an evil desire.

In the next section, I draw on recent studies suggesting that such inferences about higher-order affective attitudes do affect moral judgment in important ways. I have suggested how respondents might make inferences about a higher-order attitude from the interaction between the specifications of belief, (first-order) desire, and action in Cushman's vignette about Jenny. If this account is cogent, and if I am right that such inferences to higher-order affective attitudes are even more important to mental-state responsive moral judgments than information about first-order desires, then we would predict that in Cushman's (2008) study the belief variable would explain more of the variance than the (first-order) desire variable. This just what Cushman found. Even in judgments of punishment and blame, which respond in a significant way to outcomes as well as to information about mental states, the agent's belief still accounted for 50% of the variance, while her desires only accounted for 13%. In judgments of wrongness, which Cushman found to be more purely responsive to information about the agent's mental states, the desire variable accounted still only accounted for 21%, while the belief variable accounted for 62% of variability. While we cannot take this as a direct index of the importance of higher-order attitudes to ethical assessment, this evidence is suggestive.

Some of the most powerful evidence in favor of analyzing the function of ethical evaluation in terms of higher-order affective attitudes comes from a study by Pizarro et al. (2003). One set of vignettes described an agent, Jack, either doing the positive deed of giving his coat to a homeless person, or else the negative act of smashing the window of a car parked in front of him, and as doing these acts either impulsively or else deliberately. Pizarro and colleagues found an interesting asymmetry: subjects discounted negative evaluations of negative acts that were performed impulsively as opposed to deliberately, but did not do the same for positive impulsive acts. Further

analysis suggested that perceptions of second-order desires were responsible for this asymmetry. In one experiment, without being giving explicit information about the agent's second-order desires, subjects were asked to indicate on a Likert scale the degree to which they thought the agent "would rather *not* have had an impulse...", "wanted to have an impulse...", and "really wanted to do what he did". Soliciting such perceptions revealed that subjects assumed that "the second-order desires of agents who performed positive acts were consistent with their impulses, whereas they believed that agents who performed negative impulsive acts had conflicting second-order desires" (Pizarro et al., 2003, 271). This assumption meditated the asymmetry. Moreover, when subjects were explicitly told that the agent performed a negative or positive act impulsively, yet wished not to or had the second-order desire not to possess such an impulse, they discounted praise for positive acts, and further discounted blame for negative acts. This strongly suggests that the perceived higher-order desires or preferences of an agent are important to ethical evaluation.

More recent results are also consistent with this higher-order account. Critcher et al. (2012) found that descriptions of a moral or immoral decision as being made quickly or instead as made "after long and careful deliberation" signaled to respondents information about certainty. Follow-up experiments suggested that people perceive such certainty as carrying information about whether the agent feels conflicted or not about performing a moral or immoral act, and adjust character assessments accordingly. Woolfolk et al. (2006) found that respondents judged a killing more harshly when the agent was taken as identifying with his action than when he was taken as not identifying with the action. The respondents read vignettes that described an agent, Bill, who was forced by hijackers to kill a friend, Frank, with the twist that Bill has just discovered evidence that Frank is having an affair with his wife. In the "high identification" condition, Bill is described as being identified with the action, "He wanted to kill Frank. Feeling no reluctance, he placed the pistol at Frank's temple and proceeded to blow his friend's brains out." In the "low identification" condition, although Bill was "beside himself with distress, he reluctantly placed the pistol at Frank's temple and proceeded to blow his friend's brains out." Respondents judged Bill more responsible and more blameworthy in the "high identification condition".

It might be possible to explain the Woolfolk and Critcher results in terms of ethical judgments about the relative strength of various competing first-order desires. The relevant desire opposing the killing might be a first-order one not to kill Frank, rather than a general higher-order affective attitude against having the desire to kill in such a circumstance. And maybe having motivation sufficient to result in actually killing Frank is not as bad if your desire not to kill Frank is almost, but not quite, strong enough to stop you. If the moral judgments concerned only whether there was sufficient motivation or not to carry out the act, however, this leaves unexplained why they would vary dependent on the degree of identification.

If an explanation in terms of first-order desires won't suffice, alternatively, a character-based approach to moral judgment might do better. In articulating the implications of their finding on evaluation of "wicked desires", Inbar et al. (2012) suggest that in evaluating acts that put an actor in a position to benefit from harm, respondents might infer that "only a bad person would do such acts, because of the desires they require an agent to adopt", and that similar character-based inferences might at least in part explain moral evaluation of violations of purity norms such as cleaning one's toilet with the national flag (Haidt et al., 1993). Elsewhere, Pizarro (2010) suggests that in the cases of such agents, "the inference that he or she intended a negative outcome seems reasonable (because bad people, by definition, are likely to desire and intend bad things)." However, Inbar et al. found that although ethical judgments of putting oneself in a position to benefit from harm impacted subsequent character judgments, the reverse was not true. Negative character assessment did not predict negative judgment of future actions. Perceived desires for harm to occur, on the other hand, did predict action evaluations. Thus although important, character judgments may not be the central focus of such ethical evaluation.

A more parsimonious and precise way of understanding character evaluations in these cases begins from the search for what it is about being a bad person that makes one likely to desire and intend bad things, as Pizarro puts it. Notice that in the Woolfolk study the description of the low identification condition of the reluctant killer Bill is much like that of Frankfurt's reluctant addict, who does not identify with his first-order desire for narcotics. On the one hand, it is clear that Bill

has sufficient desire to motivate him to pull the trigger and that the addict has sufficient desire to procure his fix, as evidenced by their actions. On the other hand, the language of desire can be used in a second-order sense. Just as we can say of the addict that, despite the fact that he really doesn't want to, he does continue to use, likewise we can say of Bill that despite the fact that doesn't really want to, he does pull the trigger and kill his friend. This suggests that the reason Bill is judged less harshly, when we are told that despite pulling the trigger he doesn't really want to, is that our moral judgment is responsive to Bill's higher-order affective attitudes.

Nichols (2004) raises an important empirical objection to accounts that explain moral judgment in general in terms of higher-order attitudes, such as Simon Blackburn's (1998). The argument begins from noting the distinction between conventional and moral judgment. In one classic study, Tisak and Turiel (1984) tested childrens' responses to vignettes involving transgressions of moral rules, such as stealing lunch money or pushing another child and causing her to scrape her knee. They compared these with responses to vignettes in which a child transgresses a prudential rule against running, and ends up scraping her knee. They found that that children were more likely to say that moral norms applied "in another city" than prudential ones. Moreover, in studies by Nichols and Folds-Bennett (2003), children treated moral norms as more generalizable and less authority-dependent than paradigmatic response-dependent properties. For instance, children who responded that onions are icky but might not be in other countries still typically judged that pulling hair would be wrong in other countries. Children as young as three demonstrate understanding that in comparison with conventions such as not chewing gum in class, norms against causing harm (against pulling hair, for instance) are not only less permissible and more serious, but also more generalizable and less authority-dependent (Smetana, 1981). And even young babies have negative reactions to agents who harm others (Hamlin et al., 2007). However, as Nichols points out, the higher-order aspects of moral judgment seem to emerge much later. In one experiment in a study by Nunner-Winkler and Sodian (1988), when presented with a case in which one child pushed another off a swing, children under six predicted that the pusher would feel happy, while older children predicted negative affect. In a subsequent experiment, one of the children committing a

moral violation had a happy face and another felt bad. Above six years of age, children tended to rate the child who displayed joy after a transgression as worse than the one who felt bad for their action; younger children were at chance.

Nichols takes the results by Nunner-Winkler and Sodian as evidence that the second-order aspect of evaluative judgments develops later than the capacity to make moral judgment, and uses this to challenge accounts that explain moral judgment in general in terms of higher-order attitudes. Nichols is correct that higher-order accounts fail as an explanation of the judgments of the outcome-responsive system of moral judgment. This is the system of moral judgment that we would expect to be employed by young children, in light of the developmental evidence noted above from Cushman et al. (2012). But this doesn't show that higher-order judgments dissociate from intent-based moral judgment. On the contrary, according to that Cushman et al., the intent-based system of moral judgment comes to the fore by about the age of seven, which is precisely when Nunner-Winkler and Sodian find the emergence of higher-order judgments to the effect one who takes joy in harming is bad. This is just what we would expect if higher-order affective attitudes play a crucial role in the mental-state responsive system of moral judgment.

Psychological Mechanism and Social Function

Why would higher-order affective attitudes be so important to ethical evaluation? One plausible answer is that an individual's higher-order attitudes affect how they behave. In the psychology of an agent, first-order affective attitudes might be selectively sustained and defeated by higher-order affective attitudes. If I have not only a preference for altruistic acts but also a preference for preferences for altruistic acts, a weak or intermittent intention to help others may be revived when absent and strengthened when weak, so as to be much more likely to result in altruistic action. Conversely, if I have a preference for seeing a certain person dead, but also have a preference not to have a preference of that type, in cases where my preference would otherwise dispose me to action, the higher-order preference may weaken or defeat that motivation such that it does not have the requisite strength or continuity to result intentional action.

My suggestion is, first, that mental-state responsive ethical judgments express the higher-order attitudes held by the person making the judgment, and that in virtue of this, they also affect the behavior of agents indirectly, by putting social pressure on higher-order attitudes that differ from those expressed. The proposal is not that this is the only way in which ethical judgment can operate, nor that all ethical judgment operates in this way. On the account developed above, the outcome-responsive system driving blame and punishment judgments is driven primarily by first-order reactive attitudes. Higher-order attitudes play their role only in the more developmentally advanced system of mental-state responsive ethical judgment. Moreover, even when we do make the sort of ethical judgments that focus primarily on mental states such as the agent's intention, the claim is not at all that we think of ourselves as expressing and affecting higher-order affective attitudes. Rather, the proposal to be tested is that such judgments operate in this way on the levels of psychological mechanism and social function, quite apart from how we experience or conceive of these processes.

If this is right, one thing adult human beings do in judging acts as "right" or "wrong" is to express their own affective attitudes towards specific types of (first-order) affective attitudes. We say that the desire to kill another is an evil desire just in case we find such an emotional motivation undesirable, or more precisely, recalling the distinction from Section 2.1.1, just in case we have affective attitudes against affective attitudes in favor of deadly harm befalling another. If we perceive an agent as sharing this higher-order affective attitude of ours, then it is sufficient to remind him of this value that he and we share, by reminding him that such motivations are evil ones. Reinforcing the desire of his not to desire harm can then do the psychological work of weakening or defeating any intention to kill that he might have. In cases where we can't appeal to another's higher-order affective attitudes, their own values, to influence their actions so as to conform with our values, there may be no other recourse than excluding such a person from our social interactions. Indeed, recent work by Skitka suggests an intimate relation between higher-order emotional attitudes and preferred social distance (Skitka et al., 2005; Skitka, 2010). Expressing moralized attitudes in the form of moral judgments may implicitly act as a challenge to listeners either to adopt a similar

attitude or else suffer the resulting social exclusion and lack of social cooperation.

On the level of psychological mechanism, then, the proposal is that mental-state responsive ethical judgments are expressions of higher-order affective attitudes. On the level of social function, the proposal is that in making ethical judgments of others intentions we put pressure on higher-order attitudes that differ from those the judgment expresses. It is important to this account that the higher-order affective attitudes expressed in ethical judgment are preferences about lower-order preferences themselves, considered quite generally. While I might prefer that my romantic partner prefer not to leave me, because this is a preference about a particular person rather than about the preference itself, considered generally, the theory does not predict that we would judge such an intention to be morally wrong, even though we might not like it. In contrast, a judgment to the effect that wanting to leave one's partner is morally bad in general expresses an affective attitude about this type of affective attitude, such that it applies to the relevant emotional motivation equally in any human being. On this second-order approach to dual-system moral psychology, what matters for ethical judgment is how the person making the judgment feels about the various ways in which people (ourselves and others) feel about things. There are cases where we like things that we would prefer not to like. Even in such cases, an action we judge as evil is one we prefer that people (ourselves and others) not prefer. An action we judge as virtuous is one we prefer that people (ourselves and others) prefer.

This account would predict that a real-life version of the reluctant ("low-identification") Bill, when asked, would judge the emotional motivation to kill as evil, because he has a higher-order attitude against such a motivational state. Conversely, telling the enthusiastically murderous ("high-identification") Bill that what he is doing is wrong forces him either to convince us to adopt the higher-order affective attitudes he holds, to renounce the attitudes he currently holds and instead express ethical judgments consistent with our attitudes, or else to suffer social exclusion. Likewise, expressing to a third-party the judgment that what Bill is doing is wrong forces them either to agree, or to convince us to adopt a different set of higher-order affective attitudes, or else to risk social exclusion.

2.2 The Qualities of Heart

In the psychological literature reviewed above, while much attention has been devoted to the intentions we perceive to be behind an agent's actions, the state focused on here is taken to be characteristically cognitive. Malle and Knobe (2001), for instance, suggest that desires and intentions differ, according to the folk-psychological understanding we apply in moral judgment, in that while desires are taken to be the input to a decision-making process, intentions are taken to be the output. For this reason, intentions must represent the action to be done. In a case of deliberate action, social perceivers take the agent to have appropriately determined that the intention to be enacted fits both the way the world is and also the agent's desires, all-things-considered. I noted in the Introduction how MLK speaks of hatred and love as determining the moral valence of certain types of action. Such emotional states may serve as action potentials. If we take anger or love to be associated with or even partly constituted by characteristic physiological profiles, then perhaps the muscular tension and boiling blood characteristic of an angry state disposes one towards different sorts of actions than the sort of physiological state characteristic of the love MLK speaks of. Nonetheless, the emotional state need not represent any particular action to be done. Various Qualities of Heart could dispose us to respective action types indirectly, in virtue of changing the weights given to various of the agent's desires in the decision-making process, and in virtue of changing the salience accorded to different of the agent's beliefs. Thus when I'm overwhelmed by hatred, the belief that that Joe has slighted me and the desire to get revenge might have more salience and more motivational force than another equally accurate belief that Joe is deep down a good person, and the desire to protect and support his well-being. The feeling of brotherly love that MLK takes to be morally positive might have an opposite effect.

To get a handle on these motivational emotional states, we can proceed by identifying the psychological and physiological changes present during such episodes, as well as the activity in the brain or elsewhere in the body that sustains these effects. We need not establish which of these aspects, if any, corresponds to the folk-psychological notion of emotion. Such an approach can address the commonalities and differences between feelings of ill-will and feelings of benevo-

lence, say, while remaining agnostic about whether emotions are a natural kind. This allows us to avoid debates about whether emotions are essentially cognitive or instead body-based, and whether emotions are essentially conscious. This does not mean that substantive aspects of research into emotional reactions are left out; on the contrary, the nature and causal relations of somatic and cognitive aspects may well come into more precise focus when not lumped together, for instance. This approach also allows us to ask whether psychological processes that are especially associated with emotional reactions, such as affect valence, might nonetheless be present in cases where we would not be tempted to attribute an emotional reaction. And indeed, recent empirical work suggests that affect valence is pervasive in human psychology, being implicated in evaluative decision-making about everything from consumer choices to moral judgments (Loewenstein and Lerner, 2003; Haidt, 2007; Lebrecht et al., 2012).

On this an approach we might think of emotional episodes as often involving a cycle of initial perception, triggering associated affective and somatic responses. These in turn can trigger thoughts, which may in turn trigger further affective and somatic reactions, and so on. This opens space for a central suggestion found in psychological models in Buddhist traditions: that unwholesome states of craving and aversion arise in response to an affective tone associated with perceptual representations, rather than directly in response to the object perceived. And this, in turn, provides the crucial opening for therapeutic interventions. Through paying careful attention to our own experience, the Buddhist account claims, we can see that perceptions and their associated affective valence are separate from – and indeed separable from – reactions of craving and aversion, as well as the elaborate thought processes these can motivate (Nyanaponika, 2000; Grabovac et al., 2011).

The central point of the Buddhist model is that while we cannot change the fact that being a conscious being interacting with the world involves both pleasure and pain, we can take responsibility for the pain and pleasure we cause ourselves in reacting to the world. There horrible things that happen in the world, and so on many occasions to perceive things as they are is to perceive things as painful. My suggestion, which I take to be both a Buddhist one and also an empirical claim subject to future investigation, is that these initial affectively valenced appraisals need not

lead to further proliferation in cycles of emotional reaction. This distinction between initial appraisal and subsequent cycles of emotional elaboration is of central importance to my account, because it allows us to separate a number of questions that are apt to be conflated. An important objection to the claim that we ought not to have anger, for instance, is that anger is warranted in response to atrocities such as genocide and rape, as well as in more mundane matter. I agree that is fitting to be pained by such things, or more precisely that it is fitting to be pained by an accurate perception of such things. Empathetic pain in response to seeing another's pain is unpleasant, to take a more mundane example, but it does not follow on my account that we ought not to feel empathy. In terms of the model developed above, I agree that the initial perception and its affective valence are subject to evaluation in terms of warrant, perhaps based on causal considerations. These questions are separable, nonetheless, from questions about how we ought to proceed from there. The project I undertake in this dissertation is to provide a means to normatively evaluate the various possible ways of further reacting to an initial painful or pleasurable perception of things in the world. Any way of reacting strengthens habits of reacting in that same way, and the ethics of emotion that I seek to develop suggests that some ways of reacting to pain and pleasure ought to be cultivated, and others ought to be attenuated. The means I suggest for discerning between these two is pragmatic, even hedonist. Some ways of responding emotionally feel much better than others.

Goodwill Feels Better

Laboratory-based economic games offer one way to study the emotions that underlie ethical behavior and ethical judgments. Fehr and Gächter (2002) used a public goods game in which players were given an endowment of 20 units of money and could decide how much of this to contribute to group projects. In some conditions, after being informed of others' contributions, players could choose to punish those who did not contribute significantly, paying 1 unit from their own endowment to subtract 3 units from the endowment of another. Under conditions in which players knew that they could be punished for not contributing, the mean contribution to group projects was dra-

matically higher. Interestingly, on self-response questionnaires, players indicated a high-level of anger and annoyance towards those who failed to contribute to group projects in a significant way (80.5% indicated a level of 4 or 5 on a 7-point Likert scale), and an even higher degree of anger in those who had contributed a large amount relative to the free-riding player. Reciprocally, players indicated that they would expect others to feel anger towards them if they had similarly failed to contribute to group responses.

Similar emotions seem to drive responses in the Ultimatum Game. In one version of this simulation of economic behavior, one party, the proposer, is given the decision of how to split $10 with a second party, the responder, with the caveat that if the responder rejects the offer, neither gets any. Most equal offers are accepted. But as offers becoming increasingly less equal, they are more likely to be rejected. Chapman et al. (2009) investigated the emotional basis of these decisions by asking participants in an Ultimatum to indicate how well their feelings about the preceding offer were represented by photographs of faces displaying disgust, anger, contempt, fear, sadness, surprise, or happiness. Not surprisingly, ratings of happiness dropped rapidly from a high when offers were equal (5:5) as the offers became less equal, with the proposer offering to keep $7 and give the responder $3, or to keep $9 and give the proposer $1. In contrast, ratings of disgust and anger increased dramatically as the offers became more unequal. Moreover, the researchers found that changes in facial muscles associated with disgust were correlated with subjective identification with disgust faces. This suggests that emotions such as anger and disgust drive the behavioral response of rejecting unfair offers in economic games.

Interestingly, Kirk (2011) found that increased interoceptive awareness due to mindfulness training resulted in more altruistic and less punishing responses in the Ultimatum Game. That is, individuals with increased interoceptive awareness were more willing to accept unfair offers, suggesting that their actions were not driven to the same extent as controls by sentiments of anger and disgust. Kirk et al. characterize their results as suggesting that mindfulness serves to cultivate "rational" responses to economic exchanges in the Ultimatum Game. On the individual level, the costs of retributive response may indeed outweigh the benefits, both in these laboratory-based

economic games and also in daily human interactions. Nonetheless, the enforcement of fairness norms serves an important communal function. And more generally, sentiments such outrage or guilt driving outcome-based moral judgment may be adaptive in terms of reproductive fitness (Fehr and Gächter, 2002). In his classic paper on the emergence of altruistic behavior, Trivers (1971, p. 49) notes that moralistic aggression and indignation might have been selected for as a means to counteract the tendency of the emotional rewards of altruism to drive continual altruistic acts in the absence of reciprocity. This gives rise to a possible explanation of Kirk et al.'s result: if feeling altruistic is emotionally more rewarding than feeling anger or disgust, individuals who are for whatever reason more fully and accurately aware of which emotional motivations are intrinsically punishing, and which are not, might be less motivated to engage in retributive emotional reactions and more motivated to cultivate altruistic ones.

Altruistic behavior can be rewarding in various ways, and in many cases helping others may be merely an instrumental goal in service of getting social or other emotional rewards for oneself. In a series of classic studies, Batson and colleagues tested a range of such egoistic explanations of helping behavior against their own empathy-altruism hypothesis, that egoistic concerns are not the only motivation for altruistic behavior. Put another way, Batson's hypothesis is that some motivations for some altruistic acts do take as their ultimate goal the welfare of the other, rather than the motivation to help another always being proximate, in the sense of being merely a means to some other self-interested ultimate goal. The power of Batson's work lies in the groundbreaking experimental paradigms he used to isolate and test cases in which the posited egoistic concerns could be satisfied without altruistic action to relieve another's suffering. In some experiments, Batson and colleagues manipulated feelings of empathy in subjects by using similarity manipulations, such as providing a filled-out questionnaire indicating that the suffering person's values were similar or dissimilar to those indicated by the subject; in others they induced emotion-specific misattribution, by providing a placebo pill and information either that the pill had the side effect of inducing feelings of personal distress, or else that the that the pill had the side effect of inducing feelings of concern (e.g. Batson et al., 1981). These manipulations allow a test of the empathy-altruism

hypothesis against the competing view that all helping behavior is motivated by egoistic concerns, such as the motivation to reduce the empathic distress brought on by seeing another suffering. In high empathy conditions, individuals provided with an easy possibility for removing themselves from a situation in which they observed another suffering were less likely to choose to escape than in low empathy conditions. This counts against the idea that altruistic helping is driven by a simple goal of reducing aversive arousal. With similar manipulations, Batson and colleagues provided powerful experimental evidence against a range of more sophisticated egoistic explanations of altruistic behavior (Batson and Shaw, 1991), including social evaluation and self-criticism. Using a reaction time task, for instance, individuals in the high empathy condition in one study showed no increases in cognitive association with concepts of social reward such as PROUD, HONOR, PRAISE, but did show a positive association with victim related words. In another study, Batson and colleagues found positive changes in self-reported mood when the suffering of the other was ended, even when the possibility to help was taken away. One especially interesting study addressed a proposed egoistic motivation to gain vicarious or contagious good feelings when the suffering of another is relieved. Contradicting the conclusions of a previous study, Batson et al. (1991) again found support for the empathy-altruism hypothesis. Although there was some evidence for motivation to experience vicarious relief in the low-empathy conditions, individuals in the high-empathy condition did not increase their helping behavior when expecting to receive feedback from the person helped, and did not increase their interest in hearing about the suffering person's status dependent on the likelihood that the person's condition would in fact improve.

Batson's results, if correct, show that we cannot explain altruistic behavior in terms of expected future rewards, even internal ones. They explicitly rule out the hypothesis that subjects "learn through prior reinforcement that, *after* helping those for whom we feel empathy, we can expect a special mood-enhancing pat on the back", for instance (Batson and Shaw 1991: 117, my emphasis). Nonetheless, it is consistent with Batson's results, and perhaps also with the spirit of his proposal, to suggest that people are motivated to act altruistically because feeling altruistic feels good. That is, making another's welfare one's ultimate goal could be motivated by the hedonic

reward of present altruistic emotional motivations themselves. In is important to note that emotional motivations can be seen as distinct from the external stimulation they arise with. When one encounters suffering in the world, such an experience is, for most of us, painful. But the negative affective valence of such external stimuli can be met and held by an internal emotional motivation of friendliness. When this sort of basic friendliness comes into contact with another's success, on this account, it protects one against feeling envy, and manifest instead as what Buddhist translators have termed sympathetic joy. Equally, when this same internal quality of heart of basic friendliness and care comes in contact with suffering, it manifests as compassion, a desire to help. My suggestion is that this basic friendliness, just as it arises and before it can manifest in helping behavior, may itself be more pleasurable than the alternative ways of reacting.

It is conceivable that individuals could be motivated to cultivate an altruistic state just because it feels better to feel that way towards others than the relevant alternative options. On one construal we could take this as just another sort of proximate motivation: one is only motivated to be motivated to help because that desire to help feels good. On the other hand, one could also take this as a claim about the nature of motivation: what it is to be motivated in a certain way is that it feels pleasurable to move oneself towards that goal. On this construal, what makes it the case that helping another is one's ultimate goal is just the fact that the emotional state that results in such actions is experienced as pleasurable. My suggestion is that the basic friendliness underlying both compassion and sympathetic joy is pleasurable regardless of its downstream effects, or more precisely, that it is less unpleasant than the alternative reactions of stinginess or distress. And if individuals are motivated to cultivate this basic friendliness because of its relative ease, simply by being more often in this emotional state rather than its opposites, they would more often be disposed to act for other's welfare in cases of need.

Anecdotal evidence from a few recent investigations supports these suggestions. In a recent pilot investigation of the neural basis of compassion, Tania Singer asked the Buddhist monk Matthieu Ricard to help differentiate compassion from empathic distress by directing his mind for one hour in the fMRI scanner just towards images of suffering, feeling the distress that brings on, without

allowing this to move into the emotions of well-wishing that the Buddhist tradition suggests are so important to train. According to recent comments by both Ricard and Singer, Ricard found this one hour nearly unbearable, and practically begged Singer to allow him to switch to his highly trained manner of responding to suffering with what he describes as "human warmth... love and compassion" (Singer 2010, Mind and Life XXV, 2012). Presumably, an individual with expertise in cultivating states of well-wishing may be better able to compare these states from a first-person perspective with states of personal distress, and to arrive at a discriminating preference regarding which of these states he would rather be in. Even in the absence of explicit cultivation of particular pro-social emotions, however, an increased ability to be fully and accurately aware of the range of emotional experience from altruistic to self-interested may facilitate increased discrimination of which states feel better than others, subjectively.

Valence and Preference

The notion of affective valence, as it is used in recent empirical literature, often conflates a number of separable aspects (Colombetti, 2005). Emotions such as joy are often associated with approach (towards a pleasurable object), and emotions such as sadness with withdrawal. But the dimension of approach and withdrawal needs to be separated from the hedonic tone of an emotion. Both craving and anger motivate approach behavior, for instance, but it does not follow that these are pleasurable; on my view they are both unpleasant. Indeed, it is the unpleasantness of craving or anger that motivates us so powerfully to do whatever seems as if it will appease them. Conversely, I suggest below in Section 2.3.1 that the feeling of goodwill has a positive hedonic tone, and that we can be motivated to act in benevolent ways simply because it feels so good to have the emotion that gives rise to such actions. For my account to be viable, it is crucial that certain physiological reactions, for instance those involved in ill-will, have a negative hedonic tone, for all human beings. Importantly, I do not deny that ill-will is pleasurable for some of us, in addition to being unpleasant. We can like pain, and more generally we can have a preference for things that are negatively hedonically valenced. One way to make sense of this conflict is to suggest that certain

physiological reactions have an intrinsic negative hedonic tone, independent of whether we have a preference for or against these reactions. There empirical as well as phenomenological reasons to be skeptical of such an account of hedonic tone as intrinsic; some theorists have suggested that the (un)pleasantness of a perceptual objects consists in nothing more than that we (dis)like it (e.g Prinz, 2004; Hill, 2009). That is, our own preferences determine our pain and pleasure.

For the purposes of supporting the empirical claim for Convergence in Ethical Judgment, we can be agnostic on this issue. However, if pleasantness is determined just by our preferences, then for CEJ to be plausible, some preferences must be hard-wired and universal. For instance, plausibly, tissue damage is negatively hedonically valenced for any animal. We can make sense of masochism and similar cases either by saying that the masochist has a preference for something that is intrinsically painful, or else by saying that he has conflicting preferences. For making plausible the Claim for Empirical Convergence, either will do. As a further point, I suspect that the cognitive conflict involved in situations that activate conflicting preferences is itself negatively valenced (see Section 3.1). This is what makes it the case, on my view, that being wholehearted is better than being conflicted. The pleasure of feeling goodwill can on some occasions have this kind of purity, I suggest, in virtue of not being mixed with painful feelings. In contrast, on my account, the pleasure that one might take in feeling ill-will towards an enemy will be always mixed with the pain of the physiological reaction involved in ill-will. It is not the case, however, that we are aware of the negative hedonic valence of emotional reactions such as ill-will on every occasion we have such an emotion. Indeed, it is crucial to my account that we often are not accurately aware of the pain and pleasure of our own emotions, but that with the appropriate training of attention, we can come to feel and thus to know the relative painfulness of various emotion types.

There has been controversy over whether basic emotions have the kind of distinct physiological profiles I allude to here (e.g. Barrett, 2006), although recent evidence suggests that some distinct emotions do correlate with distinct autonomic activation (Stephens et al., 2010; Sequeira et al., 2009). Be that as it may, my claim is not about the nature of emotions, nor about specific emotion categories. Even those who hold that such categories are socially constructed may agree that there

are distinct objective features of the physiology and affective valence of the motivational states of human organisms that are then socially construed in various ways. My approach to grounding ethics does depend on the minimal premise that there are at least two sets of motivational states such that the physiology and affective valence of these two sets are both distinct enough from each other and also similar enough across human beings that to the degree individuals are Wide Awake they will converge in preferring the one set over the other. It also depends on the idea that specific motivational states, as individuated by objective measures of physiology and affective valence, will specify distinct ranges of potential action. Barrett (2006, 419) suggests that "there is no one-to-one link between an emotion and a specific state of action readiness in the absence of a specific context or situation." Even so, it could be that certain states of physiology and affective valence will rule out certain ranges of actions in any social context. A certain physiological and affective profile may sometimes be construed as friendliness and sometimes not. On my proposal, nonetheless, the physiological and affective states that are sometimes construed as friendliness are distinct enough from other states and have enough in common between human beings that, to the degree one reacts to the perception of another person with this sort of physiological and affective profile, one will be disposed not to act with the ultimate aim of harming that person. Conversely for the physiological and affective states sometimes construed as hatred; however this objective state is construed, to the degree one reacts to the perception of another person with this very physiological and affective profile, one will be disposed not to act with the ultimate aim of benefiting that person. Assuming that motivations can be individuated in this objective way by measurements of physiology, affective valence, and ranges of action potentiated, my hypothesis is that to the degree individuals are Wide Awake, they will converge on attitudes for certain such objective states and against others, and that this agreement in attitude will be expressed in agreement in ethical judgment. For instance, appealing to my own phenomenology, I suspect that people who are Wide Awake will agree in preferring to be motivated by something like goodwill or friendliness, rather than by something like hatred.

At bottom, then, my approach is based in an appeal to first-personal concerns about what feels

good to an agent, or more precisely which sorts of motivations feel less bad. This account has the somewhat paradoxical implication that egoistic motivations of greed and ill-will might turn out to be a bad strategy for gaining pleasure and avoiding pain for oneself, and that acting out of the selfless motivation to be benevolent towards others might paradoxically be much more personally rewarding. Owen Flanagan rightly goes beyond the common sense that being persistently subject to anger or fear over a long period is bad for the person subject to these emotions; he suggests that such negatively valenced emotions can also be destructive in this way even in a momentary sense. As he puts it "simply experiencing fear or anger over a short episode is destructive to him or her who experiences it, at least in the sense that it is qualitatively unpleasant and produces a sense of unease and disequilibrium" (Flanagan, 2000, 271). I go perhaps beyond Flanagan in suggesting that even emotional states many of us would normally think of us pleasurable, for instance lust, are only perceived as pleasurable because of affective biases cause us not to attend to the equilibrium-disrupting physiological stimuli of these states, and even if attention is captured, not to remember them later when we reflect on the phenomenological character of these states. In effect, I am suggesting that states such as lust and even willful delusion may be like fear and anger states that we are motivated not to experience. Indeed, this is one explanation of their strong motivational force: we are motivated to appease our cravings just because we want the state of craving itself to go away.

One might wonder why evolution would have endowed our psychological systems with such a convoluted structure. After all, we have the pleasure and pain systems that we do because they motivated action that was adaptive. If ill-will or lust were adaptive during the course of animal or human evolution, then one might think that we would be set up so as to be rewarded for having such reactions; it would make no sense for our systems to be set up such that these emotional motivations would have a negative affective valence for us. If ill-will and lust were not adaptive, on the other hand, our reward system might well have set us up such that having such motivations would be painful for us, but equally, the trait of having such motivations would not be universal among human beings in the way that it is.

This objection helps bring out the distinction between what is good for our genes, and what is good for us. Let us assume for the sake of argument that the central aspects of the affective systems that reward us for pursuing pleasure and avoiding pain came to be as they are due primarily to selective pressures. And perhaps we can make sense of evolution by thinking about our genes as having selfish interests in being reproduced (through our actions). Even so, our own interests in pursuing pleasure and avoiding pain, as human beings, can come apart from the interests of our genes in being reproduced. Think for instance of masturbation.[1] Stimulating one's sexual organs on its own serves no direct reproductive advantage. We are motivated to do such things because they are pleasurable. The fact that such things are pleasurable is nonetheless a by-product of a very central way in which evolution has selected those genomes that would tend to be reproduced, that is, by making coitus with a fertile member of the opposite sex pleasurable. Human biology and psychology are replete with such examples. Sickle-cell anemia is a deleterious by-product for bearers of the recessive trait of an adaptation that allows bearers of only one copy to survive conditions of high levels of malaria, and thus be able to reproduce in such contexts, more often than those without this adaptation. Similarly, the disposition of certain human psychological makeups to depression or to drug addiction may impart no reproductive advantage on their own. Nonetheless, these are presumably by-products of certain traits, or of the interaction between certain traits, that did tend to lead to more copies of themselves being reproduced.

Acting out of greed for resources and lust for sex may plausibly lead to greater reproductive fitness. Consider that instead acting out of selfless benevolence for others likely reduces our ability to concentrate our energies on the energy-intensive projects associated with supporting offspring, and that not ever acting out of lust likely further reduces our chances of reproduction. Nonetheless, I want to suggest, it may well be the case that being motivated by benevolence is much more pleasant, or at least much less unpleasant, for the person so motivated, than being motivated by greed and lust. Similarly, I think that vengeful ill-will feels worse than acceptance and equanimity for the person motivated in these various ways, at least to one who is paying attention in the sense I have

[1] Thanks to Fiery Cushman for this example.

called being Wide Awake. These claims would be true, if they are, just in virtue of which sorts of physiological changes are involved in being greedy or benevolent, vengeful or equanimous, and which sorts of physiological changes have for us a negative affective valence. Flanagan (2000, 271) rightly points out that a certain emotion's being destructive in the sense of being characterized by qualitative unease is entirely compatible with it also having been adaptive in the sense of reproductive fitness. Of course, one might still ask why it is that the sorts of physiology involved in ill-will and greed would turn out to have greater negative affective valence than those involved in benevolence or equanimity. If indeed I am right that this is the case, then presumably there is a physiological story to be told about why particular sorts of readiness for action will require certain muscles, blood flow, and hormonal response, and why others will not. But what matters for my story is not why, but rather *that* certain emotional states do have a predominantly negative affective valence in this way, for that is what causes people who feel more fully these respective hedonic valences to judge such motivations as ethically worse. It is just a brute, contingent fact that certain sorts of emotional motivation are characterized by a predominance of negative affect, and others are not. Nonetheless, if my argument in Chapter 5 is cogent, this brute, contingent, but shared feature of human psychology may have important implications for the justification of ethical claims to other human beings and to ourselves.

2.3 Increasing Alertness and Decreasing Affective Bias

If altruistic concern really does feel better than outrage, and the hedonic value of these emotional reactions is directly motivational in the way I have suggested, why aren't more of us more altruistic and less driven by sentiments of anger and guilt, contempt and shame? Moreover, just as first-order attitudes towards male and female genital circumcision, punishments for rape, appropriate deference to authority and so on vary widely between cultures, so too socialized attitudes toward different motivating emotional states *themselves* may vary widely. In one intriguing series of studies, Jeanne Tsai and colleagues have documented religious and cultural differences in

what they call "ideal affect". Using a self-report measure called the Affect Valuation Index (AVI), respondents are asked rate how much they "would IDEALLY like to feel" each on a range of emotion types from High Arousal Negative states such as being fearful or hostile to Low Arousal Negative States such as being dull, and from High Arousal Positive states such as being enthusiastic and elated to Low Arousal Positive states such as being calm and relaxed, as well as more neutral states. Using this approach, Tsai et al. (2006) found that European Americans valued High Arousal Positive (HAP) states more and Low Arousal Positive (LAP) states less than did Hong Kong Chinese. Investigating how such attitudes might be socialized, Tsai et al. (2007a) first replicated this finding with European American and Taiwan Chinese preschoolers. Next, in two sets of the top ten best-selling children's books for each of these societies, facial expressions were coded using the Facial Action Coding System (Ekman & Friesen, 1978). As predicted, compared with the Taiwanese storybooks, the American books had more arousing activities, more excited expressions, and, interestingly, wider smiles, even though there was no overall difference in the number of smiles displayed. Next, moving from correlation to causation, the researchers found that across cultures reading a book about an exciting character versus a calm one increased childrens' preferences for the respective activities and that seeing excited smiles versus calm ones influenced which of the respective smiles was judged as happier. Although tentative, these preliminary results strongly suggest that what I have called higher-order affective attitudes do vary between cultures. Similarly between religious traditions, Tsai et al. (2007b) found that classic and contemporary Christian texts encouraged HAP states more and LAP less than did corresponding Buddhist texts, and that contemporary Christian college students valued HAP more and LAP less than Buddhists. This is an important challenge to the proposal I have made, but I think it can be met.

One explanation for the cross-cultural variation in attitudes towards affective states themselves noted in the work of Tsai, is that through storybooks and other forms of socialization, we acquire various culturally-specific affective biases of attention and memory. For instance, it might well be that social and economic structures in Western cultures train habits of attending to and remembering the intense pleasures of highly stimulating entertainment and consumption, and train attention

away from the unpleasant aspects not only of such intense stimulation, but also of the emotional motivations for seeking it. In contrast, traditional Asian cultures, perhaps especially in Buddhist religious contexts, may train habits of attending and remembering the pleasures of being calm, and train attention towards the agitating and unpleasant aspects of more intense High Arousal Positive states. This suggests that the higher-order affective attitudes held by Buddhists can be socialized through the same mechanisms that give rise to such wide cultural variability in ideal affect and in aesthetic and moral values more generally.

To be fixated on one aspect of experience is to be unconscious of other aspects, so one can counteract affective biases of attention and memory simply being alert on a generalized level to internal interoceptive stimuli and thought imagery as well as to external stimuli. Most of us human beings, most of the time, do not consciously experience all of the stimuli that reaches our sensory receptors. When I walk, I usually do not consciously feel the touch of my feet on the ground with every step. Like- wise, when I get angry, I usually do not consciously feel all of the physiological changes involved, my shoulders tightening, my jaw tensing, and so on. Instead, in selecting among the internal and external stimuli available to our sensory receptors, we fixate attention in certain habitual ways, and many if not all stimuli that fall outside of this narrow range are processed unconsciously. We experience thing as pleasant or unpleasant. When instructed to do so, subjects can switch attention specifically to the affective valence of a stimulus, and become aware of it in the sense of being able to recall, report, and deliberate on the pleasant or unpleasant aspect (Grabenhorst and Rolls, 2011). However, this positive or negative affective valence of visual and other stimuli structures our habits of attention and our behavioral responses, even when these valences are not consciously experienced (Lebrecht et al., 2012). Of particular importance for my ethical account, we are often not fully conscious of all of the affective and physiological aspects of the emotional processes that give rise to our behavior. We might focus for instance on the pleasure that results from appeasing craving. But, in another way, focusing attention on the desirable external object could serve to distract attention from the agitation and lack of ease, the negative affective valence, that characterizes the state of craving itself. Affective biases of attention

and memory function habitually in this way to focus attention certain aspects to the exclusion of others. In virtue of this, affective biases cause certain aspects rather than others to be encoded in memory for later recall, report, and deliberation. This applies to our emotional motivations in that, to the degree one's attention and memory are constrained by affective biases, we are not fully and accurately aware of how we feel.

One form of attention training drawn from Buddhist traditions but now widespread in secular health care, mindfulness meditation, has been correlated with both increased alertness and also with decreases in affective biases of attention and memory. On the one hand, mindfulness training has been associated associated with increased reportability not only of subtle and fleeting external stimuli such as in rapid serial visual presentations (Slagter et al., 2007), but also of subtle somatosensory stimuli involved in emotional reactions (Sze et al., 2010; Silverstein et al., 2011). On its own, however, increased alertness and awareness can be associated with clinical conditions such as anxiety disorders, and in with negative aspects of heightened body awareness involved for instance in hypochondriasis and somatization (Mehling et al., 2009). For this reason, the role of mindfulness in attenuating affective biases of attention and memory may be equally essential to the health outcomes of mindfulness internventions, as well as to the tradition function of 'knowing and seeing things as they are' (Davis and Thompson, 2013). The suggestion that mindfulness decreases attentional distortion is supported by recent unpublished work by Van Dam and colleagues which found that relative to controls, a group receiving mindfulness training exhibited decreased attentional blink with emotional faces (for discussion of the paradigm see Van Dam et al., 2012). Additionally, while the control group showed small decreases or modest increases in subjective distress as a result of the Trier Social Stress Test (TSST; Kirschbaum et al., 1993), the mindfulness group showed large decreases in subjective distress. Further analysis suggested that mindfulness decreased psychological symptoms, in part by improving awareness of emotional imagery, and by generating emotional stability in response to psychosocial stress. One explanation for this and other similar results (e.g. Britton et al., 2011) is that affective biases of attention and memory increase the tendency to return again and again to mental images that spark negative affect, and that

mindfulness decreases such emotional proliferation and rumination by attenuating affective biases of attention and memory (van Vugt et al., 2012; Roberts-Wolfe et al., 2012). It is important to the function of mindfulness in allowing individuals to face hard truths that mindfulness decreases not only negative affective biases, but also biases towards positively valenced stimuli. In accord with this suggestion, Ortner et al. (2007) found that decreases in arousal to negative images were common to both mindfulness training and a relaxation training control group, but that decreases in arousal to positive images were unique to mindfulness training. Moreover, the relative presence of affective biases of attention may help explain why some mind-wandering is maladaptive Killingsworth and Gilbert (2010), and some is not McMillan et al. (2013).

Thus establishing mindfulness can be seen as a practice of attenuating the affective biases of attention and memory that narrow one's awareness, and correspondingly increasing alertness. Drawing in such evidence, mindfulness has been hypothesized more generally to decrease both positive and negative affective biases of attention (Brewer et al., 2012), and correspondingly to decrease positive as well as negative illusions (Flanagan, 2007, 181). Abstracting from any particular method that might be employed to bring about these results, I have termed this process, of attenuating narrow fixations of attention and also increasing generalized alertness, one of becoming more Wide Awake. Of particular importance for my purposes here, by being Wide Awake in this way we can come to feel our own emotional motivations more fully, and to encode more accurately information about how these various types of emotional states feel, for later recall, deliberation, and report.

My prediction, subject to empirical test, is that to the degree individuals attenuate their affective biases in general, and become more fully and accurately aware of their own emotional reactions, they will converge on subjective judgments about which sorts of emotional motivations are more or less unpleasant. I suggested above, for instance, that to the degree one consciously experiences what it is like to be consumed by ill-will or craving, no matter what one's cultural background, these feel much worse than emotions such as good-will and equanimity. These are offered as only as particularly vivid examples; I don't mean to take make substantive claims in advance of the data

about which types of emotion will be perceived as more or less unpleasant. Rather, the central proposal is, first, that simply by increasing alertness and decreasing affective biases, individuals will converge in their attitudes towards various Qualities of Heart such as love and hatred.

Increased emotional awareness may lead individuals to assign a higher reward value to easeful emotional reactions, and a higher punishment value to emotional reactions that proliferate negative affect. I have suggested that basic friendliness is pleasurable, in itself and independent of any downstream effects. Given that we each carry such biases of attention and memory, some due to idiosyncratic personal history but also some socialized and affirmed more generally in the groups to which we belong, we would not expect that everyone actually will consciously feel or be able to report accurately on the pleasure of friendliness or the displeasure of ill-will. My suggestion is that feeling friendly will be less unpleasant than feeling hatred to anyone, to the degree their attention is fully engaged and not made partial by affective biases of attention and memory. Kant develops an analogous account of aesthetic beauty in the Third Critique, as that which is judged to be "an object of *universal* satisfaction" (Kant, 2000, CPJ 5:211). The idea here is that while both aesthetic judgments and personal preferences are made in relation to pleasure and displeasure, judgments of beauty make a special claim to "common validity". Aesthetic judgments are claims about what serves as a ground of satisfaction for anyone, and as such cannot be grounded "in any inclination" or interest peculiar to a particular subject. On this approach, to say that the basic friendliness that underlies compassion as well as sympathetic joy is a beautiful mind-state - this is a direct translation of the Buddhist term *sobhana-cetasika* - is to ascribe to all human beings a grounds for satisfaction in such states. In my terms, such a judgment abstracts from the particular biases of attention and memory that we each carry, to say that for any human being to feel their own motivational states in a full and unbiased way would be to feel the ease of a state of friendliness relative to a state such as ill-will or craving. More specifically, my suggestion is that at least some human Qualities of Heart have physiological and affective profiles that are both distinct enough from each other and also similar enough across human beings that to the degree any of us are Wide Awake, we will come to converge in our preferences regarding which Qualities of Heart we would

rather our actions be motivated by.

2.4 Conclusion

The considerations I have offered in this chapter are designed to show why we might expect Convergence in Ethical Judgement (CEJ) among human beings, to the degree they are Wide Awake. By fully experiencing the relative ease or unease characteristic of various ways of being, I have suggested, one can come to (re)form her affective attitudes such that she is motivated to avoid those Qualities of Heart characterized predominantly by unpleasant negative affect, and motivated to pursue those characterized by greater ease. When she makes ethical judgements about other agents' emotional motivations, she will express the affective attitudes she has formed towards these various emotional motivations. If fully experiencing the relative unease characteristic of hatred has lead her to be motivated not to cultivate a state such as hatred in her own actions, then she will express this second-order affective attitude in the ethical judgment that no one ought to cultivate such an emotional motivation. In this way, increasing awareness and decreasing affective biases would lead to systematic shifts toward particular ethical judgments, regardless of individuals' previous socialization.

It is important to note both that the account only directly accounts for a convergence in ethical judgments about agents' intentions, but also that converging ethical judgments of intentions may constrain those aspects of moral judgment that are more responsive to the outcomes of actions, as it does to some degree over the course of normal child development. A judgment that hatred is a bad Quality of Heart will also count against types of actions that could only be motivated by hatred. Conversely, if those who are Wide Awake will converge on judging the motivational state of friendliness to be praiseworthy, then they will also converge on judging as morally praiseworthy the sorts of compassionate actions that would be done by anyone who was in this state. On the other hand, the theory does not imply any convergence on the moral value or disvalue of types of action except by implication from the moral value or disvalue of the particular Qualities of Heart behind

particular actions. Nonetheless, this gives us a sketch of how establishing mindfulness might lead to convergence in moral judgments of various human Qualities of Heart, despite radically diverging cultural mores. What remains is to show how this psychological claim could bear on the normative claim that we ought to agree with the consensus of who are Wide Awake about how human beings ought to live, if there were to be such a consensus.

Chapter 3

Acting Wide Awake: A Normative Foundation

In Chapter 2, I have drawn on empirical evidence to make plausible, subject to further testing, the empirical hypothesis of Convergence in Ethical Judgment: that to the degree human beings are fully conscious and accurately aware of the interoceptive stimuli characteristic of various emotional motivations, they will converge on ethical judgments about which emotional motivations other agents ought to act on. In the present chapter, I move from this empirical claim to a normative one. In particular, I argue that the ethical judgements human beings would converge on to the degree they are Wide Awake in this way are ones that we all ought to defer to, by our own lights.

On the approach I call Acting Wide Awake, what makes it the case that we ought to act to serve others' welfare, all things being equal, is that anyone motivated by friendliness would be disposed to act in this way. And what makes it the case that we ought to be motivated by friendliness is that we would want to be motivated by friendliness rather than the other alternatives, to the degree we were Wide Awake. Conversely, a behavior that is motivated by ill-will is one that ought not to be done (if it is), just in virtue of the fact that none of us ought to have ill-will. And what makes it the case that we ought not to be motivated by ill-will is that we ourselves would prefer not to be motivated in this way, to the degree we were Wide Awake.

AWA is also committed to the further claim that if we ourselves would prefer not to be motivated by ill-will, to the degree we were Wide Awake, that is so just in virtue of the fact that the neurophysiology involved in being motivated by ill-will has a predominantly negative affective valence for any human being. However, there are two importantly different ways to construe this last claim about the hedonic valence of particular Qualities of Heart, either as justification for ethical claims or instead as an explanation of the ethical that we should defer to. Taken in the first way, AWA offers a constructivist approach, justifying ethical norms by appeal to epistemic ones. Taken in the second way, AWA offers a reduction of ethical facts to psychological ones. I consider these metaethical issues in more detail in Chapter 5. In the present chapter, I set aside these issues, focusing primarily on concerns that affect AWA as a (first-order) ethical theory. In Section 3.2 I argue that by taking Qualities of Heart as the evaluative focal point, AWA avoids many of the standard objections to ethical egoism. However, one important question is why we should agree that a Quality of Heart that has a negative valence that we never perceive, an unconscious negative valence, is in any sense bad. In answering this objection, I appeal to epistemic norms of full and accurate awareness. Like ethical norms, epistemic ones can be given various meta-ethical construals, though I set that issue aside here. My central suggestion is that if human beings share the motivation to avoid pain, and also enough of the neurobiology of human emotions, a certain circumscribed set of ethical claims have force for all of us, on our own terms.

3.1 Knowing How to Live Wholeheartedly

In describing the state of being fully and accurately aware as correcting distortions of perception, thought, and judgment, I have been implicitly appealing to a general epistemic norm that privileges knowledge over ignorance. One neat way for cashing out in practical terms of the force of such epistemic norms is offered by Gibbard (2003). Gibbard brackets the longstanding and complex debates about the proper criterion of knowledge by recognizing that to attribute knowledge in ordinary cases is to say that we can rely on the relevant judgment. Gibbard suggests that when

a speaker, Steve, says that someone, Sally, knows something, he is expressing a plan to rely on Sally's judgment in this matter. Steve might not know about things such as the soil on Mars or neuroscience of affect, and when he says to himself or to another, Sam, that Sally does know about these things, Steve means that if Sam did need for some reason to get reliable information about such things, then Sally is someone whose judgments about such things he could rely on. Interestingly, Gibbard cashes out ethical norms in a similar way. When Jessie says it would be wrong for Joe to stone someone for having been raped, Jessie is expressing a plan for what to do if in Joe's shoes in the relevant situation; specifically she is expressing a plan not to stone people for being raped. Jessie is thereby disagreeing with those who say it would would be honorable or obligatory or even permissible to do such a thing, since they are expressing a plan for what to do if in Joe's shoes that is inconsistent with Jessie's plan for that situation. On Gibbard's analysis, to say that Jessie *knows how to live* is to express a plan to rely on Jessie's judgments about which plans for living to adopt. The notion of wisdom may be much more complex than this. Nonetheless, if the affective attitudes that dispose us to rely on one person's judgments rather than another's are in principle psychologically tractable phenomena, Gibbard's approach opens the way for a naturalistic account of at least one of the characteristics we are tracking when we call someone wise.

I have suggested that to the degree a person is fully and accurately aware of internal and external stimuli, they will be a reliable judge of which sorts of emotional motivation involve great unease and which do not. There might be other ways to gain knowledge about the relative ease and unease characteristic of various qualities of being alive. Suppose we can, using present or future technology, get a reasonably reliable empirical handle on the physiological and neural activity that realizes negative affective valence. What this would mean is that when and only when the objective indicators of negative affective are present, if we ask subjects to attend in the right way (as, for instance, Grabenhorst and Rolls (2010) do), then the subjects will report an unpleasant aspect in their experience. Importantly, however, it would often be the case that these neurophysiological measures are present when subjects don't attend in that way and so don't report experiencing any

unpleasantness; we would in this case have an objective measure of unconscious affective valence. Such technology would afford us one means of knowing, in a rough sort of way, which Qualities of Heart involve negative affect valence. But it might not make us wise. The reason, I will suggest, is that objective indicators matter only insofar as they are indicators of states that any of us would feel subjectively in a certain way, and therefore be motivated by, were we in the right subjective position.

Suppose I hook up my neuroimaging setup to myself, and come to know through this technological means that the emotional state of greed or of ill-will I am currently in is actually characterized by preponderance of negative affective valence. Still, being fixated on the thought of the thing I am hating or lusting after, I do not consciously experience any unease of being in this emotional state. Others who wanted to know about the relative unconscious affective valence characteristic of various momentary states of being alive, from making love to getting in argument to volunteering help to someone in need, could ask me, and as I long as I am hooked up to the machine, my judgments about such things might be reliable. Still, I might be knowledgable about which ways of being alive are characterized by various affective valences without thereby knowing how to live.

Compare a case in which I set aside the neuroimaging machine and just carefully, persistently attend to the full realm of present experience so as to heighten generalized alertness and attenuate affective biases of attention. As with the technological setup, it might be that the emotional state of greed or of ill-will I am currently in is actually characterized by preponderance of negative affective valence and yet, being fixated on the thought of the thing I am hating or lusting after, I do not consciously experience any unease of being in this emotional state. With this natural technology of attention, as with the neuroimaging setup, I can become knowledgable about which ways of being are characterized by positive or negative affect. From the first-personal standpoint, however, I do this by becoming conscious of more of the external and internal stimuli reaching my sense receptors. I come to consciously experience negative and positive affect that was previously unconscious. The difference between the two technologies lies in the fact that consciously feeling negative and positive affect, as unease and as pleasure, are directly motivational in way

that watching indicators of affect on a neuroimaging device cannot be. Armed with knowledge from neuroimaging my unconscious affect, I *might* make plans for living that are aimed to avoid being alive in ways characterized by negative affect, *if* by imagining the unpleasantness of such situations with sufficient vividness, I successfully motivate myself to plan so as to avoid them. In contrast, to the degree I fully experience the unease characteristic of being in a state of greed or ill-will, I can't help but be motivated to avoid being in such situations again. In a moment when I am Wide Awake in this way, I *will* make plans for living that are aimed to avoid being alive in ways characterized by negative affect. If Gibbard is right, such plans will be expressed in ethical judgment. By fully experiencing the relative ease or unease characteristic of various ways of being, one can come to (re)form her affective attitudes such that she is motivated to avoid those characterized predominantly by unpleasant negative affect, and motivated to pursue those characterized by greater ease. When she makes ethical judgements about agents' emotional motivations, she will express the affective attitudes she has formed towards these various emotional motivations. If fully experiencing the relative unease characteristic of ill-will has lead her to be motivated not to cultivate such states herself, then she will express this second-order affective attitude in the ethical judgment that no one ought to cultivate such an emotional motivation.

These are meant as empirical hypotheses, rather than as normative claims about how we should be expressing our attitudes. When King suggests that "violence is immoral because it thrives on hatred rather than love," one can understand this proposal as appealing to an intuition that deep down people would rather be motivated by love than by hatred, by what we might call positively valenced "qualities of heart" rather than negative ones. In noted in Chapter 2 recented empirical evidence suggests that people may indeed have such intuitions. Pizarro et al. (2003) found that Western participants assume that an agent has positive meta-desires, such as the desire to desire to help or the desire to desire not to hurt, unless this assumption is explicitly cancelled. Knobe and colleagues (unpublished) have found evidence that this default attribution of positive meta-desires is meditated by a more fundamental and more general attribution to others of a deep or fundamental goodness that comes out in certain ethical decisions (cf. Newman, Knobe, and Bloom,

in preparation). It is unclear what precisely is being attributed in such cases. But if my way of making this suggestion precise in Chapter 2 is correct, there is a kernel of truth to the folk conception. In particular, I proposed there that the sort of ethical evaluation of the emotional motivations of agents that is evident in adult human beings' judgments is an implicit expression of the speaker's own attitude toward the emotional motivation in question. The claim there was a descriptive one, that this is the way this aspect of adult human judgment works. The prediction of systematic shifts towards this mental-state responsive system of ethical judgment and away from outcome-responsive judgments driven by anger, due to increased emotional awareness, is also a descriptive claim. When subjects in Inbar et al. (2012) judge the disaster-bond investor's desire for harm to occur as an ethically bad emotional motivation, my account suggests that what is going on is an implicit assumption that deep down the investor himself would prefer not to be motivated in this way; it is in virtue of this implicit assumption that we judge that motivation as a bad one.[1] So far, these are all descriptive, empirical claims.

Gibbard's project is self-consciously a descriptive theory about the psychology behind ethical judgment, whereas mine can be construed as in part a normative project, of spelling out which grounds we should appeal to in justifying our ethical judgments. This is a complex issue that I examine in more detail in Chapter 5. One notable difference between Gibbard's and my own naturalistic approach to claims of ethical knowledge is that Gibbard's anti-realist account ultimately gives us nothing to adjudicate radically different evaluative worlds. Two radically different sets of plans for living may each be internally consistent. If so, the radically different evaluative judgments such planners make, expressing their radically different plans, will each have an equally legitimate claim to evaluative truth, or at least evaluative quasi-truth. AWA supplies a ground from which to adjudicate such disputes. In particular, AWA takes it to be the case that all human beings, at least, share the basic project of avoiding unease. This is a concern built into the affective systems of human psychology, and perhaps more broadly animal psychology as well. It is a concern that

[1]This is why, as I suggest in Chapter 5, for beings where we can't assume this sort of emotional makeup, all bets are off. In such cases, there is no reason to expect that our intuitions of right and wrong have any purchase.

we can plan to override in certain specified cases. However, if it is affective valence that serves as the basic currency of human and animal motivation, then we can only override the motivational pull of one affective state by employing the opposite force of another affective state. The ascetic who plans to deny himself every sensual pleasure must use a thought of some end to do so, a thought that must itself have enough affective force to beat out the opposing affective pulls on his motivational system. The thought of some eternal reward might be pleasant, but at the very least the thought of giving in to the pursuit of sensual pleasures must have for him a stronger negative affective valence than the thought of not giving in in this way, however he conceives of those two. If so, it is practically inconsistent to plan to override the motivational force of affective valence in every case. To be internally consistent, otherwise radically different plans for a human life must nonetheless each be consistent with a plan that includes being motivated by affective states. These basic reasons for action may have force for psychopaths and perhaps much more broadly in the animal kingdom. My suggestion that the various Qualities of Heart have distinct hedonic reward and punishment values, in contrast, is intended to be much more narrowly constrained to human beings, or perhaps to some mammals sufficiently like us. In idealized cases of psychopathy it seems that even if there were sensitivity to the affective valence of external stimuli sufficient for instance for the the psychopath's pursuit of pleasure to be constrained or channeled to some degree by negative social consequences, still there might not be sensitivity to the affective valence of emotional motivations in general (see e.g. Holmqvist, 2008). Alternatively, if emotional motivations such as hatred must overcome the negative feelings associated in us with imagining from another's perspective suffering we might cause them, this form of affective restrain might be absent in psychopaths (see e.g. Decety et al., 2013). As I suggest below, however, such sensitivities may be a condition for being included in the interpersonal practices of making and debating ethical claims. If so, and various Qualities of Heart do have distinct hedonic reward and punishment values, this puts a further constraint on which plans for living can be internally consistent for those human beings whose moral judgments we care about.

Without these constraints, AWA as an ethical theory might look like just an account of one way

in which a particularly odd sort of social group might construct a system of values for themselves. Instead of just socializing values for or against homosexual activity, for or against certain patterns of resource distribution, in the normal ways of transmitting sentiments from one generation to another, this community socializes a particular set of sentiments for or against particular emotional motivations, in favor of Qualities of Heart such as benevolence and against Qualities of Heart such as greed and ill-will. This particular socialization happens to be effected by directing attention so as to consciously experience affective aspects of these emotional states that were previously unconscious. But nothing about this quirky means of socializing its values would seem to give this group any authority to dictate how the rest of us ought to evaluate things aesthetically, much less ethically. What hasn't been shown is why individuals outside of this group *should* take emotional states with negative affective valence to be bad, even at times when they don't consciously experience this valence as unpleasantness. After all, a different social group might equally construct an utterly opposed set of values by socializing habits of being perfectly unconscious of the affective valence of any of their emotional motivations, and so never judge the associated negative valences as bad. Indeed, my account in Chapter 2 of how we can become Wide Awake could also be used to suggest a way in which we could become instead narrowly asleep. That is, one might cultivate habits of attention and memory so as to be selectively unaware of how it feels to have the motivations one has.

However, there are deep asymmetries between being perfectly sensitive and being perfectly insensitive. First, to the degree one is unable to feel the force of emotions such as friendliness and guilt, one may thereby forfeit a great many human goods. Considering cases of extreme emotional insensitivity, such as in psychopathy, it is difficult to make the case that such individuals are better off than more normal individuals even on a purely hedonic level. Even if by opting out of the emotional pull of ethical principles one can avoid being subject to an ethical imperative to be at least somewhat sensitive in this way, there may still be an overwhelming pragmatic reason to do so. Secondly, at least for those of us with enough sensitivity to the pulls of our emotions to engage in normal human exchanges of friendship and the like, I am doubtful that perfect insensitivity,

enough to avoid entirely the effects of unconscious cognitive conflict, can be accomplished through an individual's own cultivation of habits of delusion in this way. We are vulnerable to the vagaries of attentional capture: the affective valence of our emotional reactions can capture our attention, and force itself into our conscious experience, despite our best efforts to distract ourselves. I have suggested that, due to the vagaries of attentional capture, we may not be able to completely avoid feeling the negative affective valence of an emotional motivation such as hatred. If so, to the degree a human being chooses to cultivate and act out of Qualities of Heart that are characterized by negative affective valence, she is subject to experiencing that valence as unease. Being motivated as we all are to avoid such states, for a human being to hold a plan for living that involves cultivating and acting out of Qualities of Heart she experiences as uneasy, is to make herself subject to a persistent kind of cognitive dissonance. More poetically, my suggestion is that we can never be truly wholehearted about a plan of life that we see as requiring us to act out of Qualities of Heart that are characterized by unease. On a psychological level as well as on a logical one, there may be a deep asymmetry between being wholehearted and not being wholehearted: one could only be half-heartedly committed to a value of being half-hearted in all that one does, whereas it seems at least logically possible that one could be wholeheartedly committed to being wholehearted in all that one does.

Maintaining theoretical consistency in one's plans for living is itself an epistemic value that we cannot assume to be shared. However, the motivation to avoid cognitive dissonance may be more basic and more widely shared. Although Festinger's (1957) theory of cognitive dissonance as a core drive has been controversial, a recent review by Gawronski (2012) argues that this core drive imposes a universal constraint on belief systems across cultures and species. He notes, for instance, that the motivation to avoid cognitive dissonance seems to be present even in rats (Lydall et al., 2010). These sorts of results, in contrast to some of the earlier research, seem to suggest a kind of dissonance that is not dependent on language, but is instead a more basic and thereby broadly shared human motivation. There is also anecdotal evidence in humans that the effects of cognitive dissonance on unconscious levels may surface in symptoms that are consciously felt. Jonathan

Bennett takes Heinrich Himmler's remarks in a 1943 speech to indicate that the principles the architect of the Final Solution held to be moral ones forced him to overcome emotional reactions of sympathy that he did nonetheless feel for the living beings he was responsible for exterminating. Himmler apparently suffered from nausea, stomach-convulsions, and various other complaints; Bennett (1974, 129) quotes and endorses the characterization by Kersten, Himmler's physician, of these ailments of Himmler's as "the expression of a psychic division which extended over his whole life".

The question of whether it is in fact possible to be more wholehearted by choosing to cultivate and act out of Qualities of Heart such as friendliness rather than hatred would have to be established empirically, perhaps by employing the construct of cognitive dissonance and its absence. In advance of the evidence, nonetheless, it seems plausible that for a human being the choice to devote oneself to a life of evil, for instance to devote oneself to a life of cultivating and acting out of greed and hatred rather than good-will, would result in a pervasive if subtle form of inner conflict and unease. Of course, human conditions being less than ideal, any life path we could choose is going to involve some amount of unpleasant experience. I have suggested that although some of the negative and positive affective valence we are subject to is due to our external situation, in fact much more is due to our internal reactions to these external valences. Sensual pleasures are fleeting: we spend much more time wanting what we haven't got. Nonetheless, any plan for a human life will include conditions such as separation from the things and the people we are fond of, illness, and death. So one might suggest that on my view adopting any plan for living will involve some amount of cognitive dissonance. That might be right, and still to minimize the dissonance might require choosing plans for living that minimize the suffering we cause ourselves by reacting to these negative conditions with proliferated negative emotions. Alternatively, it might be that we can be wholehearted about choosing the path that is the least bad hedonically, and that living one's life from the motivation of easeful qualities of heart is the least bed plan hedonically, even though it has some hedonic costs. For instance, I suspect that being motivated by the kind of love we see in leaders such as Martin Luther King gives one a life that feels less bad on the whole not only than

a life of being motivated by hatred to fight injustice but also that feels less bad than a life being motivated by apathy not to fight injustice. Even though being motivated by apathy to stay out of the fight might allow one to avoid being imprisoned and many other negatively valenced experiences, still the hedonic weight of one's own qualities of heart is greater. Because it is our own emotional motivations, of love or of hatred or of apathy, that we must live with in every moment, the hedonic valence of these would have the most pull when imagine different plans for living in a full and unbiased way on the basis of a full and unbiased experience of what it is like to be motivated in various moments in these different ways. I think it is because we can't help but feel the force of this that we can't imagine planning a life devoted to hatred in the wholehearted way that we can plan a life devoted to love (carrying those out being of course another matter).

Ultimately, it is this hedonic asymmetry, between choosing to cultivate Qualities of Heart characterized by ease and those characterized by unease, that underwrites my claim that all of us have reason to be one way and not the other. The sorts of beings that take sides in the ethical debates we get into, human beings, are each subject to the motivational force of ease and un-ease. The value of this project of avoiding unease, therefore, is a shared premise on both sides of any such debate. And I have suggested that to the degree one is Wide Awake one can come to be a more reliable judge of which ways of living are characterized by ease rather than by unease. Such a characterization is compatible with a number of empirical and philosophical accounts of wisdom. Summarizing empirical work on folk notions of wisdom, Staudinger and Glück (2011) note characteristics such as "experience-based body of broad and deep life knowledge", "a balanced manner of regulating their own emotions rather than getting carried away by strong feelings", and that such individuals "transcend their self-interests and care deeply for the well-being of others". In a more philosophical examination of moral expertise, Driver (2006) likewise suggests that some of the marks we use to identify moral experts include breadth and depth of actual experience with a particular area of human life. In the ways she details these characteristics, one who does not fail to be conscious of their own emotional motivations but is instead more fully and accurately aware of internal as well as external stimuli would seem thereby to gain a certain kind of moral expertise. But we can draw

out the epistemic force of AWA's proposal in perhaps a more stark way just by considering the relevant alternatives. Suppose that two human beings have equal epistemic access to the external world; in particular, suppose they know all the same facts about the impacts of their own actions on other's welfare and suffering. Still, on my account, they may make different judgments about which sorts of actions ought to be done because they may not have the same preferences regarding various Qualities of Heart. One who favors compassion may express this attitude in positive moral evaluation of actions intended to bring about welfare and not harm, whereas, one who prioritizes other Qualities of Heart may not agree. The point of the arguments sketched above is that because the project of avoiding unease is one that we all share, we all ought to privilege the judgments of those who are more fully and accurately aware of which Qualities of Heart are characterized by ease over the judgements of those who are less Wide Awake in this way. To borrow Gibbard's locution, we can say that to the degree an individual is Wide Awake, they know how to live.

To the degree I recognize that I am sometimes lacking in wisdom, I may decide to rely on those who are wise. Even if I am not on my own motivated to act as they are, I can use their example to as a motivation. And even if I am not intuitively inclined to make the same judgments about how to live that they are, I can decide to rely on the judgements they make more than I rely on my own. The burning question, of course, is whose judgement to trust. AWA provides a ground for such trust that is conditional on the results of a sort of empirical prediction, a prediction that to the degree we manage to be more fully and accurately aware of own own lived experience, our judgments too will converge with those who are already knowledgable in this first-personal way. If sometimes I am more Wide Awake in this way and at other times I am not, then I can rely on the judgements about how to live that I make when I am Wide Awake in this way, even a times when I am not so Wide Awake. And if certain individuals are in general more Wide Awake in this way than others generally are, then those of us who are not so Wide Awake can rely example of those who are better first-personal judges of the relative ease afforded by various ways of being. We can rely on their actions, and also on their speech acts of endorsing certain ways of being and criticizing others. In short, we can *act* Wide Awake, even when we are not. This is the account of

the grounds of moral judgment that I have called Acting Wide Awake (AWA).

3.2 Objections to Hedonism

To the degree AWA is construed as reducing ethical claims to facts about which ways of being are hedonistically preferable, standard objections to hedonistic views may be applicable to Acting Wide Awake as well. In Chapter 5, I offer an alternative constructivist construal of AWA. Nonetheless, those who see independent reasons to reject constructivism in favor of naturalistic reduction can still embrace the conclusions of AWA, to the degree it can overcome the standard objections to canonical hedonist accounts, which are more focused than AWA on the affective valence of external stimuli. Here, Fred Feldman's (2004) survey of possible hedonist approaches and the objections to them will be of particular service.[2]

The first and perhaps most obvious objection to hedonism I will call Pleasure in the Bad. The basic problem is that some people (and pigs) take pleasure in what is base, bad, and evil. Noting variants of this complaint from Aristotle, Broad, Brentano, and Moore, Feldman refines from these sources two examples designed penetrate some superficial sorts of response from the hedonist camp. A terrorist who bombs a schoolyard and is thrilled by the misery of his victims serves as an example of pleasure in evil. A pig who wallows in sexual debauchery with the sows in the pigpen serves as an example of pleasure in what is base. To bring out the objection here, we can note

[2]There are differences between the conception of hedonism Feldman is operating with and my proposal, Acting Wide Awake, even construed as a reduction of ethical claims to naturalistic ones about hedonistic concerns. In particular, Feldman takes hedonism as a doctrine about makes a life go well or not, specifically, the doctrine that pleasure and the lack of pain is what a makes life go well. I prefer to take the momentary quality of being alive, rather than the sum of the pleasure or pain of a lifetime, as the evaluative focus of at this foundational level. The quality of being alive can change from moment to moment. Nonetheless, there are deep similarities. Just as the hedonist theories Feldman surveys take pleasure to be the only "intrinsic" good, AWA can similarly assume that subjects could judge the relative desirability of a state of affairs, its goodness or badness, in virtue of the e relative ease and unease characteristic of experiential states involved. The unpleasant feeling associated seeing others' suffering cause one to judge such a state of affairs as bad. Likewise, it is the ease being that is characteristic of being motivated in a benevolent way that makes this a good state to be in.

that a pig could conceivably have more intense and long-lasting ecstasy in his debauchery than a philosopher finds in her intellectual reveries. And philosophers have tended to doubt that the pig's pleasure is equally valuable (admittedly, we have yet to hear back from pigs on this one). The objection can be made more forcefully, perhaps, for the pleasures of terrorism. It seems possible that the terrorist is equally thrilled by killing small children as another man (or the same man, the terrorist himself) is thrilled by being a good father to his own children. But intuitively, it seems absurd to claim that because the two are equally pleasurable they are equally ethically good. Transposed to take direct aim at my empirically-based theory, the objection is that even if subjects were to judge that they experience equal amounts of pleasure from terrorism and debauchery as they do from fatherhood or philosophy, they would not thereby judge the experiential states involved in these to be equally ethically good. If so, judgments of the ethical goodness of various ways of being cannot depend solely on relative pleasantness. A different objection to AWA is that some people *would* judge the pleasure of terrorism to be equally (or more) ethically valuable as those of fatherhood; if so, so much the worse for basing a theory of value on empirical facts about the judgments people do or would make. The fact that some people take pleasure in causing others harm is precisely the problem, so it seems that an ethical theory based ultimately on the hedonic valence of emotional motivations, or anything else, falls into absurdity.

It is a neat feature of AWA that it overcomes both of these apparently opposite objections with the one and the same point. What it is to be Wide Awake, on my definition, is to experience fully external and internal stimuli that reach one's sensory systems, in the way that is allowed by the attenuation of affective biases of attention, and to be able to accurately recall, deliberate, and report on one's experience in the way that is allowed by the attenuation of affective biases of memory. Such ideal perceptual observers will be accurate judges of the relative ease or unease characteristic of various ways of being. I grant that given certain conditions, those of a pig or of a terrorist, the tactile and visual stimuli caused by wallowing in the mud with sows or seeing children blown to bits might be pleasurable. However, as noted above, I predict that individuals who are Wide Awake will judge whatever positive affective valence these external stimuli have for them to

be vastly outweighed by the affective valence of their internal motivational states. In particular, I predict that individuals who are fully and accurately aware of how it feels to be consumed by lust or by hatred will judge the unease of these states to vastly outweigh any pleasure that could be gained by acting out of these motivations as the pig or the terrorist does. In this sense the pleasures of the pig or the terrorist are base or evil, on my account, just because they are actions that involve much more unease than than ease for the pig or the terrorist. It is because the pig and the terrorist don't realize the great unease they are choosing that they act as they do; and it is because one who is Wide Awake feels fully how much unease is involved in such actions that he would not do such things. Judges who are accurate about such things would not judge that lustily pursuing sows in the mud as on the whole pleasurable, on my account, because being consumed by lust is such a uneasy way of being. The corresponding objection proposes that since people have all sorts of different moral values, some would judge blowing up children up to both pleasurable and also ethically valuable. This objection we can overcome by denying that judges who are accurate about such things would judge the emotional motivations required to intentionally blow up children up to be either pleasurable or ethically permissible.

Feldman considers two further objections to hedonism based on a comparison of possible worlds. The first is raised by Moore, who asks us to consider on the one hand a perfectly beautiful world and on the other a "heap of filth". In neither world are there any subjects to be pleased by beauty or be repulsed by the ugly. Yet, the objection goes, it would be rational to consider the beautiful world better than the ugly one. If so, pleasure is not the only thing good. Like Feldman, I consider this argument to be without force. Perhaps unlike Feldman, I think that impersonal, logical considerations about intrinsic goodness and rationality obscure the issue in these sorts of cases. The sentimentalist has a ready explanation of why we would judge the beautiful world to be better than the ugly one. What Moore has done is to explicitly ask each of us to imagine a situation we, in particular, have positive affective dispositions towards, "Let us imagine one world exceedingly beautiful. Imagine it as beautiful as you can; put into it whatever on this earth you most admire..." (Moore, *Principia Ethica* (1903/1962) 84 in Feldman 2004: 51). Despite Moore's stipulation, it is

not the case that no one is being pleased by this world: *you* are, the person he asks to imagine and to judge this world. If we express our own pleasure in the world we have imagined to suit our tastes by judging it as good, this should come as no surprise on the level of psychological explanation. And on the philosophical level, it hardly shows that there is goodness in the absence of pleasure.

Interestingly, Feldman takes as much more powerful a related sort of thought experiment due to W. D. Ross. The idea here is two imagine two states of the universe perfectly equal in the total amount of pleasure and pain, and also equal in the total amount of virtue and vice. However, in one world it is the virtuous people who have all the pleasure, while in the other it is the vicious. Call this the Objection from Unjust Worlds. "Very few people would hesitate to say," Ross suggests, that a world in which the virtuous are happy and the vicious miserable is a better world than one in which the vicious are happy and the virtuous miserable (Ross, *The Right and the Good* (1930), 138 in Feldman 2004, 52). The first thing to note about this is that Ross is making an empirical prediction, of the sort amenable to the methods of experimental psychology, by claiming that a majority of people would make a particular judgment. And no doubt his prediction is a safe one, when made for a normal population of American adults. But the thesis of AWA is that we shouldn't trust a normal population of subjects to determine what is good or right, given the affective biases of attention and memory that human beings normally inherit and are socialized into. Indeed, those of us who, like Ross, hope for some sort of consensus on a set of ethical claims can't trust the judgments of normal populations of subjects to deliver this, because the variability between different human groups and subgroups is so high. The AWA approach explains these radical divergences between human groups as due to affective biases of attention and memory, biases that can be attenuated. In Chapter 2 I noted evidence that outcome-responsive moral judgments of blame and just-desert are driven fundamentally by emotional reactions of anger. And I noted suggestive evidence that by developing emotional awareness, punitive reactions are decreased (Kirk, 2011). Feldman takes Ross' objection so seriously that he formulates a "desert-adjusted" adjunct to his hedonist account, specifically designed to accommodate the intuition that a world in which there are just deserts is a better one than a world in which there is equal suffering not apportioned to those who have acted

badly. I say instead that the judgment that a world in which there are just deserts is better is due to a particular type of emotional reaction on the part of the person making that judgment, and that since the particular type of emotional reaction in question is one we ought not to have, the judgment, too, is one we ought not to have. Justice and Fairness do not have a deep meta-ethical grounding on my account. This is not to say that they have no weight as normative factors. If it happens, as seems likely due to conditions of human psychology, that rewarding people for cultivating bad intentions has the consequence that more such bad intentions are cultivated, then we ought not to set up our societal institutions in this way. Likewise, if unequal distribution of resources creates conditions in which bad Qualities of Heart proliferate, then we ought to act to change things, just because we ought to act out of care for others' welfare. But this hardly refutes hedonism, for according to AWA the foundational reason that either of these institutional factors count as legitimate factors in ethical deliberation turns on the hedonic quality of the emotional motivations involved.

The objections to hedonism surveyed above try to show that there are some pleasures that are not good. A different sort of objection tries to show that there are some things that would be judged as good that are not pleasurable. Feldman gives the example of a Stoic who prefers a life without pleasure or pain, because both would disrupt the peace he wants as an end in itself. The objection to hedonism here is that if for the Stoic a state of not having pleasure is more desirable, better, than a state of having pleasure, then pleasure cannot be what makes these states of peace good ones. Call this the Objection from Non-Existent Pleasures. This objection brings out an important difference between AWA and the hedonistic theories Feldman surveys. My proposal focuses on unpleasantness and its absence, giving priority to these over considerations about pleasantness and its presence. In effect, this makes a state that totally lacks negative affective valence more valuable than a state with the most pleasure possible. I use the term ease to mean a lack of unpleasantness, rather than the presence of pleasure. Total ease is a complete lack of unpleasant experience, regardless of whether this is accompanied by pleasure. It is an empirical prediction of mine that individuals who are more fully and accurately aware of how various ways of being feel will judge ease to be more desirable than pleasure in this way. In part, this move is inspired by suggestions

found in Buddhist traditions of practice, which aim at an ideal of being Wide Awake in the way I have suggested, that such ideal judges will judge the very presence of pleasure to be subtly agitating. But this Buddhist suggestion is a plausible one in light of the bias of the affective motivation system, which prioritizes the avoidance of pain more highly than the pursuit of pleasure, no doubt because of evolutionary causes. In my terms, the suggestion is that from a first-person perspective, even subtle forms of pleasure involve subtle forms of unease, and thus a state that lacks both pleasant and also unpleasant affect is the most desirable.[3] In effect, AWA can agree with the Objection from Non-Existent Pleasures that some things that are not pleasurable are still judged good, in particular the absence of unpleasantness may be even more highly valued than the presence of pleasure. If so, however, this consideration lends support to the ethical value AWA attributes to various Qualities of Heart in particular to the degree that they are free of negative affective valence.

A further sort of objection is given force by thought experiments such as Nozick's experience machine, which gives one a merely virtual experience of having lived a pleasurable life replete with accomplishments and relationships that seem perfectly real to the one hooked up to the experience machine. Nagel gives the less outlandish example of a deluded businessman, one who dies contended, thinking of himself as having enjoyed the love of his family, the respect of his community, and the good fortune and good teamwork to have grown a successful business. Unbeknownst to him, we are told, these people were merely feigning love, respect, and camaraderie. In both cases, then, the subject takes himself to have a perfectly pleasurable life, but since we have

[3]In one early Buddhist dialogue (AN.IX.34), a monk named Sariputta, taken by the tradition to be Wide Awake in the sense I define it, tells us that *nibbāna*, the goal of practice and the highest ethical ideal, is pleasurable (*sukkha*). Another monk, bewildered because this goal is often described as involving the cessation of feelings of pleasure as well as of pain, asks how it could be that there would be pleasure in a state in which there is nothing felt. Sariputta replies, "Just that, friend, is pleasurable in that, that there is nothing felt," and goes on to discuss how pleasures based on external sensory objects, and even the pleasures of deep concentration, can each be a more or less subtle affliction for him (*ābādho*). "Just as for one who is healthy, to the degree disease were to arise, that would be an affliction." Just so, even the subtle cognitive states associated with the pleasure of being deep concentration are a subtle form of affliction. "And whatever is an affliction, friend, is said by the Blessed One [the Buddha] to be suffering (*dukkha*). By this line of reasoning, *nibbāna* can be seen as pleasurable (*sukkha*)."

information the subject does not, we cannot bring ourselves to judge his life as a good one. Call this the Objection from Unknown Deceit.

We can respond to the Objection from Unknown Deceit by challenging the stipulated conditions as unrealistic, and so our intuitions about such cases as untrustworthy. The experience machine is obviously outside of the domain with which our systems of ethical judgments were developed to cope, but even the case of the deluded businessman is remarkable, requiring that he go through years or decades of his life without taking note of telltale signs - in the fleeting expressions, tone of voice, and daily habits of his most intimate associates - that every one of them is deceiving him about their most basic attitudes toward him. If this is an empirical possibility at all, it would require a great deal of self-delusion. It would require on the perceptual level that the businessman not pick up on the subtle and fleeting micro-expressions that betray people's feelings before they are able to inhibit them (Ekman and Friesen, 1969), for instance. On the level of reasoning, it would require that he not recall the subtle but pervasive sorts of information that ought to give him doubts, and that when he does recall them, he fixates instead on the sort of information that assuages these doubts. We do do this sort of thing. Indeed, perhaps most of us do this most of the time. But these are just the sorts of delusion that are maintained by affective biases of attention and memory, and so just the sort of delusion that individuals are rid of to the degree that they are Wide Awake. In effect, we can grant that the deluded businessman's life is not a good one, and say that even someone *inhabiting* such a life would judge it from the inside as not good, to the degree he was more fully conscious and more accurately aware of the stimuli available to his sense receptors. Of course we ourselves would only judge this life not to be a good one from the inside to the degree we were Wide Awake. But I don't think it is necessary to getting on in life that we be less than Wide Awake, in part because the very kinds of alertness and affective biases that normally prevent us from seeing and remembering many unpleasant aspects of life are also responsible for allowing the proliferated negative emotional reactions we have to the painful things we do see. This is where the empirical evidence of decreases in rumination and empathic distress and the like with mindfulness training are helpful, since that would suggest by becoming

more Wide Awake, we not only make ourselves more fully vulnerable to seeing the painful sorts of things in the world that activate empathic concern, we also by the same token make ourselves less prone to empathic distress. More generally, the idea is that it is possible to be live very Wide Awake. This is an empirical hypothesis, subject to empirical disconfirmation. The observation that we do not attribute a good life to the deluded businessman does not threaten Acting Wide Awake, though, since a central claim of the theory is that our ethical evaluations of how to live implicitly assume a perspective that privileges the subjective perspective of one who is fully and accurately aware of which ways of living are characterized by negative and positive affect.

A final family of objections can be raised, by recalling a thought experiment offered by Socrates in *Philebus* (21a). Imagine a life filled with the greatest pleasures, but lacking even the basic cognitive resources to know that one was pleased, to recall past pleasures, and to forecast future pleasures so as to decide on a present course of action. This would be the life of an oyster, not a man. Socrates' interlocutor Protarchus, at least, is left speechless by this consideration, unable to bring himself to judge such a life even eligible as a good one. Moore takes the problem raised by Socrates here to be that "if we are really going to maintain that pleasure alone is good as an end, we must maintain that it is good, whether we are conscious of it or not" (Moore, *Principia Ethica* (1903/1962), 89 in Feldman 2004, 44). Even if Moore is wrong on the exegetical suggestion,[4] the question he raises is an important one. How could unconscious pleasures be good? Feldman is puzzled by the very notion of unconscious pleasures, and so sets Moore's version of the question aside. For AWA, however, this challenge may be one of the deepest.

After all, my proposal is that it is precisely because we are not fully conscious of the affective valence intrinsic to the various Qualities of Heart that we fail to judge their aesthetic and ethical goodness accurately; to the degree individuals pay attention in a less biased way they will

[4]Socrates take's Protarchus' speechlessness as an opportunity to suggest that, for a human being, the life that is best is one integrating the capacity for feeling pleasure together with the capacity for thinking and judging clearly. One can take this as starting point for basing ethics not in considerations about pain and pleasure but instead in (allegedly) objective considerations about what is necessary for the flourishing of the thinking, feeling beings that we are. I argue against one such approach in Chapter 4.

consciously experience more fully the positive and negative affective properties of emotional motivations such as ill-will or benevolence, and will thereby will converge in judging ill-will to be the sort of Quality of Heart one ought not to have, and benevolence to be one we ought to have. For the sake of ease of exposition, I have sometimes spoken in an somewhat imprecise way of our capacity to feel more fully the relative ease and unease of various Qualities of Heart. We can put the proposal more precisely in psychological rather than phenomenological terms. The particular negative and positive affective valence a certain perceptual object has for us, given our particular conditioning, structures many aspects of our behavior, such as whether we purchase one brand or another, how we relate to a new acquaintance, and what sort of ethical judgments we make (Lebrecht et al., 2012). Affective valence affects behavioral response even (or perhaps all the more) when our attention is focused elsewhere, but we can attend so as to make these affective properties conscious; when we do so, we experience negative affective valence as unpleasant; positive affective as pleasant. It is this dispositional property, to be experienced when made conscious as pleasant or unpleasant, that makes certain unconscious valences positive and others negative. What I claim is to be gained in terms of ethical judgment by being Wide Awake is that one will be more fully conscious, and thereby more accurately aware, of the otherwise unconscious affective properties of emotional motivations such as ill-will and benevolence. The emotional motivation of ill-will involves physiological states that, for any human being so motivated, are characterized by a predominance of negative affective valence. We can know this because (if my prediction is correct) to the degree that individuals feel more fully their own emotional motivations of ill-will, they will converge in judging that state to be one that is predominantly unpleasant. AWA as a normative theory turns on the idea that a state such as ill-will is bad, for any of us, even if none of us happen to consciously experience its negative valence. But how could unconscious negative affective valence be bad? Call this the Objection to Unconscious Badness.

A first response to this objection has already in effect been given in the explanation immediately above. We can't encode, for later deliberation or report, information about the affective valence of a perceptual object unless we consciously experience that valence. For this reason, a negative

affective valence that has not been consciously experienced will not be judged by the subject to be bad. The sense in which it is nonetheless bad is that whenever this valence does become conscious, it is so judged. A more complete response can be developed by returning to the arguments in Section 3.1 for an asymmetry between sensitivity and insensitivity. First, to the degree one could succeed in being perfectly insensitive to the emotional pull of one's emotional reactions, as perhaps in an idealized extreme case of psychopathy, one might well thereby forfeit the pleasures of many other human goods. This kind of perfect insensitivity may be a logical possibility and yet not a psychological one, moreover, because we are subject to the vagaries of attentional capture by intense emotions. And if so, a human plan for living that involves developing insensitivity may be less internally coherent than a plan for living that involves being sensitive to which Qualities of Heart feel good to us when we are Wide Awake, and choosing to act out of those.

3.3 Factors and Foundations

It is important to note at that my claim is not that we do, nor that we should, cite in the course of moral deliberation only considerations about ease and unease, or considerations about what emotionally aware people would do. Rather, the idea is that whatever factors are relevant in ethical decision-making, their relevance is justified by such considerations. The distinction between what Shelly Kagan (1998, esp. 17-22, 189-191) has called "normative factors" and their "foundational theories" can help here. When we try to decide what to do in some particularly difficult case, we may appeal to various factors. Take the classic trolley case, in which we are faced with the decision whether to push a fat man off a bridge to his death, thereby stopping a trolley and saving the lives of five people lying on the track ahead. Perhaps the fact that this action would involve intentionally pushing one man off a bridge to his death makes doing so wrong, even when that would save the lives of five others. Perhaps the fact that there is a net benefit to aggregate welfare makes the action right. The type of volition involved in killing, the type of character traits that would be required to do such an action, and the consequences for aggregate welfare are all among the sorts of factors

that people appeal to in justifying their decisions about how one ought to act in such cases. These are normative factors. What makes difficult cases difficult is that it is not obviously clear which normative factors should be given more weight than others. Debates about such difficult cases thus lead quickly questions at the foundational level. What is it that makes a particular factor have more weight in a particular situation than another? In virtue of what could a given normative factor have any claim on our ethical decision-making at all? Foundational theories attempt to answer these questions.

Acting Wide Awake is intended as a foundational theory, an account of what it is that makes certain factors relevant in ethical decision-making. The fact that an act would harm another is an important reason not to do it. So too the fact that a certain sort of action is respectful towards those who are worthy of respect might be an important reason in favor of doing it. Likewise, if an act would support a social institution that is worthy of support, or bring an end to an evil sort of social structure, these things ought to be done, all things being equal. We should be thinking about such factors when we are considering whether to act offensively towards a wise elder, or whether to implement a policy that takes wealth from some to support those in greater need. Although I do not take substantive stands on these first-order normative questions here, it is important to note that AWA is pluralist at this level of normative factors. Nonetheless it is a monistic theory at the level of normative foundations. Specifically, AWA takes the justification for the weight of any such factors in ethical decisions to be decided on the basis of which sorts of Qualities of Heart are good ones to be motivated, and in turn on facts about what which Qualities of Heart we would want to have to the degree we were Wide Awake. I noted above, for instance, that the justification for giving weight to considerations about the welfare of others, according to AWA, turns on the fact that compassion and goodwill are a good Qualities of Heart, and in turn on the claim that we ourselves would want to be motivated by these rather than alternatives such as apathy, to the degree we were Wide Awake.

Claims at the level of foundational theory are closely linked to claims about which normative factors are relevant to ethical decision making. On the one hand, it is often the conviction one has

that certain factors are relevant, or more relevant than other factors, that motivates one to give a particular account of what it is that makes these particular factors more relevant than others. On the other hand, the account one gives at the foundational level of what makes certain factors relevant clearly has implications for which factors do have weight. So we would expect that the answers one gives at the level of foundational theory would have strong implications for which normative factors one takes to be important, and visa-versa. What is interesting about Kagan's analysis is that he shows how far these two can come apart. His discussion of ethical egoism of is of particular relevance to my account.

The idea that an action is right if and only if it promotes the pleasure and reduces the unpleasant experience *of the agent* might seem to be at home only the mouths of mafioso and other evil characters. This is ethical egoism at the factorial level, and it is by my lights implausible as a theory of how one ought to live. Ethical egoism at the level of foundational theory, however, is a very different animal. The idea here is that whatever factors turn out to be relevant to ethical decision-making, the relevance they have will be due to their providing or supporting the welfare of the agent. This view at the foundational level might lead to ethical egoism at the factorial level as well. But it also might not, as Kagan shows. On the one hand, we might take actions as the primary evaluative focal point, and evaluate individual actions on the basis of whether a given one will lead to the agent's welfare. With these assumptions in place, Kagan suggests, it is plausible that the only permissible actions are those that promote the agent's own well-being; in this way ethical egoism at the foundational level may imply ethical egoism at the factorial level.

Nothing about making hedonic welfare the ultimate ground of ethical claims dictates that we take *actions* as the primary evaluative focal point, however.[5] If instead we take *rules* as primary,

[5]Indeed, my account above suggests that taking intentional action as a primary evaluative focus is an unstable theoretical position, where this is taken to be unitary construct including both the volition behind an action as as well the causal consequences of behaving in that way. This is because if my account in Chapter 2 is correct, the evaluation of action involves the interaction of two separable psychological systems, the one responsive to outcomes for which the agent is causally responsible, the other responsive to the agent's intentions. Different ethical systems can make appeals to one or the other of these systems; for instance consequentialists may appeal to intuitions delivered by the outcome-responsive system, while Kant may equally legitimately appeal

and ask ourselves which sorts of rules when followed absolutely would lead to the agent's welfare, Kagan suggests that we may come up with something that looks remarkably like common-sense morality. Because of the social consequences of lying, and the social benefits of giving away resources to others, rules dictating such actions might indeed maximize the welfare of the agent. Because such a view takes rules as the primary focus of evaluation, an act is not justified by the fact that it would maximize the agent's welfare; even if it did so in a particular case, if it required following a rule that, maintained absolutely, would not maximize the agent's welfare, then the action would not be permissible. Alternatively, we could ask which *traits* of character would maximize an agent's welfare. Because of the social consequences just adduced, honesty and generosity might be traits that would indeed maximize an agent's welfare.[6] And if so, again, a rule against lying, an institution that promoted honesty, or an intention not to speak a falsehood might be justified indirectly, on the grounds that it promoted the virtue of honesty. But then again, such a rule, institution, or intention might not be justified if it did not promote this trait, for on this system traits would be the primary evaluative focal point. The important point for our purposes here is that rule- and virtue-based egoist views at the foundational level can accommodate multiple different sorts of factors at the normative level. More generally, monistic theories of value at the

instead to the intention-responsive system when he begins his meta-ethical project with an ode to the overwhelming importance of goodwill, to the exclusion of outcomes. Since the two systems involved in human moral judgement interact at precisely the point where intention meets causal consequence, it makes sense that much of human judgment would be focused on action. But this analysis equally implies that what is required at the foundational level to settle the ethical status of an action is to decide which system, that focused on intentions or that focused on outcomes, is to be prioritized. And that is precisely to take as a primary evaluative focus one of these, the volition for an action *or* the causal consequences of an action, rather than holding intentional action as a unitary construct, and focusing evaluation on that.

[6]Kagan does not suggest, but we might, that the evolutionary processes that were involved in giving rise to common-sense morality may have operated along some of these same lines, selecting dispositions that invariably motivate individuals to follow those rules that would, when followed invariably, maximize the welfare of agents who maintained them. Such forces could have selected not only to intuitive reactions against lying, but also for dispositions towards angry revenge and making sure that aggressors get their just deserts, if having such a disposition would ensure over the long run that one is not taken advantage off in the competition for scare resources. Even if this descriptive account were true, however, that would do nothing to support the normative claim that we ought to adopt ethical egoism at either the foundational or factorial levels.

foundational level can lead to pluralist views at the level of normative factors.[7]

Acting Wide Awake takes a monistic stance at the foundational level, and a pluralistic stance at the level of normative factors: considerations about which character traits, rules, societal institutions, as well as intentions are right or wrong can all figure legitimately in justifying a particular ethical judgment. Nonetheless, a trait of acting generously is a good one, on my account, if and only if it involves Qualities of Heart, such as benevolence, that are less unpleasant than other relevant alternative emotional motivations, such as greed. Likewise, in cases where maintaining a particular sort of rule, e.g. against stealing, would serve to guard against cultivating Qualities of Heart that are bad in this hedonic sense, then maintaining that rule is the right thing to do. For similar reasons, maintaining a societal institution that alleviated individuals' need to take from others, for instance a system of social supports, might be justified if it were to serve in reducing the predominance of the emotional motivations of greed or ill-will that result in stealing. Importantly, however, taking Qualities of Heart as the primary evaluative focus also means that a certain trait, rule, or institution might under most conditions be the right thing, if it cultivated good qualities of Heart, and still under certain conditions be wrong, namely those conditions in which this factor functioned to cultivate hedonically bad Qualities of Heart.

As Kagan points out, however, a monostic foundational theory does not allow room for taking the ethical valence of the primary evaluative focus to be dependent on other conditions. Taking Qualities of Heart as the monostic foundation means that is never the case that we ought to cultivate Qualities of Heart that are bad. On this construal of AWA, if a certain sort of emotional motivation is hedonically bad for human beings, than none of us ought to cultivate that sort of motivation

[7]Of course, one can also be pluralist at the foundational level, insisting that there are multiple considerations, all equally ultimate, in terms of which considerations at the factorial level are to be justified. The challenge here is to articulate the principles that dictate which of the foundational considerations are to be given priority over others in the hard cases where these principles overlap or come into conflict. One can simply deny - at least in theory, if not in practice - that there are answers to be given in difficult cases about why certain factors ought to be prioritized over others, or why certain factors ought to be given any weight at all. At one extreme, one can reject all attempts to give any foundational explanation of why certain factors are to be given more weight than others in our decision-making.

in thought, speech, or bodily behavior, regardless of other positive consequences. If ill-will is a unwholesome sort of emotional motivation in this way, for instance, then it is never the case that being motived by ill-will is the right way to be, even if being so motivated would strengthen institutions or character traits that would normally be considered praiseworthy on AWA. The reason is that these institutions or character traits are considered praiseworthy just in cases where they serve to cultivate those Qualities of Heart that are good.

Alternatively, one could take the foundation to be the general hedonistic properties of external as well as internal stimuli, and then say that as a matter of psychological fact the hedonistic weight of emotional motivations vastly outweighs that of external conditions, in general. If so, there might be isolated cases where it cultivating a bad Quality of Heart would be the right thing to do. But if these cases are indeed not unified by any principle that could be moralized, then it still might be the case the only norms that would survive as moralized norms among such a population of individuals who are Wide Awake would be those that are based on these foundational considerations about Qualities of Heart. Given the diversity of human cultures, even among individuals who are Wide Awake, there might be a diversity of conventional norms about how to eat and how to mate. But although we don't have a choice about which social world to be born into, we do have some degree of choice about which social worlds we continue to live in and to mold ourselves in accordance with and to sustain for those who follow us. Because we can choose in this way, the choices made by others are subject to criticism from us, and our choices are subject to criticism from others. Given this, one interesting proposal for empirical investigation would be that the only grounds on which individuals who are Wide Awake would criticize the choice of which social world to continue in or to adopt would be based ultimately in considerations about which Qualities of Heart are cultivated.

3.4 Normative Implications as Empirical Questions

I showed in the previous section how taking Qualities of Heart as the focus of ethical evaluation can justify the relevance of many different sorts of considerations as factors in ethical decision-making. A final implication of this account is that even if we can give a basic code of how we should be motivated, how precisely we should *act* may in most cases be uncodifiable. To the degree we are each Wide Awake, we will naturally prefer to be motivated by Qualities of Heart that are good, in the sense defined in earlier sections. But what this implies for which sorts of actions we ought to engage in and to avoid is a complex issue, because human situations are so complex. When a human being finds herself in a situation, and is alert and unbiased enough to take in the relevant information and to feel fully the Qualities of Heart she is motivated by, then she will naturally be motivated in the right way and so naturally be motivated to do whatever follows from being motivated by good Qualities of Heart. That said, it may not be possible to codify this sort of ethical expertise in anything except a very minimal and very rough way. Perhaps certain actions could never be motivated by Qualities of Heart that are good ones. Some Buddhist traditions hold that voluntary sexual activity can only occur when motivated by unwholesome Qualities of Heart, for instance. But even among Buddhist traditions there is controversy over this point. On feature of AWA is that points the way towards an empirical means to adjudicate such disputes. Suppose that some Buddhist master claimed to be Wide Awake, or let others think he was, and then also claimed that his sexual activity was motivated wholly by good Qualities of Heart. We can't take him at his word, of course, because self-report can be unreliable, and especially in such circumstances. Fortunately, one feature of AWA is that it provides in principle an empirical means of settling such questions. First, we can measure in an objective way both an individual's level of alertness to internal and external stimuli and also his level of affective bias in attention and memory. This provides a first check on how much we can trust someone's self-report. But even without assessing the individual's introspective reliability, if there were some objective way to measure the affective valence of an individual's emotional motivation at a particular point in time, then in principle we should be able to tell objectively whether the Buddhist master was indeed engaging in a certain sort

of action out of Qualities of Heart that were characterized by ease, or instead that the action was motivated by a Quality of Heart that was characterized by negative affective valence, unconscious or unreported though it might be. This approach can serve not just to intra-Buddhist debates, but much more widely. One might wonder for instance whether there could ever be a case where killing another human being is ethically permissible. Here again, the relevant test would be whether it is possible for one who is Wide Awake to act in that way, and whether they would judge the emotional motivations involved as characterized by ease or not.

It is important to note that for an emotional motivation to be hedonically preferable, in principle, it need not be at all pleasant. It need only be less unpleasant than the other attitudes that are possible to take up. This is especially important because we are often motivated to turn our attention and our thoughts away from the suffering of others by aversive reactions to seeing the pain of how things are in the world. A hedonistic basis for ethics might be taken to be in conflict with an ethical claim that we ought to act out of compassion for others suffering; after all since the empathic reactions reactions triggered by seeing others suffering are painful, perhaps ignorance is bliss. Drawing from Buddhist tradition, my suggestion is that the emotional reactions of aversion that drive us to turn away from suffering are themselves negatively affectively valenced, and will be reported as unpleasant to the degree that one is fully and accurately aware of them. The unpleasantness of emotional reactions of turning away may seem subtle, and may not seem to ordinary subjects to outweigh the unpleasantness of that we feel by not turning away from the suffering in the world. However, I suggest, by increasing emotional awareness individuals will come to feel more acutely the hedonic weight of internal emotional proliferation. If I am correct, in many cases where they were previously motivated to avoid perceiving others' suffering, having developed more full and accurate emotional awareness, individuals will feel it more unbearable to turn away than simply to feel their own negative affective appraisal of others' suffering. If this is right, then to the degree individuals cultivate increased alertness and decreased affective biases in the way I have suggested, they will be *not more insulated* from seeing the painful aspects of human behavior but instead *more vulnerable*. This does not mean that they would thereby be more subject

to the debilitating effects of empathic distress, since according the research reviewed in Chapter 2 the very mechanism I have suggested for feeling more fully one's emotional responses also works to counteract habits of proliferating sorrow and distress. Instead, to make oneself vulnerable in this way is just to put oneself in a situation to be motivated to act to alleviate others' suffering. Because being motivated in this way internally is characterized by much more ease than turning away, being concerned for others in the way that leads forcefully to action is the better way to be. My suggestion, subject to empirical disconfirmation is that ignorance is not bliss; in particular, it is worse than being Wide Awake, on a purely hedonic level.

Aside from the internal suffering caused by emotional reactions to other's suffering, one might wonder about the unpleasantness of the external conditions one gets into by acting out of compassion. In particular, one might wonder whether and how we should be motivated to change unjust situations. Central here is the question of the ethical status of righteous anger. As an example of Qualities of Heart that differ in hedonic and ethical valence I have used throughout the contrast between ill-will and good-will, or between between hatred and friendliness. These are perhaps not very precise categories, and at any rate I mean them only in a provisional and imprecise way, since AWA suggests that it is the judgements of those who are Wide Awake, rather than my own less than fully Wide Awake phenomenological sense, that will give us precise distinctions regarding which sorts of emotional motivation ought to be cultivated and which ought not to be. Nonetheless, I do think that somewhere in the vicinity of anger and hatred there is a certain type of emotional motivation that would be judged by any observer, to the degree that she feels more fully her own Qualities of Heart, as one that ought not to be cultivated. At first glance, at least, any thesis in this general vicinity will be in conflict with certain central Western ethical values. Aristotle, for one, holds that we ought to have an emotional motivation something like anger (ὀργ-) at the right things, towards the right people, in the right way, at the right times, and for the right duration (EN IV.5). Many modern people share an intuition that we ought to be outraged by social injustices. I think there is something right about this, and also something wrong.

Outrage can often motivate people to act forcefully, and at considerable risk to themselves,

to stand up against institutional structures that cause great suffering. One might think here of the protest movements that brought to an end the British Raj in India, the Jim Crow Laws in the United States, or South African apartheid. One thing that might motivate forceful action is a kind of hatred of those who perpetrate injustice, a desire to see them get their just deserts. I suspect that such hatred or ill-will involves physiological changes that are deeply unpleasant to the person in these emotional states. If this is so, then to the degree judgements of justice or injustice are based in emotional reactions of anger, these judgments will not be made by those who are Wide Awake, just in virtue of the fact that they will not want to cultivate in themselves nor want others to cultivate reactions of anger. This need not mean giving up on holding individuals responsible for their actions. Drawing on experience in clinical treatment of disorders involving drug abuse, self-harm, and anorexia, Hanna Pickard (2013) argues for a model of responsibility without blame. She holds that we can and must, for treatment to be effective, hold clinical patients responsible for their actions, but that we can do so with "detached" rather than "affective" blame. I suggest a way to improve on this way of categorizing things below. Nonetheless, Pickard's points about the clinical context may hold much more generally. She points out that "compassion and empathy push the negative emotions constitutive of affective blame aside." Moreover, Pickard's account affective blame of involves a "feeling of entitlement – of being in the right, in relation to another's wrong". I would likewise suggest that blame and anger more generally involve affective biases of focusing on the faults of the other, and correlatively ignoring one's own faults. In extending Pickard's model to the justice system, Lacey and Pickard (2012) note a specific issue in this regard, our collective biases against noticing the societal responsibility we bear for the developmental conditions that lead to anti-social behavior. If affective biases cause fixation on the faults of another to the neglect of our own faults, decreasing such affective biases in itself may reduce allow us to feel more empathy with the other, in part by seeing how we too are disposed to act out of bad Qualities of Heart. But such a stance does not entail giving up on holding someone (ourselves or others) who has transgressed legal or ethical principles responsible for their actions. On the contrary, a motivation of compassion may dictate that we do provide negative as well as positive incentives to

engage in better behavior. One implication of AWA is that to the degree perpetrators of violence or injustice are motivated by greed or hatred they are causing themselves harm internally, even before they begin to harm others outwardly. The Buddhist suggestion is that we can, and we ought to, eliminate these kinds of destructive tendencies in ourselves. Critically examining this proposal, Owen Flanagan (2000) notes Strawson's (1962) suggestion that to eliminate reactive attitudes such as gratitude and resentment is to eliminate the expression of our own humanity and the acknowledgement of others, serves as a classic example. Flanagan points more generally to the Aristotelian supposition evident in Western moral philosophy and perhaps more generally in Western thought, that eliminating the reactive attitudes may not be possible and moreover that even if it were it is not desirable. "But perhaps that picture is just that - ours," Flanagan (2000, 279) puts it well, "and cherished primarily for its powerful roots in our history and not because it is rooted in deep wisdom, in the best picture of normative self-construction. Maybe."

My argument in this chapter and in this dissertation more generally is designed precisely to question this Western supposition. In comparing the merits of the Buddhist proposal, it is crucial to note that the suggestion is not that we should eliminate all reactive attitudes, even if this were possible. Rather the suggestion is that some are (much) better than others. In particular, the idea is that attitudes rooted in delusion, attachment, and hatred are worthy of elimination, and that to do so allows other attitudes to flourish in us and to motivate our actions, attitudes such as love in the brotherly sense MLK means as a contrast to hatred. This proposal suggests a way to construe Pickard's insight in less cognitive way than she does. By drawing on these traditions of nonviolent action, one might suggest the possibility of forceful action motivated not by hatred or anger, but instead by a basic kind of friendliness and compassion for the suffering that perpetrators causes themselves, as well as the suffering they cause their victims. Indeed, being motivated by such emotions is what seems most powerful, and wise, about figures such as Martin Luther King Jr., as opposed to, say, Malcom X. AWA provides a neat rationale for this intuitive sense. The point is not to move from affective blame to a more cognitive blame, but rather to switch from one type of affect to another. In responding to actions motivated by hatred, for example,

the suggestion is to respond out of care so as to precent the perpetrator from harming himself as well as his victims. If instead we turn away or else hate the perpetrator for hating will, according to King and to the Buddhists, in either case we will by causing ourselves suffering even as we allow others' suffering also to proliferate. This possibility too, can be taken as an empirically testable hypothesis. As above, if there were some objective way to measure the affective valence of individuals' emotional motivations as they act forcefully to stop injustice and hold perpetrators responsible, then in principle we should be able to tell objectively whether there are any individuals who can act forcefully in these ways out of Qualities of Heart that are characterized by ease, or whether instead such actions are always motivated by a Quality of Heart that is characterized by negative affective valence, unconscious or unreported though it might be.

3.5 Conclusion

I think that many human beings can feel the intuitive force of taking Qualities of Heart such as anger and blame to be ones that we should not cultivate; even if asked they would not agree, sometimes other behavioral cues can be indicative. Let me close with one example. The Zen teacher Joan Halifax tells of leading a silent sit-in, a kind of bearing-witness at the execution of a New Mexico man convicted of raping and strangling a seven-year old girl. Alongside these silently mediating Buddhists at the gates of the correctional facility was a group of friends and relatives of the young girl who had been murdered. When an officer came out and announced that the man had been executed, at first the group of relatives cheered, expressing their anger at this man for what he had done. But with the Buddhist group staying silent and present for the pain of the whole situation, those cheering the execution quickly became quiet. There could be many explanations for this. But one plausible explanation is that the possibility of a different sort response being made perceptually salient, those cheering the execution felt the relative unease involved in cultivating and acting out of anger; they felt for themselves that there was something off about their own actions. Perhaps they even felt the possibility of a compassionate, if forceful, response to both the suffering

of the perpetrator and that of his victim.

In the immediately preceding section I have focused on how ethical questions, including as the appropriateness of such righteous anger, might be operationalized as empirical hypotheses. I based that analysis on the claim of Section 3.1, that among Qualities of Heart such as friendliness, greed, compassion, and hatred, we can distinguish those that are ethically better from those that are ethically worse on the basis of the relative ease or unease they involve for a human being so motivated. Taken one way, the plausibility of AWA as a normative theory turns on the empirical tests I have proposed coming out a particular way. Suppose that, my argument in Section 2.2.1 notwithstanding, it turns out empirically that hatred is characterized by more ease than friendliness, such that to the degree human beings are Wide Awake, they will prefer to be motivated by hatred rather than by friendliness. If so, according to AWA, an emotional motivation of hatred, and the actions that result from it would be more ethically praiseworthy than an emotional motivation of goodwill or friendliness, and the actions that result from that. To the degree common sense is committed to hatred being ethically worse than goodwill, this would be embarrassing for AWA as an ethical theory. If this is how things turn out empirically, then perhaps we should reject the theory rather than our common sense. But this argument can also be turned around in AWA's favor. It is no fault of a theory that its conclusions mirror folk intuitions to some degree. After all, maybe we do get some things roughly right. In particular, if our ethical judgments include an implicit assumption that ethically good Qualities of Heart such as goodwill are good for the person so motivated in the way I have suggested, maybe they are right about that. This is an empirical question, and I take it as a strength of AWA that the proposal stands or falls on such empirical grounds.

Chapter 4

How Not to Ground Ethics, from the Human Point of View

In Chapters 2 and 3, respectively, I outlined the empirical and philosophical considerations in favor of the normative account Acting Wide Awake. This approach has three central aspects. First, ethical claims are grounded subjectively; they are true or false in virtue of facts about the person making the judgment in relation to facts about the actions in the interpersonally accessible outside world, rather than in virtue of facts about about the world independent of the subject making the judgment. Secondly, the aspects of subjectivity in which ethical claims are grounded are affective and more generally experiential ones, facts about the rich and vivid conscious feelings and perceptions of an individual human being, rather than facts about the rational principles by which a human being is constrained in virtue of being a rational being. Thirdly, ethical claims are grounded not in the idiosyncratic ways in which one individual one group of individuals experience the world, but rather in aspects of human experience that are universal among human beings in virtue of the biological and psychological make-up that we share and that shape our experience in common ways. In this sense, we can talk of universal features of *the* human point of view, whereby we mean those experiential features that are common to each individual human being's world of experience.

It would be hubris to suppose of any current theory of the mind or morality that it gets all the details right. The best we can hope for a theory is that it does less bad than competing theories. In the present chapter, I argue that AWA does just that. In particular, I survey four prominent approaches to ethical theorizing in Western philosophy, noting points of convergence with and divergence from the AWA account. I bring out the relevant distinctions between these approaches by focusing on the three very general axes noted above, noting for each approach whether the meta-ethical ground of ethical claims is subjective, whether it is affective, and whether it is universal (see Table 4.1). The approach I have called subjectivist might be contrasted with the type of approaches to ethics often called realist. In the absence of further specification, however, this latter usage is misleading. On a Humean theory such as Prinz's (2007) view, for instance, the rightness or wrongness of an act is determined by the emotional dispositions of the person (or moral community) making the moral judgment. Since emotional dispositions are real, empirically tractable phenomena, Prinz's theory can be characterized as realist about response-*dependent* moral facts. Foot's Aristotelian approach, in contrast, is a version of realism about response-*independent* moral facts.

Table 4.1: LOCATING ETHICAL VALUE

The ground of ethical claims is... *(for caveats see text marked by symbols below)*					
	AWA	Kantian	Humean	Utilitarian	Aristotelian[§]
Subjective?	**Yes**	Yes*	Yes	No[‡]	No
Affective?	**Yes**	No	Yes	Yes	No
Universal?	**Yes**	Yes	No[†]	Yes	Yes

Like AWA, Kant takes a human being to be subject to laws "given by himself". The ground of ethical value is subjective in this sense (*); if an action is wrong, it is wrong at least in part in virtue of facts about the person making the judgment. For Kant, the relevant fact about the person making the judgement is that she is a rational being, subject to the demands of reason; it is in virtue of being a rational being that she must necessarily judge certain sorts of actions as obligatory and others as impermissible. While the Humean approach agrees with the Kantian one that the meta-ethical ground of ethical claims is subjective, for the Humean the facts about the

person making the judgement that are relevant to making an ethical judgment true or false have to do with his passions rather his reason; the meta-ethical ground of ethical claims is affective. The meta-ethical ground of ethical claims is also affective on Utilitarian approaches; the rough idea is that pleasure is the good, and that one ought to maximize what is good. However, whereas for the Humean the truth-makers for ethical claims are the emotional attitudes and responses of the person *making* the judgment, for the welfare consequentialist what determines whether an action is right or wrong are the emotional responses of those who are *affected* by the consequences of the action. The occurrence of the pains and pleasures that ground ethical claims are to be assessed from an impersonal standpoint, as facts about pain and pleasure in one person on a par with the pains and pleasures of any other person (‡). Thus although for modern Utilitarians the meta-ethical ground of ethical claims is affective, it is not subjective (though, as we will see, Mill himself seems to have proposed a subjectivist ground for his Utilitarian normative conclusions). Foot's neo-Aristotelian approach takes this rejection of subjectivism a step further, holding that emotional reactions have no special role in grounding ethical claims. McDowell (1998, 167ff) suggests that Foot's variety of naturalism is not Aristotle's, but that exegetical issue will not worry me here; for my purposes the point is just to examine the merits of Foot's variety (§). Her approach rejects the claim that the meta-ethical ground of ethical claims is subjective because it also rejects the claim that this ground is affective. This difference notwithstanding, for both the Utilitarian and the neo-Aristotelian, the rejection of subjectivism serves as the basis for defending the idea that the meta-ethical grounds of (at least some) ethical claims are universal to all human beings. Despite his subjectivism, Kant also shares this universalist position; the laws that reason demands each rational being give to himself are the same for any rational being. Thus, all three reject the arguments of modern Humean theorists (who may, however, differ on this point from Hume himself), which move from the claim that the ground of ethical judgments is affective and subjective to the conclusion that such grounds are culturally relative, and not universal (†).

It is important to note that these views are competitors to AWA only in so far as they are taken as theories about the ultimate ground of ethical claims. Drawing on Kagan's distinction

between normative factors and foundational theories, I pointed out in sections 3.3 and 3.4 that AWA can accommodate as important in ethical decision-making considerations such as aggregate welfare, what is required for optimal human functioning, and the constraints put on our actions by the need to respect others. However, according to AWA, the justification for giving such factors normative weight turns ultimately on considerations about which Qualities of Heart any human being would choose act out of, to the degree she is Wide Awake. AWA thus offers one example of an approach that takes the ground of ethical judgments to be subjective, affective, and also universal. This position is distinctive; each of the leading competitors rejects one or another of these three claims. My aim here is show that even the ethical theories that have been taken to have the widest plausibility by Western philosophers have been plagued by problems stemming from not holding the ground of ethical claims to be subjective, affective, and also universal. If AWA does manage to coherently integrate all three of these aspects, it would thereby avoid the central obstacles the leading contemporary theories.

4.1 Foot's Aristotelian Account of Natural Goodness

In terms of the features charted in Table 4.1, Aristotelians adopt an approach to grounding ethical claims that is the converse of the approach adopted by modern Humeans. Where the recent sentimentalist theories discussed in Section 4.4 hold that the ground of ethical claims is subjective and affective but not universal, Philippa Foot holds that the facts that make ethical claims true or false *are* universal, and precisely because they are neither subjective nor affective. Indeed, in *Natural Goodness* (2003), Foot takes aim squarely and explicitly at at Humean approaches to grounding ethical claims, in a way that would apply not only to her own earlier views and the likes of Prinz and Blackburn, but also to my own subjectivist universalism. Nonetheless, like that of the modern sentimentalists, Foot's neo-Aristotelian approach is deeply empirical. On the view she develops, it is in relation to empirical facts about the human species that the ways of acting and being that are adopted by individual human beings can be judged as excellent or as bad.

Foot's strategy is to appeal to the continuity between, on the one hand, evaluations to the effect that a human being is not acting as she should or not being the sort of person she should be, and on the other hand evaluations of defect and excellent in plants and animals, as when we say that a tree whose roots are too shallow and weak to obtain nutriment is not as it should be, that the individual tree is in that regard defective. The evaluation of excellence and defect in animals and plants depends not (just) on statistical norms, but rather is relative to the way a particular life-form gets along. Foot offers the example of a peacock's tail; "it *matters* in the reproductive life of the peacock that the tail should be brightly colored" (2003: 33), and a peacock that lacked such a tail would be not be as it should be. Foot (35-6) analyzes four components of such "natural norms". She begins with (a) the life cycle, focusing in the case of plants and animals on the processes of self-maintenance and reproduction. Then there is the set of empirical propositions about (b) how these processes are achieved in a particular life-form: how nourishment is obtained, how it is employed in development, defense, and reproduction. From this set of empirical propositions is derived (c) norms, requiring for instance "swiftness in the deer, night vision in the owl, and cooperative hunting in the wolf." Finally, it is relative these sorts of norms that we judge an individual as excellent or defective.

The central suggestion of Foot's account is that the evaluation of human excellence and vice shares a similar logical form. Bees, like wolves and human beings, live cooperatively, such that the processes of self-maintenance and reproduction in these life forms depend on cooperation. In this way, the case of a honey-bee dancing to communicate the location of a source of food is relevant to the evaluation of human cooperativeness. "No doubt an individual bee that does not dance does not itself suffer from its delinquency, but *ipso facto* because it does not dance, there is something wrong with it, because of the part that dancing plays in the life of this species of bee" (35). Expanding the point, we could say that even if there was some cost for the individual bee in not dancing, on Foot's account that is not why not dancing counts as a defect in a bee. Analogously, even if to flaunt the norms of human cooperation and communication does have some cost to the individual who does so, still that is not why failing to observe these norms counts as a defect in

a human being. Moral defects are not (just) defects of egoistic practical rationality, a point which Foot explicitly intends as a counter to a Humean theory of morality. Rather, Foot's suggestion is that virtues of trustworthiness, for instance, play a necessary part in the life of human beings in the way that dancing does in the life of bees.

We say, as Foot notes (15), that "it is necessary for plants to have water, for birds to build nests, for wolves to hunt in packs, and for lionesses to teach their cubs to kill." These are necessary in the sense that some good depends on individuals belonging to these life-forms functioning in these ways; not to do so is in that sense bad. The Humean will likely counter at this point that to say something matters, that some good hangs on it, just means that it matters to someone, that it is good for individual or group of individuals. But the example of normative evaluation of plants makes Foot's case nicely, for here there can be no serious appeal to what the plant desires, or is trying to achieve. And just as to suggest that what makes it a bad thing for lionesses not to teach their cubs to kill is the attitudes, beliefs, or desires of the person making the judgment, it is unconvincing to suggest that what makes the shallow roots defective ones is the conative states of the person making the judgment. What makes bad roots bad, on Foot's view, is that those sorts of roots don't have the qualities that are required for the flourishing of that form of life, regardless of whether there was anyone around to judge it so. More generally, excellent and defect in a particular tree is relative to norms derived from facts about how nourishment and reproduction are carried out for that life form.

For Foot, the way to go about finding out what excellence and defect are for a human being, just as for sub-rational beings, is to obtain data about the life cycle of the life form in question, and what qualities that particular form of life requires. The human form of life requires qualities that are not required of non-human animals or plants, and so makes possible defects as well as virtues that are not possible for other life forms. On finding that an adult human is unable to use language, we ask what went wrong; perhaps there was a genetic condition or some physiological damage or neural lesion. On finding that an adult peacock cannot use language, however, we do not ask what went wrong; not to be able to use language is not a defect in a peacock because using

language plays no role in that form of life. Likewise, though there may well be many forms of life in which to be anti-social is not a defect, because we are social animals, like bees and wolves, to be anti-social is a defect in a human being. Moreover, we can use language to facilitate cooperation, as when we make promises. Because we must depend on the keeping of promises in all but the most direct forms of exchanges of goods or services, much human good depends on being the sort of person who keeps promises, in being trustworthy. In relation to such a form of life, an individual who does not keep promises is defective. In other words, being trustworthy is one of the virtues. Others Foot notes (45) include "loyalty, fairness, kindness, and in certain circumstances, obedience."

Despite the appeal of such an approach, it is important to bear in mind that we don't in general seem to take whatever is natural to be good; even if dispositions to rape turned out to be natural in the sense of having been selectively reproduced in the course of evolution, this wouldn't make rape right. Foot's strategy is to start with on cases such as the evaluation of the roots of trees, where we do unproblematically make evaluative judgments on the basis of whether a particular feature is required to sustain a particular form of life. But does appealing to the corresponding considerations about the human form of life tell us whether one ought to live a family life or a monastic one? Can it adjudicate disputes over whether abortion is permissible or not? To sustain the form of life that is human, human beings must reproduce. And if so perhaps being a monastic and having an abortion are equally offenses against this natural order, expressing defects in an individual of a species whose form of life includes being fruitful and multiplying. But many of us just don't see the fact that the human life form depends on reproduction as bearing on the question of abortion or celibacy in any foundational way. A Humean, in particular, wants to counter that unless one antecedently felt good about doing whatever is natural, or doing whatever is required to sustain a particular form of life, claims that a particular characteristic does so would have no force at all.

One way to adjudicate ethical disputes over foundational values, in line with the thrust of Foot's suggestion, would be to turn to objectively measurable qualities of human functioning. There are

many objectively measurable qualities of human functioning available, from memory, attention, and cognitive intelligence tests, to cholesterol and stress, from cooperative behavior in economic games to total number of offspring. For example, Foot notes promise-keeping as a virtue relative to the human form of life. It may well be a practical impossibility for there to be a group of human beings for whom promise-breaking is an excellence, for then the institution itself would break down, and with it all sorts of other human goods. Some of these would no doubt be measurable in terms of decreases in economic cooperation, nutritional intake, and thus quite possibly also reproduction and the sustain of the human form of life. The analogous argument is perhaps harder to make for the case of rape. Since on the species level rape might conceivably turn out to be an evolutionarily sustainable strategy, some measurable qualities of optimal human functioning might be positively correlated with forcible impregnation. Nonetheless, it could also be that for psychological reasons, having the sorts of emotional dispositions that would allow one to rape also means that one is not functioning with optimal cognitive excellence. And so such considerations might conceivably suggest that being untrustworthy or being a rapist are defects relative to a human form of life. Assuming that one's emotional dispositions do have effects on cognitive functioning, it might even be the case that such objective measures of human functioning could bear on the question of whether we ought to have emotional dispositions to hatred or instead to brotherly love.

Still, the problem with such an approach to adjudicating ethical disputes is precisely that there so many sorts of objectively measurable qualities of human functioning available. And a first-order ethical dispute about whether one ought to get angry about injustice or not will just replicate itself in a debate over which aspects of human functioning should be given weight in measuring optimal functioning. Certainly, we can pick some objective measures rather than others on the basis of our intuitions about what a good human life is. But this basis is subjective. At an extreme, nothing about this approach prevents one from simply picking the qualities that define the ethical ideal one already holds, and taking a measure of those qualities alone as the measure of optimal functioning. Even if it were the case that a great number of measures of human functioning were to correlate with one another, in a culture that holds that one ought to be celibate or that one ought to feel

honor-bound to stone rape victims, if some of these measures were to be anti-correlated with the way of life one regards as optimal, one might simply reject those measures as not being measures of optimal functioning. Intuitions about what sort of life is ethically optimal diverge between cultures, sub-cultures, and individuals, for instance about the value of a celibate life versus one filled with offspring. A theory that proposes to give us the means to answer questions about how one ought to live has to adjudicate those disputes. The Aristotelian adoption of an objective point of view on human functioning, ironically, leads to a failure to adequately respond to the specter of moral relativism.

An Aristotelian might respond to the problem of moral relativism in a different way, by distinguishing different levels of description and evaluation.[1] At more fine-grained levels, there is the appearance of cultural or even individual relativity: in some cultures, it is considered polite to burp after a meal, whereas in others it's considered quite rude. In some cultures it is considered a question of honor whether you stone your daughter after she's raped; in others, even considering this possibility is outrageous. At a higher level of description and evaluation, though, the appearance of relativity might evaporate, the thought goes: regardless of whether you should or shouldn't burp, both cultures value politeness. Regardless of whether you should or shouldn't stone your daughter, both cultures value integrity. This higher level of description tends to be couched in virtue terminology, which is one of the attractive things about such language; cultural divergences could then be construed as differing understandings of what virtue requires.

This levels-of-description move is interesting, but it doesn't give us any answers to the question of how one ought to live. Because of this, it also fails to answer the challenge of relativism, the suggestion that there are no perfectly general answers to the question of how a human being should live. If we all agree that one ought to be the sort of person who has integrity, but haven't answered whether that amounts to stoning one's daughter for being raped or not, then we really haven't answered the question of how one ought to live. One might suggest that even so, the levels of description move serves to defang descriptive relativism somewhat by showing that it has to do

[1] Many thanks to Mark Alfano for this suggestion.

with substantive disagreements about what virtue requires, not metaethical disagreements about which traits are virtues. But this seems to me just a semantic dispute about how to characterize what is in either case an insurmountable problem for the 'levels-of-description' strategy. It fails to defang the spectre of relativism. For instead of saying that the disposition to maintain one's honor is a virtue, we could equally say that the disposition to stone one's daughter for being raped is a good one, or alternatively that it is deeply evil. A theory that doesn't tell us which it is to my mind fails to meet the desirata of an ethical theory, and whether one characterizes this fatal problem as ethical or meta-ethical is beside the point.

In looking for a "species-wide notion of human good" Foot herself (92) turns to cases of evil, and those who face it. She asks, taking the case of the serial killers Frederick and Rosemary West, whether someone who had helped to facilitate their crimes of abuse and murder would have thereby benefited them, since to benefit a person seems to be to do something that is good for him. In our refusal to say that to aid someone in such crimes would have been to do something good for the perpetrator, Foot rightly finds a conceptual connection between virtue and a certain conception of human flourishing, or, in one particular sense, human happiness. Conversely, in the letters of Nazi resisters about to be executed, Foot is struck (95) by the "extraordinary sense of happiness they radiate." And she is right to say in a footnote that although certainly they sacrificed the happiness of their lives with their families by not going along with the Nazis, still to see the alternative acceptance of Nazism as happiness "they would have had to have changed, and they would not accept the prospect of such a change" (96). This last paranthetical remark does better, I think, than the main line of Foot's argument, in giving us a means to answer the sorts of ethical questions that trouble us. The problem is not that the good and the right are unconnected to considerations about what flourishing is for the form of life that is human, but rather that such considerations, even if crucially important, still vastly underdetermine the sorts of issues we need an ethical theory to adjudicate. Foot's footnote in effect praises as a virtue the Nazi-resisters' refusal to accept the prospect of changing their feelings about what is evil. And in this I think Foot is not only correct, but has in spite of herself hinted at something that might sufficiently ground judgments about what

a good will amounts to, that is, the sorts of attitudes the person making the judgment has about the sorts of attitudes they or others could have. Indeed, this is just the sort of higher-attitude subjectivist account that Humeans such as Blackburn, and I in a more qualified way, want to endorse.

Foot rightly suggests that the crux of the difference between her way of grounding ethics and a Humean one turns on questions about motivation. She notes that when we want to understand why someone does what he does, instrumental explanations cannot go on forever, if he does A for the sake of B, and B for the sake of C, the regress must end somewhere. She describes the case of someone who throws away his pack of cigarettes because he wants to quit smoking. He does this because he wants to live longer, but somewhere such a series of practical reasons must come to an end. Foot rejects the Humean thesis that such practical reasoning must ultimately be grounded in some conative element, something the subject wants. Instead, she asks us to consider the question, "why should we not take the recognition of a reason for acting as bringing the series to a close?" In this case, the smoker recognizes living longer as a reason for action. Foot suggests that we need not always to explain this in terms of a desire to live longer. After all, she notes, "no special explanation is needed of why men take reasonable care of their future; an explanation is needed when they do not" (22-23). The point is not that it is impossible to give an explanation for a man or woman's actions in more basic psychological or neurophysiological terms. Rather, the idea seems to be that for the purpose of justifying an ethical claim, we don't need to. After all, if a smoker throws away his cigarettes and we ask him why, Foot is correct that we may well be satisfied if he tells us that his reason for quitting is that he understands he is likely to live longer as a result. Indeed, it might be somewhat beside the point, perhaps even offensive, to go on pressing the issue and asking of him then, "but do you want to live longer?".

This is fine so far as it goes. Nonetheless, as Foot notes, an explanation *is* needed when things go wrong. In cases of *akrasia*, for instance, one sees a certain end as good, but out of weakness of the will, fails to act so as to bring about that end. A different sort of defect is present when one in fact has a reason for action, but fails to see it as a reason for action. For instance if health is one of the qualities necessary for sustaining the human life cycle, when someone doesn't take care

of themselves, this is an instance of a defective will according to Foot. In such case not seeing a reason for action can count as a vice. As Foot rightly puts it, giving the example of an arms dealer who chooses not ask where his wares will be resold, "ignorance may itself be voluntary" (70).

Importantly, when a person fails in this way to see as a reason for acting something they ought to, asking that person why she fails to see it as such won't give us the sort of explanation we are looking for. Rather, in such cases we have to descend below the interpersonal level of exchanging reasons, to look for factors that if all was going well would be motivational, but due whatever psychological conditions are not impacting the person's motivational states in the right way. And if when things go wrong, we cannot stop the explanation at the interpersonal level of exchanging reasons, this suggests that even when things go right, there may be something important about the explanation to be had at the psychological and neurobiological levels. Take the case of an ethical dispute over whether it is better to live the monastic life or the family one. In the course of such a dispute, to say that one chose the family life because he recognized having children as a reason not to be a monastic is to take a stand on one side of the debate, it is not to cite some independent means of settling the issue. In ethical disputes between individuals or between moral communities, precisely what is at issue is whether a certain way of being motivated counts as a successful recognition of a reason for action or instead as a failure to see other (more important) reasons for action. So if we need psychological explanations for what goes wrong in cases where one fails to see as reasons factors that one ought to see as such, then in order to settle the debate about which factors one would see as reasons for action if all went right, we need to descend below the level of interpersonal exchange of reasons for action. We can't stop at the reasons a person is apt to cite, because one side or the other (or both) may be failing to seeing as reasons some features that are reasons. Given that there is dispute, the obvious approach is to withhold credence from the claims of both sides, and investigate too see what if anything has gone wrong on either side, just as we would look for an explanation of what has gone wrong in cases where we are sure that someone has failed to see as reasons factors that he ought to see as such. The account I offer adopts precisely such an approach, proposing that when things go right psychologically, one feels fully one's own

motivational states, and that when one does so, certain factors and not others are seen as reasons for action. If the Buddhists are correct, for instance, to the degree one is wide awake in this way, one will see the ability to devote one's life to helping others as a more important reason for action than the desire to have children or to perpetuate the human form of life. This approach allows us to explain why there is such divergence across cultures in the moral values that are held, and how it is possible nonetheless to find in universal aspects of human emotion a certain circumscribed set of answers to the question of how a human being ought to live.

4.2 Mill, Greene, and the Ironic Evolution of Utilitarianism

The Utilitarian approach differs from Foot's objective account of natural goodness in that Utilitarians ground ethical claims in facts about pain and pleasure. This approach, as it is standardly understood, moves from the conceptual claim that pleasure is the good to the normative claim that we ought to maximize the good. Interestingly, J. S. Mill seems not himself have taken the foundation of his welfare consequentialist ethical system to rest on this sort of conceptual claim. Mill begins instead from the very Humean contention that the ultimate ground of moral justification "is but a subjective feeling in our own minds" (Mill, 1863, 34).[2] He focuses in particular on sympathy for others. According to Mill, given the "comparatively early state of human advancement" - as of 1861, at least - people cannot feel "that entireness of sympathy with all others that would make any real discordance in the general direction of their conduct in life impossible" (1863, 34). My sympathies are more Buddhist: I think that whether this sort of perfection is possible directly depends only on the state of an individual's mental development and that the historical development of social institutions is only indirectly related and in any case not so teleologically progressive. I do agree that for those who have some development of this sympathy, "it possesses all the characters of a natural feeling. It does not present itself to their minds as a superstition of education or a law despotically imposed by the power of society, but as an attribute which it would not be well for

[2]Jacobson (2008) agrees, if for slightly different reasons, that Mill may in fact have been much less consequentialist, and much more sentimentalist, than has been appreciated.

them to be without," and "few but those whose mind is a moral blank could bear to lay out their course of life on the plan of paying no regard to others except so far as their own private interest compels" (1863, 34). Of course, no superstition presents itself to the mind as such. Moreover, Mill takes the considerations he offers about what feels natural as sanction for the greatest happiness morality. What I want to note about his approach, however, is that the question of which sentiments we ought to have seems ultimately to rest on what sentiments certain people - specifically, those whose minds are a not moral blank and have some development of sympathy - take as not well for them to be without, as feeling natural, as unbearable to act against. That is, the ultimate sanction for any principle of morality rests on the way someone with the proper emotional sensitivity is disposed to feel about the way they themselves are disposed to feel. So stated, this is just the sort of the claim that I want to endorse.

In contrast, the standard reading of Bentham, Mill, and Sidgwick takes these classic utilitarian theories to locate value in the consequences of an action. In particular, these theories are taken to accept hedonistic act consequentialism, the view that what makes an act good or bad is its actual, as opposed to expected, net consequences for aggregate pain and pleasure (Sinnott-Armstrong, 2009). Whether a particular form of hedonistic consequentialism takes acts, rules, character dispositions, or another type of object as the target of normative evaluation, in any of these cases the ground of ethical claims is human affective responses. Despite this affective basis, consequentialist accounts resist the subjectivism of Humean theories. For the consequentialist, the truth or falsity of an ethical claim is determined not by the emotional attitudes and responses of the person *making* the judgment, but rather by the emotional responses of those who are *affected* by the consequences of the action.

Whereas for the Humean the ground of ethical claims are empirically testable facts about how affective responses influence ethical judgments, the Utilitarian system is usually taken to turn ultimately on the conceptual claim that pleasure is the good. In taking conceptual rather than empirical claims as the basis for ethical theorizing, however, Utilitarianism opens itself to the same sorts of critique of rationalist approaches that I level at the Kantian approach in Section 4.1. This is one

price that Utilitarianism pays in rejecting subjectivism as a means to securing universality.

A second price that standard Utilitarianism pays for basing its approach on the intuitive appeal of the conceptual claim that pleasure is the good comes out in the observation that opposing considerations also have considerable intuitive appeal. It is standardly taken to be an implication of that conceptual claim, in conjunction with simple math, that we ought always to maximize aggregate welfare. If so, it has pointed out, this would seems to have the consequence that we ought to do things like torture babies or smother them if that is what is required to save a large number of lives. And many of us have at least as powerful intuitions to the effect that such acts are wrong as we do that pleasure is the good. What has given the thought experiments developed by Phillipa Foot and Judith Jarvis Thompson and now famous as the 'trolley problems' such life in both philosophical and empirical contexts is that they serve to bring out this clash of intuitions. Where the Utilitarian reasons that we ought to push the fat man off the bridge in order to stop the trolley and save five others, Kant's deontological framework provides a rationalist argument against doing so. If all we have on both sides of the debate are appeals to intuitions that justify taking one rationalist route over the other, it is difficult to see how this stalemate can be settled on philosophical grounds in any decisive way.

In order to break out of this philosophical stalemate, Joshua Greene moves away from basing the defense of utilitarian conclusions on logical grounds, premising his account instead on a psychological basis. He has pioneered the application of social psychology and neuroscience to the classic trolley dilemmas as well as vignettes such as that of the 'crying baby', in which subjects judge the appropriateness of smothering one's baby in order to prevent soldiers from discovering and killing oneself, one's baby, and others. Greene takes his neurophysiological data to suggests that it is in fact intuitive emotional reactions that drive deontological judgments against killing that is "personal", while subjects making utilitarian judgments use cognitive resources to override these initial emotional reactions (Greene et al., 2001, 2004). Aligning results from his own neuroimaging work with cross-cultural studies by Haidt et al. (1993), Greene (2008) further suggests that those with the cognitive resources bestowed by education, westernization, or adulthood are

more inclined to consequentialist moral judgments. As noted above, Greene takes these empirical results as evidence that deontological theorizing in fact expresses post-hoc rationalizations of common sense intuitive reactions determined by evolved emotional reactions rather than by rational principles. This is the allegedly "secret joke of Kant's soul".

Greene thus employs his empirical data to normative ends, claiming that it helps show Utilitarian judgments in these dilemmas to be superior to the classic deontological ones. One theme of these arguments centers on the contention that the intuitions driving our common sense judgments are evolved. The idea here seems to be that exposing certain intuitions as based in evolved emotional reactions reduces their evidentiary value for adjudicating ethical disputes, for instance over trolley problems. Drawing on Greene's empirical results, Singer (2005, 351) likewise proposes "the ambitious task of separating those moral judgments that we owe to our evolutionary and cultural history, from those that have a rational basis." As he admits, "even to specify in what sense a moral judgment can have a rational basis is not easy." Nonetheless, to do so is the only way, as he sees it, to avoid the skeptical conclusion that there are no answers to the question of how a human being ought to live.

Greene may be more modest about this goal. "Utilitarianism," he claims, "follows not logically from the truth about morality, but psychologically. It's what you're likely to want once you know the truth" (Greene, 2002, 318). Indeed, in his philosophical work, Greene adopts an error-theoretic approach to ethics. On his view, clinging to the view that certain acts really are right or wrong, intrinsically, is bad metaphysics and worse social policy. Nonetheless, to maintain utilitarianism as a practical guideline for what we ought to do one need not suppose that it is a good guideline to follow in unrealistic cases and distant possible worlds. What we want in morality is not a set of eternal truths, after all, but a practical sort of guideline for what to do in the sorts of situations we find ourselves. And a utilitarian guideline is the one we will want, Greene suggests, once we see our ethical intuitions as contingently evolved and situationally variable, rather than as delivering eternal moral truths.

In this regard, Greene (2002: 328) cites Singer's example of a severely debilitated infant whose

life would be a painful, drawn-out struggle culminating in early death, burdensome to its parents and a net loss to society as a whole. Understanding why natural selection would be inclined to implant in us a strong and perfectly general sentiment against killing infants, we will be able to discount that sentiment when we reflect on the particularities of a case. To this, Greene adds his rejection of the idea that there are properties of right and wrong out in the world independent of us, a rejection I endorse. Understanding this metaphysical truth, Greene says, we will know that there is no such thing as a right to life, nor a property of being sacred that life in general could possess. What is left is just a reasoned consideration of what things we care about. And one thing we do, all, care about is the happiness and desires of the relevant parties.

The first thing to notice about these debunking arguments is that they are essentially negative. If we discount whatever intuitions we may have about difficult ethical choices, we still need some other guideline for deciding how to act. So one theme of Greene's arguments thus centers on considerations about which of the factors influencing moral judgment we would on reflection endorse as morally relevant and which we would not. In regard to the trolley problems on which his empirical work as focused, for instance, Greene suggests that the difference between a certain killing that requires up-close "personal force" and an otherwise identical killing that requires only the remote pushing of a switch is not a morally relevance difference. Reasonable people will agree, Greene thinks, that two killings differing only on this factor ought to be equally permissible or impermissible. Greene and colleagues' more recent work shows, however, that personal force is not the only nor the most relevant difference between the trolley conditions. Greene et al. (2009) tested various iterations of these dilemmas varying whether the victim is killed as a means for preventing the deaths of others or instead as a side-effect, and secondly whether force generated by the agent's own muscles directly impacts the victim. In judgments of whether the action was "morally acceptable" they found that the effect of personal force depended entirely on whether or not the death of the victim was intended or instead merely a side-effect. Moreover, Cushman and Young (2011) find that this means/side-effect distinction affects moral judgment to the degree that subjects attribute to agents the intention to cause harm.

Intuitively, intent does seem a morally relevant factor in making moral judgments. This is a factor notably absent from the list Greene (2002: 328-9) offers of things we might care about and therefore consider in deciding whether to end the life of a severely debilitated infant. If the empirical evidence I reviewed in Chapter 2 is correct, however, agents' intentions are one thing that adult human beings do care about. Of course, on Greene's line of thought, the fact that we happen to have been endowed by biological or cultural evolution with intuitions to the effect that intent is import to ethical judgments says nothing about whether the sorts of ethical judgments this intuition drives are the correct ones to make. As an argument for welfare-consequentialism, however, this debunking move has its own irony. For two central and inter-related conceptual foundations of consequentialist ethical theorizing are themselves based on intuitions that are subject to the same critique. First, the impersonal 'view from nowhere' from which right and wrong are to be judged, and secondly, the focus on outcomes as the criteria for making these judgments, can each be shown to have evolutionary origins at least as plausible as the reconstructions Greene provides for deontological intuitions. And these are precisely the points on which my Buddhist-inspired account differs from Greene's utilitarian one.

In Chapter 3, I have noted evidence suggesting that both over evolutionary and also developmental trajectories the outcome-based system of moral judgments comes to be constrained in adult human beings to various degrees by a system responsive more purely to the motivations of the agent of an action. Consequences are outcomes, and so consequentialist judgments are unavoidably outcome-based. It is crucial for classic as well as more recent utilitarian theories that the decision procedure, the procedure one employs to decide which actions or motivations are the right ones, is separable from (although on some theories an approximation of) what actually makes an action or motivation right. Thus Sidgwick writes, "It is not necessary that the end which gives the criterion of rightness should always be the end at which we consciously aim" (1907, 413 in Sinnott-Armstrong 2009). This is important to the defense of such utilitarian theories because no one individual could possibly know with certainty the net effect of a particular act on aggregate welfare. Thus when we evaluate people's virtue or vice, we often judge them on the basis

of whether they acted in the way that they themselves expected would lead to aggregate welfare. Nonetheless, for a consequentialist theory, what makes an action or a motivation right or wrong is the actual effects it has in the external, inter-subjectively available world.

One can have a consequentialist theory on which the primary unit of ethical evaluation is the motivational state of the agent; Adams' (1976) "motive utilitarianism" is one example. Nonetheless, on such a view, the actual consequences of having a certain motivation are the criterion that determines whether such a motive is right or wrong. By way of comparison, according to AWA, acts that are (necessarily) motivated by hatred are wrong just in virtue of the fact that one ought not to have such motivations, and hatred is wrong just in virtue of the fact that it is a motivation that feels bad, to one who is fully and accurately aware of their own emotional states. Although my theory has implications for which actions in the world are right and wrong, and I claim that the criteria of rightness and wrongness is common to all human beings, nonetheless the view-point from which rightness or wrongness is to be determined is essentially first-personal. This is what makes the normative basis available, in principle, to any of us human beings. On a consequentialist metaphysics of value, in contrast, what makes an act or a motivation right or wrong are features of the act that, in effect, only an omniscient agent could fully calculate.

Some of the earliest utilitarian theories were indeed theologically based (Driver, 2009), and it is plausible that the "view from nowhere" featured in modern utilitarian reasoning derives historically from the Christian conception of a divine, omniscient being. In an insightful article on the adaptive value of costly religious displays, however, Slingerland et al. (2013) identify this kind of moralized view from nowhere as a more general social-psychological construct underlying many different religious institutions. They note that as in Western feudal societies, the early Chinese notion of a "Mandate from Heaven", *tianming*, played a powerful role not only in legitimizing early Chinese dynasties such as the Western Zhou, but also in compelling elaborate and costly ritual displays by the rulers as well as their subjects. The authors claim that in early Chinese thought, much as in Abrahamic religions, Heaven also functioned as a moralized agent with full epistemic access, "aware of and prone to judge one's actions and inner thoughts."

Slingerland et al. suggest that such "supernatural surveillance" represents an important factor in packages of cultural adaptations conveying a selective advantage to certain societies. In other work, Henrich et al. (2010) demonstrated that participants in one-off interactions in the Ultimatum Game and two variants were more likely to make equitable offers and more likely to punish inequitable offers if they came from developed societies with large community size. In light of evidence that god concepts prime increased pro-social responses in such one-off economic games (Shariff and Norenzayan, 2007), it is plausible that as part of a cultural package, the moralized view from nowhere may help to facilitate economic activity and coordinate action among large populations. In small-scale societies, reputational considerations can dissuade cheating and motivate cooperative exchanges. In such contexts, one-off exchanges with individuals outside of the society are at least as likely to be antagonistic or exploitative as to provide an opportunity for mutually beneficial economic exchange. In contrast, large scale societies provide and depend on possibilities for mutually beneficial one-off interactions with strangers, as in the global marketplace. Henrich et al. extrapolate from their evidence to argue that expectations of equitable treatment and costly punishment of unfair treatment provide net societal benefits in the context of such large-scale societies. Thus Slingerland et al. suggest that "as we move from small-scale to large-scale societies, supernatural agents become increasingly morally concerned, more effective at monitoring norm violations (omniscience), and better equipped to provide punishment and rewards (heaven and hell) according to prescribed behavior."

The notion of a morally concerned, omniscient perspective as determining the moral valence of human actions naturally leads to implicit and explicit claims that one's moral views are reflected in the objective structure of the world, though the combination of this moral realism with the notion of a supernatural author of the moral code can also lead to the tension indicated in Plato's *Euthyphro*. In the case of welfare consequentialism and other modern ethical theories, what I have called above realism about response-independent moral facts has survived the abandonment of theological convictions. Nonetheless, Slingerland et al. suggest that their account to may provide "a cultural evolutionary explanation for the emergence of the moral realism that now pervades both

religious and secular discourses."

In contrast to classical Utilitarians, Greene himself explicitly rejects this sort of realism, as we have noted. However, in offering a positive account of the considerations we ought to take into account in deciding difficult cases, he nonetheless highlights considerations about the consequences of the action for aggregate welfare, to the exclusion of the considerations about the agent's intention that my account focuses on. In this, his account continues to function as if the rightness or wrongness of an action is determined by the sort of net aggregate welfare that could only be known from a perspective of omniscience. I propose to replace the impersonal, response-independent guideline for judging how one ought to act, by instead grounding claims about what ought to be done in claims about which sorts of emotional motivations we ought to have, and in turn to ground claims about which sorts of emotional motivations we ought to have - for instance the claim that we ought to have compassionate responses - in the predicted convergence, among individuals who feel fully their own emotional motivations, regarding which emotional motivations they would want to have.

One way for the utilitarian to miss the point is to reply to the effect, "Look, it's a fact that people care about each other. If you want to call that a first-personal grounding of utilitarian concerns, fine. My project is to figure out what we should do, given that interest." It is indeed a fact that people care about each other, in certain situations. But this is not sufficient to establish that utilitarian ethical judgments are the right ones in difficult cases. People also have strong dispositions not to push each other off bridges, and not to torture or to smother babies, even when aggregate welfare would be so maximized. In addition to compassionate impulses, human beings also exhibit strong tendencies to turn away when others are in need, to exploit other people for material benefit and for sexual pleasure, to enslave whole peoples or to exterminate them. We act on all sorts of concerns. The job of an ethical theory is to tell us which of the motivations we have we ought to act on.

Greene appeals to considerations about which of our emotional reactions we owe to evolution rather than reason as a way to break the impasse between the intuitions appealed to in philosophical debates between the two systems. I have granted for the sake of argument that the debunking strategy that Greene and Singer employ against the intuitions that drive (some) deontological judg-

ments might work, but countered that if it does, it applies also to any consequentialist intuitions we might have to the effect that outcomes matter for the ethical valence of an action, rather than just intentions. As social-psychological (and political) constructs, modern ethical theories may appeal to common intuitive reactions that have propagated over the course of human religious history because of the adaptive advantage these constructs bestow on the group. If Slingerland et al. are correct, the moralized view from nowhere that underlies the consequentialist metaphysics of ethical value may take its appeal from intuitions that are now common-sense precisely because they have succeeded in propagating themselves over the course of cultural evolution. For if those arguments work, the fact that we have intuitions to the effect that consequences for aggregate welfare matter to the judgment of whether a given act ought to be done should be given just as little weight as the fact that we have stronger intuitions against causing harm in an up-close and personal way than we do to causing equal harm by flipping a switch. If so, the empirical considerations that Greene has brought to bear in the attempt to break the stalemate of intuitions between deontological and consequentialist camps are not up to that task.

Nothing about this refutes Greene's utilitarianism, of course. The *tu quoque* move just serves to take away the legitimacy of an appeal to evolutionary considerations as way of debunking deontological intuitions and leaving consequentialist ones standing. For all I have said, one might still find either logical or psychological considerations that do better. Such considerations might lead us to favor an approach focused on volitions and based on rational principles, as Kant proposes, or instead might lead us to favor of an approach such as Greene proposes, one focused on aggregate consequences and based in hedonistic considerations of empirical facts about pain and pleasure. Alternatively, novel logical or psychological considerations might lead us to favor an approach that combines the strength of both of these, one taking volitions as the primary evaluative focus, with Kant, but basing this evaluation of various motivations for action on hedonistic considerations of empirical facts about pain and pleasure, with Greene. My bet, of course, is on this last hypothesis.

Greene appeals to psychological predictions about which sort of an ethical system reasonable people will choose once they know the psychological and metaphysical facts, the truth about moral-

ity (Greene 2002). And like Green, I appeal to an similarly empirical sort of ideal observer account. So one might think that the *tu quoque* I have leveled against Greene's attack on deontological intuitions should apply equally to my own account. After all, I base my account on how various sorts of emotional motivations feel to the person so motivated. My claims for convergence among human beings on ethical principles turns on the shared physiology of these emotional states, and the shared affective reactions to this shared physiology. Perhaps this very universality, having no doubt come to be as it is in virtue of contingent evolutionary processes, should serve to debunk the epistemic value of such reactions as a way of knowing what is ethically right and wrong. Unlike Greene's approach, however, nothing about the argument I offered in Chapter 3 requires appeal to genealogical considerations. The ultimate justification for my ethical system is not genelogical nor logical considerations, but rather as with Mill's account, the ultimate ground of moral justification "is but a subjective feeling in our own minds".

The ironic evolution of Utilitarianism is two-fold. The classic utilitarian move begins from the intuitive claim that pleasure, impersonally conceived, is the good. The trouble is that in difficult cases, conflicting intuitions are equally powerful. In order to break this stalemate, modern Utilitarians like Greene move from a logical to a psychological basis. Greene's move is to base utilitarian values on a prediction that when things go right epistemically and psychologically, people will converge in endorsing welfare-consequentialist conclusions. However, there is no reason to suspect that knowing the truth about the evolutionary and childhood development of moral judgment, people will endorse outcome-responsive moral judgments over intent-based ones. Just the opposite, if my argument above is cogent. The second irony is that Mill himself seems already to have endorsed a psychological basis, rather than a logical one, and a more plausible psychological basis than Greene does. Mill's idea seems to be that when things go right epistemically and psychologically, people will feel for themselves that feeling sympathetic to others' suffering is a good way to be. In other words, on Mill's account, it is not that aggregate welfare is the ultimate good and compassionate feelings are instrumentally good because they lead us to maximizing aggregate welfare. Rather, the argument seems to run the other way: one ought to act to maximize aggregate welfare

because that is what one would try to do if one was motivated by compassion, and compassion is the sort of motivation we would prioritize as a good one if all was going well psychologically and epistemically. Above in Chapter 3, I proposed just such a conclusion.

4.3 Kant and the Ground of Subjective Universal Principles

In the previous section I have surveyed the empirical research of Greene's own and others that he has drawn on with substantial rhetorical effect to attack Kantian deontological ethical conclusions. Although Kant is famous as the ultimate rationalist, on Greene's account this is a deep irony. Science tell us, Greene would have it, that deontological theorizing in fact expresses post-hoc rationalizations of plebeian intuitive reactions determined by evolution rather than reason. Paraphrasing Nietzsche, Greene (2008) calls this "the secret joke of Kant's soul". I detail in Section 4.4 how this argument of Greene's can backfire, undermining as well some central intuitions underlying welfare consequentialist reasoning. And the normative account developed Chapter 3, building on the empirical results surveyed in Chapter 2, offers a defense of central aspects of the Kantian approach, in particular the normative primacy of an act's volition over its consequences. But before we turn to these questions, a point on Kantian exegesis.

If Greene is accusing Kant of using reason to defend common moral judgments, there is no secret joke here. Kant explicitly claims to be doing just that. A normative theorist cannot hope to operate from any basis other than common sense morality, Kant suggests in *The Groundwork of the Metaphysics of Morals*, though he can muddle the task. "A philosopher, though he cannot have any other principle than that of common understanding, can easily confuse his judgment by a mass of considerations foreign and irrelevant to the matter and deflect it from its straight course" (GMM 4:404 in Kant, 1998). For this reason, there is something splendid about the "innocence" of common sense. The problem, nonetheless, is that common cognition "cannot protect itself very well and is easily seduced." In particular, common sense does not have the tools to protect itself from philosophical onslaughts such as Greene's, and it is easily seduced by "a propensity

to rationalize against those laws of duty... and, where possible, to make them better suited to our wishes and inclinations" (4:405). It is for these practical reasons, so that common moral cognition may "escape from its predicament about claims from both sides" that Kant embarks on his project of finding the supreme principle that already underlies it. The point of making this supreme principle explicit is not to reform the fundamental ground of common moral cognition, but on the contrary to make accessible and available the ground already implicit in common moral cognition for use in those cases when it needs to defend itself either against speculative obscuration or against motivated rationalization, or, we might add, against both acting in concert.

Kant sets out on the project of providing this groundwork for the meta-physics of morals by appealing, famously, to intuitions about the overwhelming importance of the quality of the will in making an act good or bad. This appeal specifically aims to undermine any relevance to moral worth that prudential considerations about outcomes might be thought to have, in particular the sort of considerations about hedonic consequences that would sway one emotionally towards favoring a particular outcome.

> A good will is not good because of what it effects or accomplishes, because of its fitness to attain some proposed end, but only because of its volition... if with its greatest efforts it should yet achieve nothing and only the good will were left (not, of course as a mere wish but as the summoning of all means insofar as they are in our control) - then, like a jewel, it would still shine by itself, as something that has full worth in itself. (GMM 4:394 in Kant, 1998)

This far, the Buddhist-inspired that I am proposing is very much in sympathy with Kant. Both endorse volition as the primary unit of normative evaluation, though as I note below, these two approaches depend on two very different notions of what a goodwill consists in. Nonetheless, the initial appeal of either of these theories may depend on very similar intuitions. For Kant, being trustworthy or benevolent has an inner worth that "does not consist in the effects arising from them... but in dispositions, that is in maxims of the will that in this way are ready to manifest themselves through actions" (4:435). For me, the physiological and affective states that directly

dispose one to action are the means that one must summon, and it is these sorts of dispositions that are essentially good. Thus I see thoughts about ethical principles as only indirectly related to action, in virtue of their ability to summon the physiological and affective states that directly dispose one to behavior.

That difference notwithstanding, a further similarity lies in the subjective but universal grounds my approach and Kant's cite for claiming that judgments of intention have more normative force than judgments of outcome. We both suspect that the reason past theorists have failed to "discover the principle of morality" is essentially that it "never occurred to them that [the human being] is subject *only to laws given by himself yet universal*" (4:432). More precisely, we agree on the idea that the moral norms to which a person is subject are "given by himself", and that these are "yet universal", though where Kant holds that it is laws of reason that subjects universally give to themselves, I hold instead that it is universal aspects of human emotional reactions that structure human evaluative worlds in common ways.

There are three connected claims that serve as premises for the arguments that Kant uses to move beyond the point he makes about the goodness of a goodwill to the further claim that this goodness is grounded in *a priori* principles of reason. Interestingly, these three premises turn out to be empirical ones, and it is by taking opposite stances on these three claims that my more Humean account of the goodness of goodwill diverges from the Kantian one. I discuss each in turn, and the evidence available to adjudicate between them.

An empirical refutation of Kant might seem to miss the point. After all, one of Kant's central points in the Preface to the *Groundwork* is the irrelevance of empirical considerations.

> Everyone must grant that a law, if it is to hold morally, that is as a ground of obligation, must carry with it absolute necessity; that, for example, the command "thou shalt not lie" does not hold only for human beings, as if other rational beings did not have to heed it, and so with all other moral laws properly so called; that, therefore, the ground of obligation here must not be sought in the nature of the human being or in the circumstances of the world in which he is placed, but a priori simply in

> concepts of pure reason; and that any other precept, which is based on principles of mere experience - even if it is universal in a certain respect - insofar as it rests in the least part on empirical grounds, perhaps only in terms of a motive, can indeed be called a practical rule but never a moral law. (AK 4: 398; trans., Gregor)

The conclusion of this argument, that laws that are properly moral cannot rest in the least part on empirical grounds, depends on the premise, that everyone must grant that moral laws hold not only in the circumstances in which humans find themselves, but absolutely, as a conceptual truth. One might take this as a claim about what the application of our moral concepts ought to be. In this case, even if there were people who happened not to use moral expressions in this way, the claim would be that they are wrong not to. But it is unclear on what basis one could decide how it is that we *should* use moral concepts, as opposed to how we do. One might appeal here to a Platonic thesis about concepts, holding that there are objective grounds for assessing the referent of moral concepts of right and wrong, independent of how any human community actually does use the relevant terms. I cannot address here the merits of such an approach; suffice it to say that an ethical system premised on such a metaphysical thesis would rest on grounds that are at best deeply controversial.

In any case, Kant does not seem to take his claim about the referents of moral concepts in this way. The dialectic of the *Groundwork* proceeds by first analyzing "common cognition", starting "from our ordinary ways of thinking about morality... to discover the principle behind them" (Korsgaard in Kant, 1998, xi). This suggests that when Kant claims that "everyone must grant" a moral law to be one that has absolute necessity, he means this as an empirical claim about how we commonly use moral concepts. It might not be sufficient to falsify this thesis just to find an instance of some speculative philosopher or other who has convinced himself of a way to use moral concepts without such absolute necessity. However, it would be more damaging to the argument if it turns out that common folk routinely use moral concepts in ways that do not imply this sort of absolute necessity.

Kant's claim for the irrelevance of empirical considerations thus rests on an empirical premise,

and, it turns out, one on which the available evidence is not wholly in his favor. It is true that one of the most robust characteristics of moral judgments are their generalizability. In the classic study by Tisak and Turiel (1984) noted above, children were more likely to say that moral norms applied "in another city" than prudential ones. Moving slightly more broadly afield, Nichols and Folds-Bennett (2003) show that children tend to take harm and disgust violations as wrong also "in another country or someplace far away." On the one hand, such evidence is consistent with Kant's claim that moral claims apply not to only to human beings, but rather to rational beings as such. On the other hand, it is equally consistent with the sort of account that Greene and I favor, on which the human practice of moral judgement emerged to deal with human situations, and does so without making any substantive claims about remote possible worlds. Moreover, some recent evidence suggests that objectivist intuitions about moral properties depend on implicit assumptions about shared subjective values: Sarkissian et al. (2010) gave vignettes to undergraduates in South Carolina and in Singapore in which two individuals diverged in their judgments about whether a certain act was wrong. In cases where the individuals in the vignette making the moral judgments came from radically different cultures or ways of life, respondents were less willing to say that one party or the other had to be mistaken. If so, this suggests that people use moral concepts in ways that do not require everyone to grant that moral truths hold for all rational beings as such.

Kant takes his conclusions about the supreme principle of morality to be *synthetic a priori*, resting on an analysis of how the subject must conceive of moral concepts, and therefore not derived directly from analysis of concepts. Nonetheless, the premise from which he derives his conclusions, namely that moral judgments are taken to apply categorically to all rational beings as such, he does take to be shown "by mere analysis of the concepts of morality" (4:440). And in the absence of evidence about how other rational beings use moral concepts, or even the ability to perceive moral judgment in non-human rational beings, we are left with only the evidence about how we human beings use moral concepts to adjudicate questions of how such concepts apply. Moreover, the bold claim that we do use moral concepts to apply with absolute necessity to all rational beings clearly turns on empirical considerations of contingent fact. And since it

serves as a premise for Kant's argument, the burden for showing this empirically rests with the Kantians. Nonetheless, it is not at all clear how one would go about substantiating this claim. For even a vignette involving moral transgressions by Martians might only trigger human practices of moralizing to the degree the beings involved are perceived to share certain human characteristics. It is an empirical question whether when presented with a perfectly logical machine, lacking eyes or a face or the ability to express emotion, we would apply moral judgments, in particular evaluations of a good or bad will. What is worse for the Kantian is that however the empirical results happen to come out on this question, facts about how we use moral concepts are contingent, and so the range of application this usage implies is also contingent. For this reason analytic truths about what reason implies for morality turn out to rest on a fragile basis. Even if it were true that we currently use moral judgments to apply with absolute necessity, all that is needed to change such an analytic truth is a change in human practices of applying moral concepts. If it is easier to change such a social practice than to change the underlying biology and psychology of our emotions, then it is emotional universals and not conceptual truths that serve as the more steadfast ground for constructing an ethical system.

This brings us to a second and related empirical claim, regarding which psychological faculties are under our control. In offering his famous example of the altruist, Kant notes that some people's cognitive systems just happen to be structured such that they "find an inner satisfaction in spreading joy around them and can take delight in the satisfaction of others so far as it is their own work" (4:398). But the trouble with such inclinations is that they are fickle, contingent on conditions that are beyond our control. The same person when overcome by grief might not be so inclined, and yet another might never have been, due to biological happenstance. Kant's argument that inclinations do not have the moral worth held by the commands of reason is premised on the claim that inclinations are not under our control, that the only thing under our control is whether or not we act from duty, irrespective of whether our inclinations happen to be congruent or incongruent with what it requires. Thus he comments that the Biblical passages in which we are "commanded" to love our neighbor, even our enemy, must mean that we are to promote others' happiness not

from inclination but from duty.

> For love as an inclination cannot be commanded, but beneficence from duty - even though no inclination impels us to it - is *practical* and not *pathological* love, which lies in the will and not in the propensity of feeling, in principles of action and not in melting sympathy; and it alone can be commanded. (4:399)

Here again, a great deal turns on how we define inclination, and what we mean by whether or not such psychological forces can be commanded. Nonetheless, it is clear that in principle this is not a conceptual claim about what it means to be an inclination or what it means to be able to be commanded, but rather an empirical claim about how various psychological forces relate to one another. To take just one instance, the evidence from research on mindfulness suggests that emotional dispositions are plastic and can be trained. If so, emotional reactions are not beyond our control. Indeed, whatever we do we are in effect choosing to train one habit of mind or another; the question is whether who are training good habits. It is because emotional dispositions to action are susceptible to training that they can be evaluated ethically.

I noted above that for Kant as for me, it is the disposition to certain sorts of action that ought to be evaluated ethically, "the essentially good in the action consists in the disposition, let the result be what it may" (4:416). We agree further that the ultimate goal of freedom depends on determining which sorts of dispositions are to be cultivated, "independently of any property of the objects of volition" (4:440). For Kant, this means that we must determine the will based on rational principles of non-contradiction and the like. The account of knowing how to live wholeheartedly given in Chapter 3 in effect offers a different route to self-mastery, a type of freedom from being controlled by the emotional forces of self-interest that depends on full and accurate perceptual awareness, rather than on ideals of rationality. Here, just as Kant and I agree that previous theorists missed the important possibility of a subjective ground for ethics that nonetheless applies to all human beings, my suspicion is that the Kantian program turns to reason as a means for freeing the human will from being motivated by external rewards, and therefore to absolute commands of reason as the grounds of ethical value, just because it overlooks a different and more direct

means. In particular, I want to suggest that this rationalist approach overlooks the possibility of finding freedom through the sort of being Wide Awake that I described in the second and third chapters. My positive proposal is that the many important points the Kantian approach gets right can be saved from the problematic dependence on ideals of reason, and in particular on social facts about whether we use moral terms in ways that imply application to all rational beings as such, by grounding the value of benevolent motivations over self-interested ones instead in facts about which sorts of motivations feel better to any human being who is feeling them fully.

It turns out that much of the need Kant sees for his problematic dependence on ideals of reason derives from the practical problem of how to counteract motivated self-delusion, from our tendency to make exceptions for ourselves from rules we do hold that others should follow. In demonstrating the practical importance of finding the ground of ethics in rational principles, "completely isolated" from empirical considerations (4:408), Kant cites a third empirical premise, regarding the psychological efficacy of reason. He remarks that it is clear that much of what we do, even when it is in conformity with duty, is actually done for the sake of "the dear self, which is always turning up" (4:407). Buddhists will of course agree with this diagnosis, if not with Kant's suggested cure. Seeing how pervasive such hidden motivations are, Kant thinks that the only thing that can defend the notion of true moral worth is the conviction that even if there have never been any empirical instances of action from duty, still "reason by itself and independently of all appearances commands what ought to happen" (4:408). An account of the *a priori* ground of ethics is not only a theoretical desiratum, but also a practical one, crucial to helping us convince ourselves to act as we should.

> For, the pure thought of duty and in general of the moral law, mixed with no foreign addition of empirical inducements, has by way of reason alone... a influence on the human heart so much more powerful than all other incentives, which may be summoned from the empirical field, that reason, in the consciousness of its dignity, despises the latter and can gradually become their master. (4:411)

Testing the claim that reason is a more powerful psychological influence than other incentives turns on the question of how we define in empirical terms the psychological faculty of reason, and it may

be especially difficult to decide how to operationalize the notion of pure reason. Nonetheless, if this is important to the Kantian account of the importance of finding a rational ground for ethics, then here again it seems the burden is on the proponents of such an approach to show how we could test this claim.

Although Kant does in places sound as if reason operates on human judgment and action independent of all affective incentives, Paul Guyer points out that especially in his later works Kant softens this point, taking reason to operate on the human will in virtue of a decidedly Humean mechanism. In particular, it is ultimately a "passion for freedom" that is to be transformed, through reason, from a passion for one's own freedom into a passion for the freedom of all. As Guyer puts it,

> the idea seems to be that the determination of the will by the moral law leads to action by reweighting our natural incentives: it makes the naturally pleasurable prospect of indulging our own inclinations painful and transforms the naturally painful prospect of thwarting our own inclinations into the pleasurable prospect of living up to the moral law, and this realignment of our prospects for pleasure and pain is what leads to our dutiful action. (Guyer, 2012, 14)

In general, this sounds very much like the sort of function I predict will be played by feeling fully one's own emotional motivations: a reweighting our natural incentives such that motivations of ill-will and of craving will come to feel uneasy for us, and relative to these, the emotional states that dispose us to helpful behavior towards others and to contentment in relation to our own needs will come to feel much more full of ease. Such an account need not deny the important role of reason in assessing how best to accomplish the sorts of goals that are motivated by feelings of benevolence. Moreover, careful discerning thought is often crucial to weighing various values, for instance in making the right decisions about which sorts of emotional dispositions to cultivate. Nonetheless, my account may part ways with the Kantian one over the metaphysics of value, since I hold that what makes certain emotional dispositions the right ones to cultivate are not facts about logical coherence or rational principles, but rather facts about the way humans beings are constructed.

4.4 Hume and the Humeans on Moral Emotions

Like Kant, Hume and his heirs take it that human beings are subject to laws given by themselves. For Hume and the Humeans, however, the ground of these laws lies not in reason but in passion. AWA adopts an approach close to Hume's own in a number of ways. Hume proposes that "we do not infer a character to be virtuous, because it pleases: But in feeling that it pleases after such a particular manner, we in effect feel that it is virtuous" (Hume, 2000, III.1.2). On its own, this proposal is cryptic enough that it could perhaps be endorsed and also rejected by almost any theorist of ethics, depending on how it is understood. One might read Hume's account as a first-order sentimentalism, on which it is just in virtue of feeling a certain way about the sorts of things that a person does that we judge those acts, and perhaps also the person, to be bad. Modern sentimentalists of this ilk have emphasized that since people's actual emotional dispositions and the judgments vary widely between cultures, if these are the sort of facts that make an ethical claim true or false, then ethical truths are also culturally relative. Hume himself seems not to have drawn this conclusion, suggesting instead that certain emotional dispositions may provide a ground for ethical claims that is both affectively based and also shared among all human beings. He notes for instance of "certain calm desires and tendencies... either certain instincts originally implanted in our natures, such as benevolence and resentment, the love of life, and kindness to children; or the general appetite to good, and aversion to evil, considered merely as such" (Hume, 2000, II.3.2).

Michael Slote's (2010) Humean account focuses on one such universally human emotional response: empathy. On Slote's view, roughly, the goodness of an act is constituted by empathetic concern for those affected by that act, and moral approval is constituted by empathy with the empathic feelings of one doing the act in question. I am sympathetic to Slote's Humean vision of a non-relativist moral system grounded in universal benevolent emotions, and much more could be said about his particular approach than I can say here. Like Slote, I think that there are facts about human emotional life that are universal enough to ground substantive ethical claims that will apply to any human being, and that these grounds can be discovered through examination of our own emotional lives. Thus, like Slote, I draw some inspiration from the notion of a moral sense, of the

sort that Hutchinson thought could ascertain the value of universal benevolence. However, there are a number of weaknesses in the argument by which Slote arrives at this conclusion. First and most centrally, I am skeptical of the reference-fixing account of moral concepts that Slote offers, precisely because of its putative strength, that is, that it rules out relativism on conceptual grounds. On the more Humean account I have offered in Chapter 3, in contrast, it is an empirical issue whether or not there is any epistemic method that leads to agreement on universal ethical truths. And if there is, this empirical approach seems to me a much more likely to fulfill the practical desiratum of providing grounds for adjudicating ethical disputes between cultural systems. In this regard, the meta-ethical basis of Slote's account is more rationalist than Humean, and is subject as such to the critiques I surveyed in particular in regard to Kant in Section 4.3. Secondly, in order to fix the reference of moral concepts Slote relies heavily on the intuitive appeal of a notion of being empathically warmed or chilled. One might develop a less metaphorical account in the terms of cognitive and affective neuroscience, for instance. But Slote does not engage closely with the empirical details of empathy and emotional contagion. If he did, the account might run into a number of troublesome empirical results (for a survey see Prinz, 2010). Thirdly, while empathy might play a role in understanding harm-based moral values, Slote does not survey other types of emotional dispositions; to ground common-sense values about transgressions against community and against the natural order, emotions such as contempt and disgust (respectively) would at least initially seem to be much better candidates than empathy.

Shaun Nichols (2004) does engage seriously with the empirical details of empathetic reactions in moral psychology, and is therefore in a position to offer a more precise account of which aspects of empathic concern could ground ethical claims. He suggests that the backing for norms about harm violations comes from forms of empathetic concern rather than the personal or contagious distress involved in empathy. Nichols rightly suggests that with both harm and disgust violations, affect plays a role in leading children to treat these as objectively bad. However, such judgment cannot be wholly explained by appealing to affect, Nichols contends. On his view, affect is neither necessary for sufficient for making a harmful action into a moral issue. Children employ

a rudimentary normative theory to specify which actions are transgressions. In motivating this suggestion for the case of harm-norms, he surveys cases in which eliciting negative affect is not sufficient for moral condemnation. Toothaches may be "bad" but not "wrong". Seeing victims of accidents or natural disasters induces emotional response but not negative moral judgment (2004, 15-16). Superficial distress cues apparently have similar results. Likewise with disgust norms, "unintentional vomiting" is not counted as a transgression, despite being disgusting (2004, 25). But Nichols' wording suggests an obvious explanation: the ascription of intention might be the deciding factor in these cases, rather than a normative theory specifying which emotionally charged actions count as transgressions. Seeing superficial distress cues or a girl skinning her knee by accident, one might well not perceive any intentional harm. And when a parent intentionally applies antiseptic to the cut, one may rightly perceive the parent as having no desire for their child to be pained, just the contrary. On the account I offered in Chapter 2, that is why applying antiseptic is not judged to be wrong. To show the further necessity of a normative theory, we would need instances where the factors of intent and causation are present, but not moral judgment. Nichols notes that "there are even intentional actions that are disgusting but not prohibited" (Nichols 2004, 25). But after giving the (sole) example of parlor tricks (and leaving it up to the reader to specify the details), he immediately concludes that "there seems to be a body of information, a normative theory, proscribing a class of disgusting behavior" (Nichols 2004, 25). This seems to me too quick.

Nichols is noncommittal on the nature of the normative theory that is needed to specify those harmful or disgusting acts that count as moral violations. Indeed, for all Nichols says, the sort of implicit, behavioral role that normative theory plays in his account could be cashed in terms of socialized emotional dispositions. If so, Nichols' brand of sentimentalism appears on investigation to approach quite closely that of Prinz (2007). On a rough first pass at Prinz's account, it is the dyad of anger and guilt responses that serves to specify which harmful acts count as moral violations. Acts that cause pain, such as putting anti-septic on a child's cut, may trigger empathetic distress or empathetic concern. But we don't get angry at people who do such a thing, nor do we feel guilty if we were to do it; just the opposite. On Prinz's account, it is these sorts of emotional

dispositions that specify which acts are wrong. And as with the rudimentary normative theory proposed on Nichols's account, these emotional dispositions are culturally constructed. Moreover, Nichols agrees with Prinz that there are no truth-makers that could ground universal ethical claims; if Nichols is less confident about the thesis of moral relativism, it is only because he is ambiguous between this thesis and an error-theoretic account on which our moral claims do not make true assertions, whatever other uses they may be put to.

On a first-order account such as Prinz's (2007), one who has the emotional disposition to feel approbation towards a certain behavior thereby has an ethical value in favor of such acts. Take someone who feels that it is right to enjoy life and make the most of one's sexual pleasures, or someone who feels that it is right or even required to do one's best to be fruitful and multiply. One Prinz's account such (first-order) emotional dispositions serve as the truth-makers for ethical claims, such that if our protagonist (or his moral community) does have pro-attitudes towards these things, then he is correct in claiming that sex is right or even required. This may lead him to take a stand regarding the preferences that other people have, perhaps claiming that living a celibate monastic life is less than praiseworthy, for instance. On a higher-order sentimentalist account such as Blackburn's (1998), in contrast, it is only when one takes a stand regarding (first-order) attitudes in this way that one expresses moral value. For Blackburn, a pro-attitude towards sex or procreation is just that. But such a (first-order) sentiment is itself subject to endorsement, or censure, by oneself or another. It is in these cases, where our preferences dispose us to criticize or praise others' preferences, that the issue is treated "as a matter of public concern, as something like a moral issue" (Blackburn, 1998, 9). On still a higher level, one is prepared to censure those who do not share one's sentiment (for instance, of approval for sex). And going up another step, one becomes prepared to censure those who tolerate (and do not censure) those who do not share one's sentiment.

Theorists as diverse as Blackburn, Gibbard, McDowell and Wiggins agree that to take cannibalism to be wrong is not (just) to feel disgust, but rather to think that it is "appropriate" to feel disgust and anger, say, in response to eating human flesh. But a further refinement to such neo-

sentimentalist views is needed, D'Arms and Jacobson (2000) urge, because very different sorts of considerations could bear on whether it is appropriate to feel a certain way. It might not be prudent to feel outrage towards a tyrant, but this makes his actions no less of a moral outrage. "To judge an emotion is fitting is not to think it adaptive but to endorse its evaluation as correct. This is a kind of a higher-order attitude toward the emotion, which we reify by wielding a vocabulary of regulative terms such as fearsome, shameful, and funny" D'Arms and Jacobson (2003, 145).

The "rationalist sentimentalism" D'Arms and Jacobson develop is rich and interesting. Indeed, my own account echoes in central ways both their suggestion that we can use evaluative terms to regulate our emotional reactions, and also the suggestion that we can in important ways be more or less "alienated" from the values our actual sentiments embody (D'Arms and Jacobson, 2010). Nonetheless, D'Arms and Jacobson seem not to be centrally interested in the question of whether their rationalist sentimentalism will provide any universal grounds to settle ethical disagreements. On the one hand, a rationalist form of sentimentalism might maintain the general Humean line that emotions are essential to morality, while still resisting relativism by grounding rightness and wrongness not in emotional dispositions themselves, but rather in the concepts we apply to them. The cost of adopting this sort of rationalist sentimentalism is that the same sorts of problems that I have detailed in regard to the Kantian form of rationalism will apply. As I noted there, if the application of our concepts is determined by human social practice, this may be less stable and indeed less universal that the features of emotional reactions that we share as human beings. If it happens that the regulative concepts we use in our moral community specify certain sorts of actions as outrageous, it does not follow that the regulative concepts used in other moral communities will specify the same set of actions. Worse, even if all humans beings happened now to agree on how to apply the concept outrageous, this would offer no guarantee of long-term stability.

On the other hand, as with my emphasis on the ethical evaluation of emotional motivations, the emphasis on the regulative role of evaluative concepts might be important to the larger story of a rationalist sentimentalism, and yet claims about what is outrageous might turn ultimately on the affective and physiological aspects of emotion. On this route, I have suggested, there might

be some hope of finding ethical claims that are universally true, in virtue of universal aspects of human emotion. This strategy for resisting relativism requires a deep and detailed engagement with the empirical details of the affective and physiological aspects of emotion, however, and for all their rich analysis of social and phenomenological dynamics of sentimental judgment, D'Arms and Jacobson's project seems not to be aimed in this way.

A Humean account can of course bite the bullet and accept that ethical truth does vary between cultures, in which case it need not provide evidence of either conceptual or biological universals. Prinz and Blackburn agree that while it is a fact that we do express sentiments, and sentiments about sentiments (and so on), using ought statements, there need not be any further fact about what sentiments one actually ought to have. Blackburn can say truthfully (or at least quasi-truthfully), from an internal perspective, that the wrongness of cruelty is independent of anyone's responses. The explanation of this judgment of Blackburn's, from an external perspective, is the fact that he, now, has strong attitudes of approval towards attitudes disapproving of cruelty. Thus imagining a case of someone (such as his future self) approving of attitudes approving of cruelty he will disprove, strongly. For Blackburn, that is all there can be to ethical debate.

For many of the rest of us, however, this conclusion is an anathema. For me, as for Aristotle, Mill, Kant, and their intellectual descendants, there is more to ethical debate than just the particular (higher-order) emotional dispositions we each happen to have due to the (sub)cultures in which we have been socialized. Even if in a certain culture one is honor-bound to stone one's daughter for being raped, the fact that people hold that attitude, or even that they hold higher-order attitudes in favor of having such an attitude, just shows that they have the wrong attitudes, we want to say. We hold out hope that there are some answers to the ethical question of how a human being ought to live, answers that apply as generally as that question.

4.5 Conclusion

In this chapter, I have surveyed four leading approaches to ethical theorizing, noting in particular their points of convergence with and divergence from the proposal I had set out in the previous chapter. The positive proposal I detail there is in part a response to suggestions by modern Humean theorists that there are no answers to the question of how one ought to live, no answers that are true for any human being from her own point of view. I share with modern sentimentalist approaches both my empirical outlook and also the empirically-driven conclusion that the moral judgments we make are expressions of the emotional dispositions we have. With Hume, but against leading contemporary Humeans such as Jesse Prinz and Shaun Nichols, I suggest that there are some aspects of emotional motivations that are universal enough to make it the case that a circumscribed set of ethical claims are true for all human beings. Thus, although I share Aristotle's conviction that empirical, natural philosophy is deeply relevant to ethical questions, I have rejected the (neo-)Aristotelian approach exemplified by Phillipa Foot as identifying the wrong sort of empirical considerations, ones that lack the right sort of motivational force to adjudicate contested questions about how we ought to live. In a similar way, although my Buddhist-inspired approach shares with Utilitarian theories an emphasis on reducing suffering, I depart from the consequentialism of modern Utilitarians such as Joshua Greene and Peter Singer by taking first-personal considerations about how one's own emotional motivations feel to be the criteria upon which we can decide which sorts of emotions we ought to be motivated by. I have argued that some leading arguments for counter-intuitive Utilitarian conclusions not only fail to offer a means for deciding whether for instance the compassionate motivation to relieve aggregate suffering ought to trump other possible motivations, but moreover that the same arguments employed by Utilitarian theorists may tell against their own consequentialist assumptions. In this light, I have noted as a point of commonality between Kantian approaches and my own the emphasis on volition, rather than consequences, as being the primary ground of ethical evaluation. A second area of convergence is that both Kant and I see identifying certain universal aspects of the subjective grounds of ethical judgment as the means to solve the riddle of ethical universals. Because both the strengths and weaknesses I have

pointed to in the Kantian approach are of such a broad and central nature, I have assumed that I could address most parsimoniously modern Kantian theorists and others attracted to similar rationalist approaches by making my complaints in regard to Kant's own work. In particular, I have suggested three empirical reasons for resisting the rationalist impulse to ward off the specters of ethical relativism and moral skepticism by appealing to dictates of rationality. Instead, in Chapter 3 I have proposed a different sort of epistemological ideal that offers a more sturdy foundation. As I suggest in the next chapter, questions about the ground of ethical claims arise in the context of disagreements over grounding ethical values that arise between human cultures and sub-cultures, and also sometimes within oneself. Because the practical point of view inhabited by human beings involves justifying our decisions about how to live to ourselves and to others, it may be that to ground ethical claims in the human point of view requires that there be some answers that apply for any point of view that is humanly possible to inhabit.

Chapter 5

Who Cares? Metaethics from the Human Point of View

Aside from whatever promise or failings AWA may have, it can serve as an example of a more general strategy for opposing cultural relativism from a naturalistic perspective. The premise of this general approach is just that there is some shared psychological structure that puts enough of a constraint on human normative frameworks to provide a ground for adjudicating some of the ethical disputes that arise between cultures, individuals, and also between parts of ourselves. Call this Shared Human Contingency. In this chapter I illustrate the promise of using Shared Human Contingency as a supplement to naturalistic approaches to ethics. I focus on three such approaches in particular, the constructivist views of the variety Street (2010) calls Humean, the simple form of sentimentalism typified by Prinz's (2007) account, and the more sophisticated quasi-realist approaches of Gibbard and Blackburn. These approaches share an emphasis on the radical contingency of the particular framework of values one happens to inhabit and express in one's ethical judgments. But precisely because these approaches take one's starting normative commitments to be radically contingent, on any of them it could just so happen for evolutionary and psychological reasons that all the beings with whom we can in fact make and debate ethical claims share enough of their normative frameworks in common with one another to adjudicate

certain of these differences. In this way I suggest that distinguishing between those normative frameworks that are logically possible and the (proper) subset of these that are psychologically possible for human beings can resolve a number of recent debates in recent meta-ethics.

5.1 The Ugly Factual

On many influential approaches to ethical theorizing, ethical truths are taken to be independent of human responses. Often this involves the proposal that there are normative properties of act-types, say, that are independent of any emotional response or ethical judgment that anyone might have. This position is often called Realism. If cogent, a realist position can take for instance the wrongness of stoning people for being raped to be independent of how anyone might feel about it. This is an attractive result, but it comes at a significant cost. The Realist proposal that there are properties of to-be-doneness and not-to-be-doneness independent of any subjective responses meets with deep metaphysical and epistemological difficulties. Mackie (1977), famously, noted the "queerness" of the idea of properties of to-be-doneness and not-to-be-doneness independent of any subjective responses. A more recent and specific charge is that made by Sharon Street (2006). The rough idea of Street's argument that the evolutionary forces that have shaped the sentiments we have are not related in the right way to the sorts of evaluative truths that Realism posits. A large variety of these popular sentiments happen to be implanted in us, because of having been adaptive over the course of human evolution. But it is absurd to suppose that evolution just happened to implant in us those sentiments that get the ethical truth right. This is the problem of Unexplained Coincidence (Drier, 2012). In short, the problem for Realism is precisely its defining feature, that ethical truth is supposed to be independent of human responses.

Abandoning this Realist picture, many modern heirs of Hume have opted to bite the bullet of moral relativism. On Prinz's (2007) view, for instance, normative properties are constituted by real, empirical, properties of human beings' dispositions to moral emotions. If one individual or moral community holds honor killings of rape victims to be obligatory and another holds such

actions to be monstrous, each in a way that is coherent with the rest of their respective normative commitments, then for one community it is the right thing to do and for another it is wrong. This culturally relative approach to ethical evaluation can be applied to intentions and motivations as much as to actions and outcomes. Suppose that those of us with an Aristotelian or more generally Greco-Judaic cultural heritage are socialized to not only to feel angry about certain societal inequalities but also to feel good about being motivated by this anger to bring about change. If so, on this relativist line of thought, for us such righteous anger is the right way to be motivated. In contrast, if Buddhists and more generally East Asians are socialized to feel bad about anger in themselves and to disapprove of anger in others, then for them being motivated by anger is the wrong way to be. We are better off facing up to the fact that ethical truths are culturally relative in this way, Prinz would have it, than hiding our heads in the meta-ethical sand.

Humean versions of constructivism, of the sort favored by Street (2010; 2012), would at first glance seem to lead to much the same sort of cultural relativism as Prinz's Humean sentimentalism. As constructivism is often understood, this family of approaches holds the normative truth to be constituted - rather than discovered - by whatever principles would emerge from a procedure of ideal deliberation, for instance from behind the veil of ignorance proposed by Rawls. Rather than being characterized by a given procedure, however, Street thinks that constructivism is better characterized as taking the normative truth to be constituted by the values entailed from within a particular practical "point of view". Kantian versions of constructivism attempt to derive substantive moral conclusions from within the standpoint of any practical point of view as such. In contrast, the versions of metaethical constructivism that Street characterizes as Humean deny that one can step out of one's set of ethical commitments, as a whole, in order to establish that one set or another is more correct. Instead, Street's Humean constructivism takes one's starting commitments to be radically contingent, "such that had one come alive with an entirely different set of evaluative attitudes, or were mere causes to bring about a radical shift in those attitudes, one's reasons would have been, or would become, entirely different" (Street, 2010, 270). On this rough characterization, such a Humean constructivism shares much of its plausibility and also many of its

more unpalatable commitments with the explicitly relativist sentimentalism Prinz argues for. For on both views, the fact that we have attitudes against stoning people for being raped (or attitudes that entail attitudes against such actions) is precisely what makes such actions wrong. It is this feature that makes such views immune to the epistemic argument Street levels at Realism. Because approaches such as Street's and Prinz's take the direction of explanation to run from subjective attitudes to evaluative truths, on such views it is no coincidence that we (often) get it right when we judge stoning people for being raped to be wrong.

This position has its own deep costs, however. For if it is the sentiments we have that constitute the ethical truth(s), this implies that were our sentiments different, so too the ethical truth would be. Such a position entails what Dreier (2012) calls "Ugly Counterfactuals". He notes for instance,

> (UC) If our sensibilities were more selfish and individualistic, feeding the distant hungry would be wrong.

Such Ugly Counterfactuals are, as Dreier comments, "hard to believe" (2012, 272). The question is why they are hard to believe. My suggestion will be that certain facts about the starting set of evaluative commitments shared by all human beings, in virtue of facts about our evolutionary history and neurobiological makeup, make it the case that such counterfactuals are never factual in any human context. However, in the case of other Ugly Counterfactuals this shared human basis is less obvious, at best. Take for instance,

> (UC) If our sentiments were different than they are (in the relevant way), stoning people for being raped would be permissible.

Unpalatable as the many such Ugly Counterfactuals implied by Street's constructivism are, she thinks this is just a bullet we have to bite to avoid the metaphysical problems of response-independent realism. Prinz embraces the further implication that some Ugly Counterfactuals are factual, in the context of foreign cultural norms. In Western cultures, people are socialized to feel horrified by the the idea of stoning people for being raped. This is why such a counterfactual is ugly, for us. On Prinz's account, such emotional dispositions are the truth-makers for ethical claims. Since these

truth-makers vary between cultures and sub-cultures, so too do ethical truths. People who are socialized as in Western cultures, such that the idea of stoning people for being raped is abhorrent, are correct when they say that such actions are wrong. Equally, in other human cultures people are socialized to feel approval towards stoning people for being raped, or to feel ashamed if they don't act in this way. For them, according to Prinz's constructivist sentimentalism, stoning people for being raped is obligatory or at least permissable. The claim that such stonings are wrong, in the mouths people of with emotional dispositions of approbation toward such stonings, would be false. In this case, we have an Ugly Factual, or more precisely a factual claim that will seem ugly to modern liberal sensibilities.

> (UF) Where other people's sentiments are different from ours, stoning people for being raped is permissible (and even obligatory).

Just as this conditional claim will seem ugly from many modern liberal points of view, from an evaluative perspective in which the family's honor takes priority, not to give it that priority is what is unthinkable. So Prinz's theoretical account will imply a a factual claim that will seem equally ugly from such a traditional perspective.

> (UF) Where other people's sentiments are different from ours, what is necessary to preserve the family's honor is not permissible.

For now, my point is not to assess whether one of these evaluative perspectives has a better claim to ethical truth than the other, but rather to show that whatever evaluative perspective one operates from, Prinz's metaethical view implies the existence of some such Ugly Factuals. From any cultural perspective, there will seem to be a cost to Prinz's theory in that it implies certain Ugly Factuals, though which conditional implications will seem ugly will vary with one's cultural location.

For this reason, we can see this cost of mind-dependent accounts with the most force from the evaluative perspective we happen to inhabit. Modern liberals will want to say that even in cultures where people share the feeling that honor killings are necessary to preserve the honor of the victim's family, and so on, still they are wrong to feel that way. So we (modern liberals) want to say,

too, that even if our sentiments happened to be like the sentiments of people in such a culture, still stoning people for being raped would be wrong. Quasi-Realist theories such as Simon Blackburn's and Alan Gibbard's attempt to make space for us to go on saying such things, without thereby taking on the problematic Realist commitment to the existence of response-independent normative properties. These theories tell one or another story about the semantics of moral practices, aiming to vindicate the practices of talking as if there were response-independent moral properties without adopting the metaphysical commitment to such properties existing in this way. If such an approach works, it allows us the satisfying prospect of "earning our right", as Blackburn (e.g., 1998, 281) puts it, to reject Ugly Factuals of the sort noted above. That is, it allows us to assert what Dreier calls "independence counterfactuals". To continue with our earlier example,

> (IC) Even if we were to approve of stoning people for being raped, it would still not be permissible.

The motivation for Quasi-Realism is to vindicate such common sense convictions while avoiding the metaphysical problems associated with Realism. Nonetheless, Street (2011) holds that quasi-realists do not earn the right to independence counterfactuals because her argument against Realism applies also to Quasi-Realism. When the Realist asserts or implies support for IC, she is taking the wrongness of such stonings to be independent of how anyone feels about such things. If Realism is cogent, it could provide justification for saying that stoning people for being raped is wrong even in a human culture where people feel that it is right (depending, of course, on how it is the independent normative truth turns out to be). And IC is just the appealing sort of statement that Quasi-Realism is supposed to vindicate. But to the degree the Quasi-Realist asserts or implies support for the IC, Street reasons, he thereby commits himself to just the sort of independence that makes Realism problematic. For then he needs an account of why it is that the sentiments we happen to have so often happen also to accord with the ethical facts that his claims imply are independent of us.

Street imagines a case of ethical disagreement between two characters, Ann and Ben. Continuing the example we have been using, suppose Ann and Ben disagree as to whether honor killings

are obligatory or instead monstrous. Suppose Ann is in favor of them and Ben is opposed. In order to mimic common practice, Street rightly points out, the Quasi-Realist has to grant that Ann and Ben can each justifiably assert that there is some truth about whether honor killings are obligatory or instead monstrous that is independent of either of their own attitudes. They disagree about what that independent normative truth is, but they agree that there is one. The Quasi-Realist grants them this kind of claim, or at least allows them to earn it by understanding from the theoretical level what it is they are doing on the normative level when they assert the that there are independent truths, namely expressing certain affective attitudes. For Blackburn, when Ann and Ben agree that there is some normative truth about honor killings that is independent of either of their attitudes, they are each expressing their own strong current attitudes against the hypothetical case in which someone holds different attitudes than the respective attitudes they each currently hold. Thus when Ben imagines a case of someone (such as his future self) approving of attitudes approving of such stonings he will disprove, strongly. Hence he will say, now, that wrongness of honor killings is independent of how his future self feels about it. And conversely, Ann will say that the obligatoriness of honor killings is independent of how she or anyone else might feel about such things in the future, since imagining (now) her future self condemning honor killings, she will disapprove strongly. Similarly for Gibbard, when Ann and Ben agree that there is some normative truth about honor killings that is independent of their attitudes, they are each expressing a plan that has the same kind of form: they are each expressing a plan not to revise their plans even for the hypothetical situation in which they themselves hold accept different plans than they currently do. Street correctly points out that in order to successfully mimic common practice the Quasi-Realist has to grant this kind of agreement. As she puts it,

> To accommodate the idea that Ann and Ben agree, therefore, quasi-realists must admit as intelligible what I am calling "talk about the independent normative truth as such"—in other words, talk which presupposes nothing substantive about what the independent normative truths in question are. (Street 2011: 10-11)

While I agree that the quasi-realist must grant that Ann and Ben agree that there is some truth about

whether honor killings are monstrous or instead obligatory, nonetheless I think there is reason to be wary of Street's locution, "the independent normative truth as such". What Ann and Ben agree on is that there is some truth about whether honor killings are obligatory or instead monstrous, and that whatever this truth is, it holds regardless of how any individual human beings happen to feel about the practice. Suppose they disagree on other matters as well, for instance abortion, the death penalty, and nuclear arms, but that in each of these cases they again agree with one another that there is some normative truth about these matters that holds regardless of how any individual human beings happen to feel about these things. The question is how the Quasi-Realist is to accommodate this general kind of agreement. Both of the Quasi-Realist strategies that I have outlined for accommodating the agreement between Ann and Ben to the effect that there are truths that hold independent of anyone's attitudes about these things, do indeed presuppose nothing substantive about what those normative truths are. On a Realist picture what allows for the agreement between Ann and Ben is the supposition that there are properties out there independent of human normative frameworks entirely. And if the Quasi-Realist supposes this, then Street is right, he has indeed avoided the Ugly Factual only by appealing to something that could rightly be called "the independent normative truth as such". And if so, the Quasi-Realist is indeed subject to the same problem that Street correctly points out in Realist views. It is important to remind ourselves, however, that the Quasi-Realist project is not intended to save the metaphysics of Realist ethical theories. Rather the idea is to save common practice. And so in assessing whether or not Quasi-Realists are committed to "the independent normative truth as such" it is worth noting, if only as anecdotal evidence, that one rarely hears anyone except metaethicists agreeing or disagreeing in those terms.

Dreier (2012) presents an interesting argument to the effect *that* quasi-realism is not saddled with Realism's problem of Unexplained Coincidence. He uses an inductive strategy of proceeding through various stages of developing a Quasi-Realist semantics, arguing at each stage that the problem Street raises has not yet entered. What the argument does not show, Dreier recognizes, is precisely *why* the problem of Unexplained Coincidence is not a problem for Quasi-Realists. In

the next section, I propose an answer to this question, and also suggest a way in which Street's Humean constructivism and Quasi-Realist approaches can come together.

5.2 Humanly Possible Normative Frameworks

In my argument for AWA in Chapter 3 above, I have noted some points of inspiration from Gibbard's expressivist analysis. I also noted there that unlike AWA, Gibbard's view offers no way to adjudicate differences between radically opposed but internally coherent (hyper-decided) sets of plans. For this reason, the resulting meta-ethical view, like Blackburn's, offers no way adjudicate radically different but internally coherent sets of ethical judgments. Notice that for Blackburn if we differ in how we feel (about how we feel) about stoning people for being raped, then we will say different things about whether such actions are right or wrong. For the reasons reviewed above, Blackburn will be entitled to say at any given time that stoning people is wrong and would be even if he himself felt differently than he does. But equally, if I approve of such stoning I will be entitled to say that stoning people is right or obligatory, and would be even if I myself held a wholly different set of attitudes than I do. Quasi-Realism offers no way to adjudicate the dispute between us.

In essence, it is because Quasi-Realist views do not on their own offer any way to adjudicate such ethical disputes over fundamental values that I take them to be inadequate, on their own at least, as a response to the practical problems that motivate meta-ethical inquiry.[1] However, it is for this same reason that they can be formulated so as to escape the commitment to normative truth as such that Street saddles them with. Indeed, although Street (2012) and Blackburn (1984) both allow the constructivist sort of Neurathian work towards internal coherence, taken on their own neither of these views offers any way to adjudicate between radically different but internally coherent normative frameworks. These views do not even allow us to get the normative truth

[1] Street (2010, 377) seems to make a similar point, suggesting that the constructivist might "take on board the expressivist's account of the meaning of normative terms, but then... argue that this account taken by itself fails to answer the traditional questions of metaethics."

wrong, in this radical, global sense, much less to get it right.

Blackburn (1984, 197) himself raises the worry whether the Quasi-Realist "musn't in some sense have a schizoid attitude to his own moral commitments - holding them, but also holding that they are ungrounded". I return to this point below. Blackburn's response focuses instead on the worry that such a theoretical stance leads from metaethical relativism to amoralism, that is to the morals of a French gangster and the like. Interestingly, Blackburn takes it that asserting the radical contingency of normative frameworks in general is harmless given the contingent but shared psychological constraints on which sorts normative frameworks human beings are able to occupy. "Just as the senses constrain what we can believe about the empirical world, so our natures and desires, needs and pleasures, constrain much of what we can admire and commend, tolerate and work for" (197). This is just the sort of shared human contingency that I suggest can be used to adjudicate a certain circumscribed set of ethical claims, within the human context. To the degree Blackburn adopts such a supplement to the mainline of his Quasi-Realism, I think he may succeed in satisfying the aspiration of common practice to adjudicate disputes over fundamental values, even while escaping the problems that Street raises for Realist suppositions about the normative truth as such.

It is Gibbard's view that Street focuses on attacking, and perhaps she is right that some aspects of his view do commit him to these problems. Nonetheless, Street distinguishes Gibbard's view from an "alternative quasi-realist proposal" that she describes only at the end of her argument against Quasi-Realism (Street, 2011, 20ff). This alternative view differs from Street's reading of Gibbard in that it does not admit as intelligible questions about the normative truth as such. The Quasi-Realist analyzes the debate between Ann and Ben, and the agreement between them that there are some ethical truths independent of anyone's attitudes, in the way indicated above, that in agreeing in this way they are both expressing attitudes against any future change of attitudes they might have. The alternative Quasi-Realist then goes on to say that questions about whether one is a reliable judge of these normative truths that are independent of anyone's attitudes can only be understood from within a normative perspective that takes substantive positions on what these

mind-independent normative truths are. It is this alternative form of Quasi-Realism that I take to be most plausible and interesting, and also to be in line with Blackburn's view as I have characterized it so far.

Blackburn himself takes his own way of putting these things to be almost wholly interchangeable with Gibbard's, and also takes his own view to reject talk of the normative truth "as such" (Blackburn, unpublished). Street contends that the "alternative quasi-realist proposal" is nonetheless committed to some notion close enough to the notion of "the indendent normative truth as such" that it is subject to the dilemma she raises for Realism. The thrust of her argument here as elsewhere is to raise skeptical worries regarding how basic human normative commitments to the value of survival and so on could track the independent normative truth that the Quasi-Realist gives us license to talk about. She extends this line of argument by appealing to a thought experiment involving humans beings and also another set of creatures much like us except with a very different evolutionary history. It turns out that the Quasi-Realist account breaks down into absurdity just the point when we and these other beings get to trading skeptical doubts about each others' ability to track the normative truth.

If this aspect of Street's argument is cogent it shows that a Quasi-Realist analysis of ethical disagreement breaks down when applied to contexts outside of a human point of view. But if the human practices of making and debating ethical claims depend implicitly on shared features of the human point of view, then this is precisely the point at which these practices *should* break down into incoherence. And if Quasi-Realism is offered as an analysis only of these human practices, then the point where we attempt to adjudicate ethical disputes from outside of this human point of view is just where Quasi-Realism *should* break down. For as Street and I agree, there are multiple (perhaps a great many) logically possible normative frameworks that are internally coherent but radically opposed to one another. What makes some of these humanly possible and not others are the details of human psychology (see especially Street, 2009). We cannot trust our human intuitions to deliver reliable results outside of such a context, precisely because this human context is normally implicit in our practices of ethical judgment. Given this kind of approach, the only

metaphysically respectable criterion by which to judge that any of us ever get the normative truth wrong, or right, is to appeal to attitudes that are shared by all those with whom we can make and debate ethical claims, in virtue of our shared human neurobiology. So suppose that human practices of making and debating ethical claims do depend in this way on an implicit assumption of a human context. To make the case more concrete, take the central suggestion of the AWA account, that any individual human being can get their attitudes wrong relative to the attitudes they themselves would have if they were fully and accurate aware. This way of assessing when an individual or group gets the normative truth wrong depends on the appeal to the idea that all human beings have shared reasons to have certain attitudes and not others, in virtue of the shared neurobiology of our emotional motivations. So suppose that human practices of making ethical judgments implicitly assume in this way that the neurobiological makeup of human emotions is similar enough that hatred would feel bad for any human being, and friendliness would feel good, as I have suggested in Chapter 2. In this case, when we make the claim that honor killings are wrong, this is intended to apply from any human point of view, and only from human points of view. The Humean constructivist can agree (or rather, *should* agree, in my view) that in relation to an implicit assumption that normative truths are *dependent* on shared human attitudes, and only in the context of this implicit assumption, the *independence* of normative truths from the attitudes held by particular individuals and human groups can be explicitly asserted.

The sort of global skeptical worry that Street raises for Quasi-Realist accounts assumes that human ethical judgments aspire to apply outside of the human context. To bring this out more clearly, note that if ethical judgments about getting the normative truth right were instead to implicitly assume a human psychological context, then it would not be intelligible to worry as Street does about whether the evolved human attitudes we can't help but share get the normative truth hopelessly wrong. And given that the practices of making, debating, and justifying ethical judgments that meta-ethical theories attempt to account for were developed among human beings, and continue to be practiced solely with other human beings, there is no good reason to suppose that ethical claims should be taken to apply outside of this context. It is true that one of the most robust

characteristics of moral judgments is that people take them to generalize. In the classic study by Tisak and Turiel (1984), children were more likely to say that moral norms applied "in another city" than prudential ones. Moving slightly more broadly afield, Nichols and Folds-Bennett (2003) show that children tend to take harm and disgust violations as wrong also "in another country or someplace far away." On the one hand, such evidence is consistent for instance with Kant's claim that moral claims apply not only to human beings, but rather to rational beings as such. On the other hand, it is equally consistent with the suggestion that the human practice of moral judgement emerged to deal with human situations, and does so without making any substantive claims about remote possible worlds. Moreover, some recent evidence suggests that objectivist intuitions about moral properties depend on implicit assumptions about shared subjective values: Sarkissian et al. (2010) gave vignettes to undergraduates in South Carolina and in Singapore in which two individuals diverged in their judgments about whether a certain act was wrong. In cases where the individuals in the vignette making the moral judgments came from radically different cultures, respondents were less willing to say that one party or the other had to be mistaken. They were even less willing, though marginally, to say that one party had to be mistaken in cases where one of the judges was an extraterrestrial with "a very different sort of psychology from human beings". In regard to the focus of AWA on the ethical evaluation of intention, it is notable that Sarkissian et al. used vignettes of actions with harmful outcomes and did not distinguish between accidental and intentional actions. Thus it might well be that the truth of judgments of actions, as morally wrong or morally permissible, is taken to be more culturally relative than the truth of judgments of intention. Nonetheless, the results do seem to suggest that people use moral concepts in ways that do not require that moral truths hold for all rational beings as such. At a minimum such evidence does seem to put the burden of proof on those who would claim that moral judgments somehow aspire to be independent of the human point of view entirely.

If human moral practices cannot be shown to be committed to truths that are independent of the human point of view, as I suspect, then neither are Quasi-Realists committed to offering an account that is so committed. Our common practices of moral judgment and dispute may still embody an

ambition to able to adjudicate disputes between radically different systems of *human* values. An approach such as AWA, if it is cogent, vindicates this ambition. Quasi-Realism, on its own, would seem not to. After all, the Quasi-Realist analysis leaves the sets of attitudes or plans we happen to inhabit and express as radically contingent. It is notable, however, that in responding to Street, Blackburn appeals to just the kind of constraints on which normative frameworks are possible for human beings that I would like also to appeal to.

> I am a little acquainted with both misery and happiness, and so I suspect are you: the former strikes me as worse than the latter—how does it strike you? Do you choose, recommend, desire and promote misery above happiness? I doubt it. I do not, therefore, have to justify my own ranking to you, for you share it. We are both of the party of mankind, and that is the only audience it is worth engaging in questions of moral justification. (Blackburn, unpublished)

It is important to note that in appealing to such shared human sensibilities, Blackburn is offering not an implied extension of his Quasi-Realist story, but rather a supplement of a different sort. In particular, Blackburn seems to be pointing to pan-human grounds that would constrain which *first*-order ethical attitudes are tenable for human beings. It is also worth pointing out that making such an appeal, the cross-cultural invariance of ease and unease may prove a more robust basis than culturally variable notions of happiness.

More centrally for my purposes here, the point should not be just that mankind is the only audience *worth* engaging with on questions of moral justification. This would seem to imply that we could in some sense engage in ethical debate outside of the human context, it's just not worth doing. I want to suggest instead that our practices of making and debating ethical claims may only be intelligible in the implied context of shared human sensibilities of the sort Blackburn points to. Note that we can deny the Ugly Counterfactual in two ways. One way is to say that stoning people for being raped is wrong across all possible worlds. A different way is to say that the presuppositions of ethical claims being intelligible depend on a particular pragmatic context. Some Ugly Counterfactuals may be ugly from any human point of view, but equally, outside of

that context our practices of making and debating ethical claims may not apply.

In a sense, the distinction I am making between logically possible normative frameworks and humanly possible ones simply transposes the Quasi-Realist distinction between the metaethical and first-order normative perspectives. Perhaps in the context of making and debating ethical claims with other human beings, our claims implicitly assume a certain shared subjective standpoint, a common set of starting commitments. If so, then by assuming such a context the other sorts of normative frameworks that are logically possible are not a relevant comparison class. When you overhear people at the next table debating about how best to relate to their ex-spouses, the question of how things would be if we all had radically different evolved sentiments were is not on the table. In contrast, to consider the nature of normative discourse in metaethical theorizing may constitutively involve making salient logically possible normative frameworks other than our own. Employing such a distinction, the Quasi-Realist approach can be refined to suggest that to assert the mind-independence of a normative truth from within a first-order normative perspective is only to assert that truth's independence from the minds of any individual or group of human beings, not to assert its independence from human attitudes in general. Correspondingly, from a theoretical perspective that makes salient all logically possible normative perspectives, and not just humanly possible ones, it is false to assert that normative truths are mind-independent. For, made in such a context in such a way, this assertion would imply just the sort of metaphysical and epistemological problems that Street raises for Realism. In response to the original Quasi-Realist proposal to avoid the problems of realism by distinguishing between normative and theoretical perspectives, Street (2010, 378) suggests that "the trouble is... that the expressivists themselves have developed resources rich enough to suggest how these worries may be recast as substantive normative worries." This is correct so far as it goes. However, if the analysis proposed above is correct, to raise the epistemological and metaphysical worries about independent normative truths as substantive normative worries is precisely to raise these in a pragmatic context in which shared human sensibilities are *implicitly* assumed. In such a context, it is perfectly legitimate for Ann and Ben each to assert that there are mind-independent normative truths, and even to assert that this at least is

a point of agreement between them. Understood as a human exchange made in the context of an assumed shared set of starting commitments, the claim that Ann and Ben agree on just amounts to the claim that there is some set of conclusions that are entailed by the starting commitments any human being has in virtue of being human. Made in such a context in such a way, Ann and Ben's agreement implies that there is some set of normative conclusions that hold for all human beings in virtue of the starting commitments we share, and are thus independent of what any individual human being or human group might mistakenly think their starting commitments are or entail. But this kind of independence, of normative truths from any particular human perspective, need not imply any kind of independence of the normative truth from the human point of view in general. Realist theorists of metaethics want to provide a means to adjudicate ethical disputes in remote possible worlds where the parties to the dispute do not share human psychological universals. I take it that normative truth "as such" is proposed as the sort of thing that might do the trick. But precisely because common practice need not adjudicate such disputes, Quasi-Realists need not either. Put in terms of Street's epistemological challenge, the point is that in the context of human ethical disputes, the relevant epistemic worry is just that an individual or group of human beings might have got the normative truth hopelessly wrong *in relation to the normative starting commitments we can't help but share as human beings*. To admit as intelligible within normative discourse between human beings a kind of skepticism that is local to this human context is not to admit as intelligible the global kind of skepticism Street raises. In particular, understanding our shared contingent human psychology to be an implicit assumption of our practices of making and debating ethical claims renders problematic the very question of whether the set of starting commitments human beings can't help but share might be hopelessly wrong. If the standards of correctness that make such epistemic worries intelligible are those given to us by in virtue of constraints imposed by contingent facts about our shared human neurobiology, then there is no reason to expect that these epistemic worries will remain intelligible outside of that contingent context, across the set of all logically possible normative frameworks. Here the right move for the Quasi-Realist to make is one in line with Street's own Humean constructivism: there is simply no place outside of a human

set of values that we can occupy in order to judge our whole shared set of values as human beings as more or less accurate than some other logically possible set of normative values. Put another way, normative truths that are independent of any particular human mind are good enough to adjudicate the only kinds of ethical disputes we actually get into; we don't need mind-independent normative truth as such. And that's a good thing, as Blackburn, Street, and I agree, since we can't have it.

5.3 The Response-Dependence Ratio

Drawing this distinction between independence from the attitudes of particular human beings and independence from human attitudes in general, as I think Street should, thus helps us to see a way forward between the horns of the dilemma she poses for Quasi-Realists. Nonetheless, this is a strategy that is available to Humean constructivists and Quasi-Realists alike. Indeed, there is more in common between the Humean approaches of Street and Blackburn than might appear at first. Combined or separately, my suggestion is that both Quasi-Realism and also Humean constructivism can and should adopt the adjunct thesis I have called Shared Humean Contingency. Shared Human Contingency, recall, is the hypothesis that the neurobiological structures shared by all human beings, contingent as they are, nonetheless impose constraints on which logically possible normative frameworks are also humanly possible. In particular, the idea is that the biology we are all born into imposes constraints sufficient to make all human normative frameworks similar enough that certain sorts of ethical disputes can be settled within this human, though not outside of it.

The account of AWA offered in Chapter 3 can be seen as one tentative way of cashing out the implications of Shared Human Contingency. On that proposal, a behavior that is motivated by ill-will is one that ought not to be done (if it is), just in virtue of the fact that none of us ought to have ill-will. And what makes it the case that we ought not to be motivated by ill-will is that we ourselves would prefer not to be motivated in this way, to the degree we were Wide

Awake. AWA is also committed to the further claim that if we ourselves would prefer not to be motivated by ill-will to the degree we were Wide Awake, that is so just in virtue of the fact that the neurophysiology involved in being motivated by ill-will has a predominantly negative affective valence for any human being. As I noted in Chapter 3, however, this last claim can be taken in two importantly different ways.

On the one hand we can take the "in virtue of" in the sense of justification. On this reading, AWA amounts to a form of ethical egoism: the *reason* that we should act to benefit others is that an altruistic motivation is more pleasant or less unpleasant than the relevant alternative sorts of motivation, and the reason that we should not act out of ill-will is that such a Quality of Heart is characterized by greater unease than the relevant alternatives. Taken this way, AWA proposes a naturalistic reduction of ethical truths to psychological ones. The account given in Chapters 2 and 3 assumes that the neurophysiology of at least some Qualities of Heart are distinct enough from each other and common enough among human beings; given this, all of us are disposed to have negative affective reactions towards certain Qualities of Heart, at least unconsciously; secondly, we are disposed to express the attitudes we do consciously feel in our ethical judgments; and thirdly, we are disposed to avoid the cognitive conflict inherent in failing to be fully conscious and accurately aware of these motivational forces. Putting these together, the idea is that the psychological structures we share push us toward a certain common ethical stance. In a sense, then, the claim is that we are disposed to be disposed to approve of certain Qualities of Heart and disapprove of others. This psychological claim can be used as a justification for ethical ones in various ways, depending on one's commitments regarding the semantics of moral terms. Adopting something like Jesse Prinz's (2007) account of the semantics of moral terms, for instance, we could then say that because these affective dispositions are common to all human beings, they can serve as the truth-makers for a certain circumscribed set of ethical claims that come out true in any human culture.

Alternatively, we can and perhaps should take the project that Prinz and (especially) Blackburn are engaged in when they appeal to affective attitudes not as providing *justification* for (some of)

the ethical claims we make, but rather as providing an *explanation* of (all of) the ethical claims we make. My claim above that certain Qualities of Heart are preferred by those who are Wide Awake in virtue of the fact that the neurophysiology involved in being motivated by ill-will has a predominantly negative affective valence for any human being can also be read in this way, as explanatory rather than as justificatory. On this reading, the fact that a motivation of ill-will is more pleasant or less unpleasant than the relevant alternative sorts of motivation serves just as a psychological *explanation* of the fact that we ourselves would not want to be motivated by ill-will to the degree we were Wide Awake. As I argued in Chapter 2, for instance, it might just turn out that whatever attitudes we have towards a particular Quality of Heart are expressed in our ethical judgments, and that to the degree people are Wide Awake their preferences and so their judgments on such matters converge. On this second reading, AWA does not appeal to the hedonic valence of ill-will as a *justification* for ethical disapproval. The appeal here in justifying the ethical claim is not to a natural facts about affective attitudes but instead to other normative considerations, for instance to epistemic norms of accuracy and full awareness, but also and perhaps more fundamentally to a norm in favor of being wholehearted. The appeal to epistemic norms here is closely akin to the move made by Valerie Tiberius' (2012) Wise Judgment Constructivism. On such an approach, as with constructivist approaches in general, the justification for ethical claims stays within the realm of the normative. Unlike more traditional Kantian constructivism, Tiberius' approach and my own do not attempt to derive substantive normative claims from the formal features of the practical point of view as such. Ours are species of the genus that Sharon Street (2010; 2012) has termed "Humean constructivism", and take the normative frameworks we happen to inhabit as radically contingent, subject not to formal constraints but just to constraint by the set of starting commitments one happens to find oneself with. This second reading of AWA's foundational claim as explanatory rather than justificatory can be usefully described as a species of this genus, taking the normative frameworks that human beings inhabit as radically contingent but nonetheless holding that as a contingent fact the neurobiological profiles of human emotional motivations puts shared constraints on which normative frameworks we are able to occupy.

There are important continuities here with the more general field of constructivist approaches. On this construal of AWA, when we offer a justification for an ethical claim we are and can only be appealing to evaluative truths that are constituted in part by who we are. For this very reason, such justification stories will not and need not have appeal outside of the community of beings who are like us in the relevant ways. But this point can also be applied against certain constructivist proposals. Rawls, for instance, suggests that certain principles for living, while logically consistent, can nonetheless be be dismissed on the grounds that they "incompatible with what we intuitively regard as the moral point of view" (Rawls, 1999, 117). While this may work as a strategy for justification within a certain cultural conception of justice or morality, when extended as a strategy for justifying claims that are controversial in the broader community of human beings as for instance Steven Ross (2004) proposes, the strategy fails. Like Ross, AWA attempts to locate a modest, circumscribed set of claims about how one ought to live that can be justified for any human being. To return to our earlier example, one question at issue is whether honor killings can be justified from any human point of view, or not. Now, the justification story in favor of honor killings will be embedded with a comprehensive evaluative outlook, just as will a modern liberal justification for the claim that honor killings are monstrous. The question at issue is whether there is any way to adjudicate between these two evaluative outlooks, and so it does not settle the question to suggest that (some necessary parts of) the justification stories in favor of honor killings do not fall within 'our' modern liberal conception of morality, any more than the converse claim on 'their' part would settle the question. Nonetheless, my contention is that the neurobiology we are born into does impose substantive constraints on which sets of commitments we can start from, and end up with. For instance, I think a human being cannot wholly avoid feeling that it is better to live in a way that is wholehearted rather than half-hearted, and for this reason cannot avoid the similarly normative constraint in favor of deferring to the judgments about how to live we ourselves make when we are Wide Awake. More generally, the point is that because such psychological facts about us as human beings have implications for how we respond subjectively, certain claims about how we ought to live and to adjudicate questions about how we ought to live will appear more plausible

to any human being, and others will be appear less plausible, more vulnerable to criticism. But again, the empirical facts about us that are causally responsible for our subjective experience being the way it is need not figure in our normal justification stories for ethical claims. Rather, the point is that these facts constrain which points of view the justification stories we offer need to be plausible for. Indeed, it is precisely because these structures of subjective experience can be implicitly assumed to be in force on both sides of any debate over fundamental values between human beings that we don't need to incorporate them as part of the justification story for an ethical claim.

I noted above that when pressed Blackburn does seem to move towards such an approach. In particular, he seems to adopt the sort of move I recommend, suggesting that we don't need to adjudicate ethical debates except with those who begin from a contingently human perspective, and that within such a perspective there is enough commonality to make it the case that things such as honor killings might indeed be monstrous, for all of us. This sort of move is available to Street as much as to Blackburn. Like Prinz's and Blackburn's approaches, Street's articulation of Humean constructivism to date has emphasized the radical contingency of the set of values, as a whole, that any valuer happens to be born and nurtured into. In emphasizing this aspect to the neglect of what is entailed by a specifically human point of view, Street fails in the same way that Quasi-Realists do to vindicate any ambition that might be embodied in our practices of moral judgment and dispute to adjudicate disputes between radically different sets of human values. However, although neither Humean constructivism or Humean Quasi-Realism on their own offer any substantive grounds from on which to to adjudicate the sorts of debates that motivate meta-ethical inquiry, they can each be supplemented by considerations that allow even the theorist, as theorist, to hold that there is enough common ground among humanly possible normative perspectives to adjudicate certain ethical debates among human beings. Defending this avenue for adjudicating ethical disputes requires clearly distinguishing this Humean strategy from its less plausible Kantian cousin, which attempts to adjudicate hypothetical ethical debates between all logically possible normative perspectives. But it is a strategy that can be adopted by Humean constructivism as much as by Humean Quasi-Realism.

Seen in this light, moreover, it may be that an approach such as Blackburn's is not in the end in conflict with Street's constructivist view. Rather the two can be seen as engaged in different and even complimentary projects. Metaethical constructivism sets itself apart from naturalistic reductions of evaluative truth to natural facts about the responses of agents under certain idealized (but naturalistically described) circumstances (Street, 2010). Thus Street denies the sort of move that Prinz (2007) makes, of taking psychological attitudes as truth-makers for moral judgments. But Blackburn's approach also denies this kind of truth-making relation. The Quasi-Realist does give an naturalistic reduction of evaluative claims to natural facts about the expression of psychological attitudes. But the Quasi-Realist keeps this reduction isolated to the theoretical perspective; when engaging with normative questions even (or especially) the theorist of Quasi-Realism is entitled to express his attitudes in the form of ethical claims. I take it that one consequence of keeping these two perspectives separate in the way the Quasi-Realist tries to is that natural facts about psychological attitudes and the like cannot are not to be used as justification for normative claims. This is just what Street wants to insist on as well, that normative claims can be assessed only within the normative perspective, and not from without. Moreover, there have been persistent questions as to whether constructivism can succeed as a metaethical view in the way Street wants it to (see e.g., Bratman, 2012; Ridge, 2012). Seen in this light, it looks as if Humean constructivism offers the sort of epistemology of value a Quasi-Realist must use when weighing which normative attitudes to hold on the basis of other normative attitudes he holds, staying as he must within the internal normative perspective. And on the other hand, it also looks as if Quasi-Realism in turn outlines a plausible account from the external, empirical perspective of what is going on psychologically when one expresses the normative commitments one reasons on the basis of and arrives at, using the sort of epistemology described by the constructivist. So one way that Quasi-Realism and Humean constructivism can come together is as two complimentary approaches both grounded in a naturalistic perspective that takes evaluative properties to be different in kind from the empirical or factual properties that are the object of natural science (cf. Lenman, 2012). Agreeing on this distinction, they carve up the terrain accordingly, Humean constructivism offering an account of how

it is we go about reasoning about reasons from within the normative perspective, Quasi-Realism offering from the theorists' non-normative, scientific perspective (the outlines of) of what goes on when we adopt the normative perspective.

This distinction, between fact and value, is a venerable and widely assumed premise in recent moral philosophy in the Analytic tradition. Water often serves as a paradigmatic example of a natural kind, since the properties of water seem to be determined by chemistry, independently of human interests and attitudes. Answers to questions about which things are accurately classified as morally admirable or as beautiful would seem importantly different. In developing an empirically-based approach to ethical questions about which Qualities of Heart we should cultivate, I have relied above on a form of this distinction employed by a number of recent Humean theorists. On this approach, evaluative properties are response-dependent, whereas natural properties are response-*in*dependent. The property of a Quality of Heart such as hatred, that it should not be acted out of, is constituted in part by human reactions to that Quality of Heart, in particular that such an emotional motivation would feel bad to any human being so motivated, to the degree she was Wide Awake. In contrast, in explaining why we - any of us - would make such judgments I appealed to natural properties, objective measures of alertness and affective bias, as well as the neurophysiological profile of (something like) hatred. As such, the structure of my argument for a circumscribed set of ethical claims that are true for all human beings, from their own perspective, depended on grounding the subjective evaluative properties in objective natural properties.

To assess the merits of this distinction between evaluative and natural properties would involve a detailed and far-ranging examination of truth and reference, which I cannot attempt here. Nonetheless, in closing I want to examine briefly how AWA might fare what Hillary Putnam calls "the collapse of the fact/value dichotomy". The intellectual trajectory of Jesse Prinz, whose naturalistic form of realism about response-dependent normative properties I have mentioned throughout, is instructive in this regard. Prinz's (2007) sentimentalist account of moral properties is founded on his account of emotions (Prinz, 2004), and in turn on his account of perception (Prinz, 2001). All three of these interesting proposals appeal explicitly to the naturalistic account of ref-

erence developed by Dretske (1981) in terms of reliable covariance. And though Prinz's various accounts extend this basis into some un-Dretskian directions, nonetheless the viability of all of these accounts require at a minimum that some such naturalistic approach will yield an account on which the determinants of mental contents such as RED, ANGER, and WRONG, can be fully specified using non-psychological, non-semantic vocabulary. The projects of making sense of perception, emotion, and morality from a naturalistic perspective that Prinz has undertaken are pinned on this hope. What is interesting, however, is that Prinz has more recently come to see this hope as a pious one, and to reject the project of finding such a naturalistic account of reference (Prinz, 2013). In doing so, Prinz follows similar shifts in the intellectual trajectories of philosophers such as Putnam and Wittgenstein. Prinz's developing position is as yet not very systematic, and it is unclear what precisely will remain of his empiricist account of perception, or of his moral relativism, when the implications of this shift are thought through. The suggestion is at any rate not an idealist one; there are still features of the world that constrain which perceptual judgments we make. Nonetheless, Prinz suggests that there is often, perhaps always, a level of "choice" involved. In his earlier incarnation Putnam famously suggested an externalist account of reference on which the term "water" referred to just that substance classified as of 1750 by reference to the chemical composition H_2O. It turns out, however, we don't use the concept of water in just that way. Peoples' judgments about whether substances such as tea and sewer water belong in the category of water are not predicted by their estimations of the percentage of H_2O present. Location, source, and function seem to figure in folk categorization in addition to chemical essence (Malt, 1994).

It is perhaps misleading to put this in terms of "choice", as Prinz does, since the use of many public concepts, including water, is heavily constrained by inherited practices, and moreover by human interests some of which may be hardwired biologically. Nonetheless, the more general point is just that natural properties are partly dependent on human responses (Prinz 2013 explicitly makes the analogy to his sentimentalist account of moral properties). If evaluative properties have a high ratio of response-independence to response-dependence, we could say, still in the case of natural properties the denominator is not zero. The world we inhabit is one selected from

many nomologically possible lived worlds. Aspects of us, in particular our habits of attention and perception, are responsible for grouping certain features of the natural world together rather than others. In one sense, this aspect of the proposal brings us full circle. I began in Chapter 3 by suggesting that diversity in moral judgments can be explained by diversity in affective biases of attention and memory. The heart of my proposal was that to the degree human beings attenuate such biases they will come to agree with one another on a certain circumscribed set of ethical claims, in particular regarding which sorts of emotional motivations we ought to act out of. Since as individuals we can indeed make a choice about whether to sustain mental habits of affective biases or instead to cultivate mental habits of attenuating these, we can make choices that affect - if only indirectly - which motivations and actions we group together as good ones, and which we group as bad.

Drawing on Putnam's later incarnation, however, we can see that there may be a deeper challenge to AWA lurking in the rejection of naturalistic accounts of reference, but also an answer to this challenge. It was crucial to my proposal for grounding a certain measure of objectivity that the world pushes back, in the form of our human nature. In particular I have suggested that the hard-wired interest in avoiding unease, in combination with certain neurophysiological features of emotional motivation shared among all human beings, makes it the case that we are not able to plan for a life motivated only by hatred, as such, in the same way as we could plan for a life motivated only by love. On this strategy, the objectivity of (even a circumscribed set of) ethical truths is secured ultimately by appeal to objective empirical facts about the shared animal motivation system, neurobiological profiles of emotional states, and affective biases of attention and memory. In other words, the objectivity of ethics is to be secured by the objectivity of biological and psychological properties. But this appeal to would seem in vain, if the fact/value dichotomy should be rejected. I appealed for instance to an intuition that the neurobiological profile of (something like) hatred differs enough from the profile of (something like) brotherly love or friendliness, for any human being, that to the degree any of us were fully and accurately aware of these emotional states, we would prefer the one over the other. But the process by which certain clusters of neurobiologi-

cal properties are delineated may themselves be value-laden. Drawing on Dewey, Putnam (2002) suggests that fact and value are inextricably entangled in this way. On the one hand, this leads Putnam to rejects proposals such as Richard Rorty's, that we give up on objectivity entirely, and settle instead for "solidarity". This is an unstable position. As Putnam puts it "the very notion of solidarity requires commonsense realism about the objective existence with the people one is in 'solidarity' with" (Putnam 2002, 100). Putnam also rejects Bernard Williams' way of arguing that resolutions to ethical questions can be resolutions only from our particular cultural perspective. Williams' suggestion that ethical claims can be true relevant to only to some social world makes sense, Putnam holds, only as a contrast to a problematic notion of it *the* objective view (or social and other aspects) of reality as it would be conceived of by a perfected physics. But judgments of scientific truth rely on epistemic values, of the coherence or beauty or simplicity of a theory. If psychological facts are subjectively determined in this way, they will will not secure ethical objectivity, but equally they cannot be used to ground an objective claim that ethical expressions are (merely) subjectively true. Both of these critiques would apply to Prinz's (2007) view, and could in a different way apply to the proposal I have developed here.

Putnam suggests that both the scientific materialism of Williams and the disappointment in this project that leads to Rorty's proposal result from asking too much of truth, and that we can settle for a kind of commonsense realism about ethical truth as about scientific truth. Close to the heart of Putnam's critique of the fact/value dichotomy and also his solution to it is a pragmatist suggestion, that experience is never neutral, that an infant begins from experiences of food and drink and cuddling *as good*, and of pain and deprivation and loneliness *as bad*, that "as our experiences multiply and become more sophisticated, the tinges and shades of value also multiply and become more sophisticated" (Putnam 2002, 103). This same point could be made for cognitive conflict; my proposal in Chapter 3 was founded in part on the idea that we begin from an experience of internal conflict as bad and being wholehearted as good, and that whatever convoluted and conflicted psychological position we end up in, from masochism to outright endorsement of evil, nonetheless carries with it the seeds of this basic, and shared, value in favor of being wholehearted.

This gives us an account of what is valued, but not what is valuable. What distinguishes the valued from the valuable, on Putnam's reading of Dewey, is that the valuations of human beings are subject to criticism by ourselves and by others, and to refinement in light of this criticism. I cannot address the details of this proposal here. What I do want to point out is that such an approach may lead in the end to a riff on the same theme I have been suggesting could supplement meta-ethical approaches such as Blackburn's and Street's. Here we might note that for all practical purposes, the criticism to which our valuations are subject is always criticism by other human beings. There is a deeper worry here, though, if there is no determinate way to refer to human beings as a natural kind. With the advent of the technology capable of achieving transhumanist goals of enhancing human psychological functioning, for instance, it might become less clear which sorts of beings' criticisms should count. Nonetheless, the force of the worry can be dulled somewhat if what matters is the degree of similarity to those being criticized. Empirical research shows that when normal populations acknowledge someone as wise, a large part of what we look for is evidence of breadth and depth of experience with complicated human situations (Staudinger and Glück, 2011; Tiberius and Swartwood, 2011). In this light, it is odd to expect criticisms made any logically possible being to have much force for us. And it would require an ambitious and value-laden account of rationality to suggest that the criticisms of rational beings as such should have force for our ethical decision-making. In contrast, it is easy to see why the criticisms of wise human beings would have force for us, if as human beings we begin from the same experiences of being cared for and being wholehearted as good, of experiences being rejected and experiences of being conflicted as bad.

5.4 Conclusion

In this chapter I have argued that both from a Humean point of view, and also from a more radical perspective that challenges the dichotomy between fact and value, our shared human contingency puts important contrasts on which ethical and metaethical possibilities are real ones for us. If

a metaethical account of our practices of giving and exchanging reasons for breaks down when applied to a conversation with beings having a radically different evolutionary history, who cares? Those sorts of beings never seem to care enough to criticize our valuations, nor is clear that we would care about their criticisms if they cared enough to offer them. Taking the question of "who cares?" not in a rhetorical but in a substantive way, then, the answer is that it is only because of having a certain shared set of human sensibilities that we care to engage in criticism of one another's valuations. If I am right, because the set of normative frameworks that is psychologically possible for human beings to occupy is a heavily circumscribed subset of those that are logically possible, there is enough shared sensibility among the community of those who care about ethics to converge on a circumscribed set of ethical values.

Above I have noted some of the possibilities that are open for Humean naturalistic accounts of moral judgments when we do assume there to be a difference in kind between response-dependent evaluative properties and response-independent natural properties. In particular it is this divide that underwrites the distinction between the internal normative perspective on moral claims and the external, empirical perspective on moral expressions that Quasi-Realist accounts make so much of. If even the application of scientific properties involves a degree of dependence on human response, however, then the perhaps the difference between natural and normative perspectives is matter of degree. Some types of concept application are highly relative to human interests and other types are determined more by the way things are independent of us. On this approach, we might say that there is a ratio of response-independence to response-dependence, rather than a difference in kind between the two.

Constructivist approaches meta-ethics would seem ideally positioned to capitalize on an approach that takes all decisions about the application of concepts to be in some sense response-dependent. Indeed, Putnam suggests that in judging claims about facts as much in judging value claims, "we always bring to bear a large stock of both valuations and descriptions *that are not in question in that inquiry*" (Putnam 2002, 102; italics in original). An account such as Street's is explicitly designed to accept or reject particular normative commitments by appeal other norma-

tive commitments; the idea is that no non-normative perspective will be adequate to adjudicate such questions. So the account should be able to take in stride the idea that there are no absolutely response-independent properties; instead decisions about one property with some degree of response-dependence will depend on considerations about other properties with some degree of response-dependence. Quasi-Realists might seem to have a harder time, to the degree their account depends on holding a firm distinction between internal normative commitments and the empirical perspective one these commitments. Be that as it may, Blackburn himself seems to anticipate the problem, at least. Indeed, he offers the quasi-realist approach to mind-independent moral truths as a model for understanding how we can speak internally, from within a theoretical perspective, of correspondence with the mind-independent world, while maintaining from a perspective external to that one that these claims about mind-independent scientific truths are expressions of our own theoretical stance (1984, 247). Of course, judgments on this level, of the relative "coherence, comprehensiveness, and control" of a particular theoretical stance, will themselves be, from yet a further perspective, expressions of our own values. If so, it is not clear what the distinction between normative and theoretical perspectives will amount to. It might be quasi-truth all the way down.

My remarks in this section have not been aimed to bring closure to this deep and difficult issue, but on the contrary to open that question up in light of the approach to ethical value argued for in this dissertation. In particular, I want to suggest that if AWA proves a novel, cogent, and promising way to move forward in persistent debates in Western philosophy over the ground of ethical claims, then approaches inspired by Buddhist thought and practice may also be due serious consideration in regard to other contemporary metaphysical and epistemological issues. The central Buddhist inspiration for AWA was to suggest the possibility of a certain type of perceptual and affective shift, one that all human beings have reason to aspire to, and that brings with it a shift in ethical commitments and thereby suggests a shift in metaethical theorizing. In particular, I proposed that this sort of approach brings to the fore in ethical decision-making and ethical theorizing questions about the Qualities of Heart motivating one's actions. Secondly, I suggested that for the sorts of purposes that get us into debating and resolving ethical questions in the first place, asking

questions with this kind of evaluative focus may prove more serviceable than the other sorts of evaluative focus that have been suggested in Western moral philosophy. In regard to metaphysical or epistemological questions as with ethical ones, the crucial shift has more to do with which questions we "commend", to use Blackburn's terms, than it has to do with which answers we give. Decisions about which lines of inquiry are commendable or valuable for beings like us must on a Buddhist approach rest ultimately on considerations about which habits of mind lead us towards or away from perpetuating unease for ourselves. But that is a story for another time. For now, I close with Blackburn's suggestive remarks.

> We have to see our concepts as the product of our own intellectual stances: how then are they suitable means for framing objectively correct, true, judgment, describing the mind-independent world as it in fact is? It is not only moral truth that starts to quake. But we can learn to approach the general problems of truth by starting with it.
>
> -Simon Blackburn, *Spreading the Word* (1984, 198)

Chapter 6

Conclusion

In this dissertation I have argued for four claims. In Chapter 2, I tried to make plausible the empirical claim that to the degree human beings from any cultural background develops alertness to internal and external stimuli and attenuates affective biases of attention and memory, this will lead them to a systematic (re)formation of attitudes for and against certain emotional motivations and thus to a convergence in the ethical evaluations they express. In particular, my hypothesis was that to the degree any of us are Wide Awake in this way, we will agree with one another as to which Qualities of Heart are good and which are bad to act out of. As I noted, this is an empirical hypothesis subject to further testing. In Chapter 3, I moved from this empirical claim to a normative one. I argued that the ethical claims converged on by those who are Wide Awake have force for all of us, on our own terms, because they express the values we would hold to the degree we ourselves were Wide Awake. Of course, some logically possible being might simply reject the value of being Wide Awake, but I suggested that this avenue is less available to human beings because we share not only the character of being motivated to unease, and also affective profiles of particular Qualities of Heart, but also a shared fundamental project of preferring to be wholehearted. I left the detailed justification for that valuation hanging until Chapter 5, while in Chapter 4 I argued that competing foundational theories in ethics, from neo-Aristotelian naturalistic reductions to metaphysically sophisticated contemporary Utilitarian positions, from Kantian rationalism to neo-

Humean moral relativism, all failed to show the promise of the a subjectivist, affective, but yet universal approach I called Acting Wide Awake. In Chapter 5 I turned to metaethical issues, raising and resolving some differences between close cousins of AWA in the Humean camp. I agreed with Quasi-Realists, sentimentalist moral relativists, and Humean constructionist that many different starting sets of normative commitments are logically possible for a moralizing being to hold. Nonetheless, I suggested that in emphasizing this radical contingency to the neglect of considerations about which normative frameworks are psychologically possible for human beings to occupy, such Humean approaches have failed to consider a central source of shared human commitments. Indeed, this point about shared human contingency may rescue a kind of objectivity in ethical claims, and scientific ones, even if the ontological dichotomy between the two were to collapse. I noted that it is a very particular subset of all logically possible beings - and perhaps even a subset of bipedal hominids, not including for instance psychopaths - who care enough to criticize the valuations we make, and whose criticisms we care about in deciding whether what we value is indeed valuable. I drew on the arguments of earlier chapters to suggest that in this context the shared project of avoiding unease that we are all born into, along with the shared neurophysiology of various Qualities of Heart from hatred to goodwill, put significant constraints on which normative frameworks we can wholeheartedly commit to. I suggested that however convoluted the psychological convoluted and conflicted psychological positions we end up in, from masochism to outright endorsement of evil, each nonetheless carries with it the seeds of a basic, and shared, value in favor of being wholehearted. Given that it is only with such human beings that we make and debate claims about how to live, and which normative frameworks to endorse, this shared psychology is enough of a foundation to justify to any human being the ethical claim that devoting one's life to cultivating to love is a better thing to do than devoting one's life to cultivating hatred, I contend, and also that this implication carries over to how we choose to live and act in every moment.

Bibliography

Adams, R. M. (1976). Motive utilitarianism. *The Journal of Philosophy*, 73(14):467–481.

Barrett, L. (2006). Are emotions natural kinds? *Perspectives on Psychological Science*, 1(1):28.

Batson, C. D., Batson, J. G., Slingsby, J. K., Harrell, K. L., Peekna, H. M., and Todd, R. M. (1991). Empathic joy and the empathy-altruism hypothesis. *Journal of Personality and Social Psychology*, 61(3):413.

Batson, C. D., Duncan, B. D., Ackerman, P., Buckley, T., and Birch, K. (1981). Is empathic emotion a source of altruistic motivation? *Journal of personality and Social Psychology*, 40(2):290.

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

Bennett, J. (1974). The conscience of huckleberry finn. *Philosophy*, 49(188):123–134.

Blackburn, S. Sharon street on the independent normative truth as such. http://www.phil.cam.ac.uk/~swb24/PAPERS/Meanstreet.htm.

Blackburn, S. (1984). *Spreading the word*. Clarendon Press Oxford.

Blackburn, S. (1998). *Ruling passions: A theory of practical reasoning*. Oxford University Press, USA, New York.

Bratman, M. E. (2012). Constructivism, agency, and the problem of alignment. In Shemmer, Y. and Lenman, J., editors, *Constructivism in Practical Philosophy*, pages 81–98.

Brewer, J. A., Elwafi, H. M., and Davis, J. H. (2012). Craving to quit: psychological models and neurobiological mechanisms of mindfulness training as treatment for addictions. *Psychology of Additive Behaviors*.

Britton, W. B., Shahar, B., Szepsenwol, O., and Jacobs, W. J. (2011). Mindfulness-based cognitive therapy improves emotional reactivity to social stress: results from a randomized controlled trial. *Behavior Therapy*.

Chapman, H. A., Kim, D. A., Susskind, J. M., and Anderson, A. K. (2009). In bad taste: Evidence for the oral origins of moral disgust. *Science*, 323(5918):1222–1226.

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

Critcher, C. R., Inbar, Y., and Pizarro, D. A. (2012). How quick decisions illuminate moral character. *Social Psychological and Personality Science*.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, 108(2):353–380.

Cushman, F. and Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science*, 35(6):1052–1075.

Cushman, F. A., Sheketoff, R., Wharton, S., and Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, page in press.

D'Arms, J. and Jacobson, D. (2000). Sentiment and value. *Ethics*, 110(4):722–748.

D'Arms, J. and Jacobson, D. (2003). VIII. the significance of recalcitrant emotion (or, anti-quasijudgmentalism). *Royal Institute of Philosophy Supplement*, 52(1):127–145.

Davis, J. H. and Thompson, E. (2013). Developing attention and decreasing affective bias: Toward a cross-cultural cognitive science of mindfulness. In K.W. Brown, J.D. Creswell, and R.M. Ryan, editors, *Handbook of Mindfulness*, page 585–597. Guilford Press, New York.

Decety, J., Chen, C., Harenski, C., and Kiehl, K. A. (2013). An fMRI study of affective perspective taking in individuals with psychopathy: imagining another in pain does not evoke empathy. *Frontiers in Human Neuroscience*, 7.

Dretske, F. (1981). Knowledge & the flow of information.

Driver, J. (2006). Autonomy and the asymmetry problem for moral expertise. *Philosophical Studies*, 128(3):619–644.

Driver, J. (2009). The history of utilitarianism.

D'Arms, J. and Jacobson, D. (2010). Demystifying sensibilities: Sentimental values and the instability of affect. *The Oxford handbook of philosophy of emotion*, page 585–613.

Ekman, P. and Friesen, W. V. (1969). Nonverbal leakage and clues to deception. Technical report.

Fehr, E. and Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868):137–140.

Feldman, F. (2004). *Pleasure and the good life: Concerning the nature, varieties and plausibility of hedonism*. Clarendon Press Oxford.

Festinger, L. (1957). A theory of cognitive dissonance. *Evanston, Il: Row, Peterson*.

Flanagan, O. (2000). Destructive emotions. *Consciousness &# 38; Emotion*, 1(2):259–281.

Flanagan, O. (2007). *The Really Hard Problem: Meaning in a Material World*. The MIT Press, 1 edition.

Foot, P. (2003). *Natural Goodness*. Oxford University Press, USA.

Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy*, 68(1):5–20.

Gawronski, B. (2012). Back to the future of dissonance theory: Cognitive consistency as a core motive. *Social Cognition*, 30(6):652–668.

Gibbard, A. (2003). *Thinking How to Live*. Harvard University Press, Cambridge, MA.

Grabenhorst, F. and Rolls, E. T. (2010). Attentional modulation of affective versus sensory processing: Functional connectivity and a top-down biased activation theory of selective attention. *Journal of Neurophysiology*, 104(3):1649 –1660.

Grabenhorst, F. and Rolls, E. T. (2011). Value, pleasure and choice in the ventral prefrontal cortex. *Trends in Cognitive Sciences*, 15(2):56–67.

Grabovac, A. D., Lau, M. A., and Willett, B. R. (2011). Mechanisms of mindfulness: A buddhist psychological model. *Mindfulness*, 2:154–166.

Greene, J. D. (2002). The terrible, horrible, no good, very bad truth about morality and what to do about it. *Unpublished doctoral dissertation, Department of Philosophy, Princeton University*.

Greene, J. D. (2008). The secret joke of kant's soul. In Sinnott-Armstrong, W., editor, *Moral Psychology: Historical and Contemporary Readings*, volume Vol. 3, pages 35–80. MIT Press, Cambridge, MA.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., and Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, 111(3):364–371.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., and Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44(2):389–400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293(5537):2105 –2108.

Guyer, P. (2012). 3 passion for reason: Hume, kant, and the motivation for morality. *Proceedings of the American Philosophical Association*, (Presidential Address delivered before the One Hundred Eighth Annual Eastern Division Meeting of The American Philosophical Association in Washington, DC, on Thursday, December 29, 2011.).

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, 316(5827):998.

Haidt, J., Koller, S., and Dias, M. (1993). Affect, culture, and morality, or is it wrong to eat your dog?. *Journal of Personality and social Psychology*, 65(4):613.

Hamlin, J. K., Wynn, K., and Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, 450(7169):557–559.

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., and Ziker, J. (2010). Markets, religion, community size, and the evolution of fairness and punishment. *Science*, 327(5972):1480–1484.

Hill, C. S. (2009). *Consciousness*. Cambridge University Press, 1 edition.

Holmqvist, R. (2008). Psychopathy and affect consciousness in young criminal offenders. *Journal of Interpersonal Violence*, 23(2):209 –224.

Hume, D. (2000). *A Treatise of Human Nature*. Oxford University Press, New York.

Inbar, Y., Pizarro, D. A., and Cushman, F. (2012). Benefiting from misfortune: When harmless actions are judged to be morally blameworthy. *Personality and Social Psychology Bulletin*, 38(1):52–62.

Jacobson, D. (2008). Utilitarianism without consequentialism: The case of john stuart mill. *The Philosophical Review*, 117(2):159.

Kagan, S. (1998). *Normative Ethics*. Dimensions of Philosophy. Westview Press, Boulder, Colorado.

Kant, E. (1998). *Groundwork of the Metaphysics of Morals*. Cambridge University Press, 2 edition.

Kant, I. (2000). *Critique of the power of judgment*. Cambridge University Press, Cambridge, UK; New York.

Killingsworth, M. A. and Gilbert, D. T. (2010). A wandering mind is an unhappy mind. *Science*, 330(6006):932–932.

Kirk, U. (2011). Interoception drives increased rational decision-making in meditators playing the ultimatum game. *Frontiers in Decision Neuroscience*, 5:49.

Kirschbaum, C., Pirke, K.-M., and Hellhammer, D. H. (1993). The "Trier social stress test": A tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28(1-2):76–81.

Lacey, N. and Pickard, H. (2012). From the consulting room to the court room? taking the clinical model of responsibility without blame into the legal realm. *Oxford Journal of Legal Studies*, 33(1):1–29.

Lama, D. (2001). *Ethics for the New Millennium*. Penguin.

Lebrecht, S., Bar, M., Barrett, L. F., and Tarr, M. J. (2012). Micro-valences: Perceiving affective valence in everyday objects. *Frontiers in Psychology*, 3.

Lenman, J. (2012). Expressivism and constructivism. In Shemmer, Y. and Lenman, J., editors, *Constructivism in Practical Philosophy*, pages 213–225.

Loewenstein, G. and Lerner, J. S. (2003). The role of affect in decision making. In Davidson, R., Scherer, K., and Goldsmith, H., editors, *Handbook of affective sciences*. Oxford University Press, USA.

Lydall, E. S., Gilmour, G., and Dwyer, D. M. (2010). Rats place greater value on rewards produced by high effort: An animal analogue of the "effort justification" effect. *Journal of Experimental Social Psychology*, 46(6):1134–1137.

Mackie, J. L. (1977). *Ethics: Inventing right and wrong*. Penguin.

Malle, B. F. and Knobe, J. (2001). The distinction between desire and intention: A folk-conceptual analysis. *Intentions and intentionality: Foundations of social cognition*, page 45–67.

Malt, B. (1994). Water is not H2O. *Cognitive Psychology*, 27(1):41–70.

McDowell, J. H. (1998). *Mind, Value, and Reality*. Harvard University Press.

McMillan, R. L., Kaufman, S. B., and Singer, J. L. (2013). Ode to positive constructive daydreaming. *Frontiers in Psychology*, 4.

Mehling, W. E., Gopisetty, V., Daubenmier, J., Price, C. J., Hecht, F. M., and Stewart, A. (2009). Body awareness: construct and self-report measures. *PLoS One*, 4(5):e5614.

Mill, J. (1863). Utilitarianism. *Indianapolis, Indiana, USA: Hackett Publishing Company*.

Nichols, S. (2004). *Sentimental Rules: on the natural foundations of moral judgment*. Oxford University Press, New York.

Nichols, S. and Folds-Bennett, T. (2003). Are children moral objectivists? children's judgments about moral and response-dependent properties. *Cognition*, 90(2):B23–B32.

Nunner-Winkler, G. and Sodian, B. (1988). Children's understanding of moral emotions. *Child development*, pages 1323–1338.

Nyanaponika, T. (2000). *The Vision of Dhamma: Buddhist Writings of Nyanaponika Thera*. Pariyatti Publishing, 2 enlarged edition.

Ortner, C. N. M., Kilner, S. J., and Zelazo, P. D. (2007). Mindfulness meditation and reduced emotional interference on a cognitive task. *Motivation and Emotion*, 31(4):271–283.

Pickard, H. (2013). Responsibility without blame: Philosophical reflections on clinical practice. In KWM Fulford, M Davies, RT Gipps, G Graham, J Sadler, G Strangellini, and T Thornton, editors, *The Oxford Handbook of Philosophy of Psychiatry*. Oxford University Press, New York.

Pizarro, D., Uhlmann, E., and Salovey, P. (2003). Asymmetry in judgments of moral blame and praise. *Psychological Science*, 14(3):267.

Pizarro, D. A. (2010). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. *Unpublished manuscript, Cornell University*.

Prinz, J. (2001). *Furnishing the mind: Concepts and their perceptual basis*. The MIT Press.

Prinz, J. (2013). Are millikan's concepts inside-out? *Millikan and Her Critics*, page 198–220.

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

Prinz, J. J. (2007). *The Emotional Construction of Morals*. Oxford University Press, New York.

Prinz, J. J. (2010). *Is Empathy Necessary for Morality?* P. Goldie & A. Coplan, Empathy: Philosophical and psychological perspectives. Oxford: Oxford University Press.

Putnam, H. (2002). *The collapse of the fact/value dichotomy and other essays*. Harvard University Press, Cambridge, MA.

Rawls, J. (1999). *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, Mass.

Ridge, M. (2012). Kantian constructivism: Something old, something new. In Lenman, J. and Shemmer, Y., editors, *Constructivism in practical philosophy*, pages 138–158. Oxford University Press.

Roberts-Wolfe, D., Sacchet, M., Hastings, E., Roth, H., and Britton, W. (2012). Mindfulness training alters emotional memory recall compared to active controls: Support for an emotional information processing model of mindfulness. *Frontiers in Human Neuroscience*, 6.

Ross, S. (2004). Real, modest moral realism. *The Philosophical Forum*, 35(4):411–421.

Sarkissian, H., Park, J., Tien, D., Wright, J., and Knobe, J. (2010). Folk moral relativism. *forthcoming*.

Sequeira, H., Hot, P., Silvert, L., and Delplanque, S. (2009). Electrical autonomic correlates of emotion. *International Journal of Psychophysiology*, 71(1):50–56.

Shariff, A. F. and Norenzayan, A. (2007). God is watching you priming god concepts increases prosocial behavior in an anonymous economic game. *Psychological Science*, 18(9):803–809.

Silverstein, R. G., Brown, A.-C. H., Roth, H. D., and Britton, W. B. (2011). Effects of mindfulness training on body awareness to sexual stimuli: Implications for female sexual dysfunction. *Psychosomatic Medicine*, 73(9):817 –825.

Singer, P. (2005). Ethics and intuitions. *The Journal of Ethics*, 9:331–352.

Sinnott-Armstrong, W. (2009). Consequentialism. *Stanford Encyclopedia of Philosophy*.

Skitka, L. J. (2010). The psychology of moral conviction. *Social and Personality Psychology Compass*, 4(4):267–281.

Skitka, L. J., Bauman, C. W., and Sargis, E. G. (2005). Moral conviction: Another contributor to attitude strength or something more? *Journal of Personality and Social Psychology*, 88(6):895–917.

Slagter, H. A., Lutz, A., Greischar, L. L., Francis, A. D., Nieuwenhuis, S., Davis, J. M., and Davidson, R. J. (2007). Mental training affects distribution of limited brain resources. *PLoS Biol*, 5(6):e138.

Slingerland, E., Henrich, J., and Norenzayan, A. (2013). The evolution of prosocial religions. In Richerson, P. and Christiansen, M., editors, *Cultural Evolution*. The MIT Press, Cambridge, MA.

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

Smetana, J. G. (1981). Preschool children's conceptions of moral and social rules. *Child development*, page 1333–1336.

Staudinger, U. M. and Glück, J. (2011). Psychological wisdom research: Commonalities and differences in a growing field. *Annual Review of Psychology*, 62(1):215–241.

Stephens, C. L., Christie, I. C., and Friedman, B. H. (2010). Autonomic specificity of basic emotions: Evidence from pattern classification and cluster analysis. *Biological Psychology*, 84(3):463–473.

Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy*, 48.

Street, S. (2006). A darwinian dilemma for realist theories of value. *Philosophical Studies*, 127(1):109–166.

Street, S. (2009). In defense of future tuesday indifference: Ideally coherent eccentrics and the contingency of what matters. *Philosophical Issues*, 19(1):273–298.

Street, S. (2010). What is constructivism in ethics and metaethics? *Philosophy Compass*, 5(5):363–384.

Street, S. (2011). Mind-independence without the mystery: Why quasi-realists can't have it both ways. *Oxford Studies in Metaethics*, 6(1).

Street, S. (2012). Coming to terms with contingency: Humean constructivism about practical reason. In Lenman, J. and Shemmer, Y., editors, *Constructivism in practical philosophy*. Oxford University Press.

Sze, J. A., Gyurak, A., Yuan, J. W., and Levenson, R. W. (2010). Coherence between emotional experience and physiology: Does body awareness training have an impact? *Emotion*, 10:803–814.

Tiberius, V. (2012). Constructivism and wise judgment. In Lenman, J. and Shemmer, Y., editors, *Constructivism in practical philosophy*. Oxford University Press.

Tiberius, V. and Swartwood, J. (2011). Wisdom revisited: a case study in normative theorizing. *Philosophical Explorations*, 14(3):277–295.

Tisak, M. S. and Turiel, E. (1984). Children's conceptions of moral and prudential rules. *Child Development*, 55(3):1030–1039. ArticleType: research-article / Full publication date: Jun., 1984 / Copyright © 1984 Society for Research in Child Development.

Trivers, R. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, 46(1):35–57.

Tsai, J. L., Knutson, B., and Fung, H. H. (2006). Cultural variation in affect valuation. *Journal of Personality and Social Psychology*, 90(2):288–307.

Tsai, J. L., Louie, J. Y., Chen, E. E., and Uchida, Y. (2007a). Learning what feelings to desire: Socialization of ideal affect through children's storybooks. *Personality and Social Psychology Bulletin*, 33(1):17–30.

Tsai, J. L., Miao, F. F., and Seppala, E. (2007b). Good feelings in christianity and buddhism: Religious differences in ideal affect. *Personality and Social Psychology Bulletin*, 33(3):409–421.

Van Dam, N. T., Earleywine, M., and Altarriba, J. (2012). Anxiety attenuates awareness of emotional faces during rapid serial visual presentation. *Emotion*, 12(4):796.

van Vugt, M. K., Hitchcock, P., Shahar, B., and Britton, W. (2012). The effects of mindfulness-based cognitive therapy on affective memory recall dynamics in depression: A mechanistic model of rumination. *Frontiers in Human Neuroscience*, 6.

Wheatley, T. and Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science*, 16(10):780.

Woolfolk, R. L., Doris, J. M., and Darley, J. M. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition*, 100(2):283–301.

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., and Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences*, 107(15):6753–6758.