Moral Motivation and the Devil

Derek Christian Haderlie

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Arts
In
Philosophy

Tristram McPherson
James Klagge
Joseph C. Pitt

March 20, 2014
Blacksburg, Virginia

Keywords: metaethics, moral psychology, moral internalism, moral motivation

Moral Motivation and the Devil

Derek Christian Haderlie

ABSTRACT

In this paper, I call into question the thesis known as judgment internalism about moral motivation. Broadly construed, this thesis holds that there is a non-contingent relation between moral judgment and moral motivation. The difficulty for judgment internalism arises because of amoral agents: when an agent both knows the right and yet fails to be motivated to act on this knowledge. Specifically, I cite John Milton's Satan from *Paradise Lost*. This is a problem because it calls into question the non-contingent relation between moral judgment and moral motivation. I argue that in order for judgment internalism to be viable in reconciling judgment internalism and amoralism, it must provide plausible accounts of both (a) the relationship between judging and motivation, and (b) the conditions for defeasibility. While crude versions of the thesis fail to do this, I provide a revised thesis which I call Narrative Internalism, which assumes a narrative theory of the self. This thesis has the dual strength that it can account for both why one would typically be motivated to Φ upon judging that it is right to Φ and also the conditions that might obtain such that one would fail to be motivated. This account of moral psychology explains both (a) the relationship between judging and motivation, and (b) the conditions for defeasibility by giving an account of plausible defeasibility conditions. I conclude that unless there are more plausible accounts of judgment internalism in the offing, which doesn't seem apparent to me, we should adopt Narrative Internalism.

Acknowledgments

Thanks first goes to my thesis advisor Tristram McPherson, who tirelessly read, commented, and discussed this thesis. My hope was to write something significant and novel—I think I succeeded in the latter because of Tristram, and while I unfortunately failed in the former, it was in spite of him. Thanks also to my committee members, James Klagge and Joseph Pitt. Furthermore, I would like to thank Michael Moehler, Steve Mischler, and Mike Zarella for their comments on earlier drafts of the thesis. Ross Baron, Whit Laga, and Justin Hosman all helped me to feel the significance of the problem of moral motivation—thanks to them for thousands of hours of stimulating conversation and for being good friends. Thanks finally goes to my wife Amanda, who was patient with my late nights and early mornings as a graduate student. And who has supported me in following my heart and pursuing my dreams. Whatever the outcomes of my work in philosophy come to, I wouldn't have been able to do any of this without her.

Table of Contents

Moral Motivation and the Devil

How art thou fallen from heaven, O Satan, son of the morning!
how art thou cut down to the ground, which didst weaken the nations![1]

Introduction

Those who hold the position that there is a non-contingent relation between moral judgment and moral motivation are, broadly speaking, internalists about moral motivation. A crude form of the internalist thesis can be formalized thus:

> **Crude Internalism**: If S judges that it is right to perform some act Φ in circumstances C, then she will be non-contingently motivated to Φ in C.[2]

Internalism stands in contrast to externalism. Externalists simply state that the relationship between moral judgment and moral motivation are contingent. The externalist might cite cases of amoralism as evidence that Crude Internalism is demonstrably false. The amoralist thesis can be stated thus:

> **Amoralism**: S is amoral just in case (a) she judges that it is right to Φ but (b) fails to be motivated to Φ and/or (c) instead is motivated to not Φ.[3]

Amoralism and Crude Internalism are *prima facie* incompatible theses. This is because, for the Crude Internalist, there is a *non-contingent* relation between one's judging that it is right to Φ and one's being motivated to Φ. But Amoralism suggests that the relationship is contingent. Assuming that there are *actually* cases of Amoralism, we should reject Crude Internalism. Still,

Aknowledgments: many thanks to Tristram McPherson, James Klagge, Michael Moehler and Steven Mischler for helpful comments both written and in conversation on earlier drafts.

[1] Isaiah 14:12, KJV.

[2] Smith (1994) p. 61.

[3] This formulation of the Amoralism thesis is intended to be broad, with the aim that it capture not only weakness of will, but also the depressive, the irrational, and those with overwhelming competing reasons (including spite or rebellion). Thus, I take it that the problem of Amoralism is a much more potent problem than just the problem of weakness of will or the akratic agent (which, presumably, would be easier to account for). See also Davidson (1980) Ch. 2.

the internalist might object: it may be possible for certain conditions to obtain that would defeat one's moral motivation, but that would not undermine the claim that the relationship between moral motivation and moral judgment is non-contingent. If this objection is to be sustained, it places a special burden on the internalist to provide an account of moral motivation that includes plausible defeasibility conditions, i.e., conditions where the relevant motivation exists (so consistent with internalism), but where that motivation is defeated by competing factors:

> **Defeasibility Condition**: S's motivation is plausibly defeasible just in case a state of affairs or condition obtains such that one's motivation to do Φ, where one judges that it is right to Φ, is defeated either (a) through rational failure or (b) by overwhelming competing reasons not to Φ (e.g., under threat of life).

Michael Smith, notably, gives such an account in his book *The Moral Problem*. Still, I believe that his account is incomplete, i.e., his account leaves a gap. My aim in the paper is to give an account of one part of moral psychology that will provide the necessary resources for plausible Defeasibility Conditions for moral motivation to fill this gap. I will take the problem head-on by considering perhaps the most conceptually difficult case of amoralism: the case of Satan's rebellion as depicted in John Milton's *Paradise Lost.*

Milton's Satan (i.e., Lucifer or the Devil) is an angel that lives with God in heaven. He is one of God's great angels: an angel whose "transcendent brightness, didst outshine / Myriads though bright."[4] To reach such an exalted station, Satan has to have been one of God's most righteous and obedient angels. Furthermore, because he was in heaven with God Satan has a privileged knowledge of the good and the right.[5] He has all the correct and relevant beliefs-- he

[4] Milton (2005) p 13.
[5] Presumably in the strongest infallibilist sense, in virtue of Satan's being with God.

"admits God was right"[6] and yet he *rebels* against God. Indeed, Satan is *the* paradigm case of Amoralism, for not only does he fail to be motivated to be obedient even though he judges it right to do so, he instead is motivated to rebel against God, in the ultimate act of disobedience.

The account I offer in this paper reconciles from an internalist perspective how Satan could both judge it wrong to rebel and yet fail to act accordingly. In other words, this account will provide plausible Defeasibility Conditions for moral motivation, both in the Satan case and in general, making it possible to reconcile the internalist and amoralist theses. The paper will consist of two major sections. In Section I, I show more explicitly how Crude Internalism fails in light of Amoralism. I ultimately argue that previous attempts to reconcile an internalist thesis with Amoralism have left gaps, making a more robust internalism necessary. In Section II, I sketch an account of a part of moral psychology. I begin by briefly explicating a definition of *understanding*. Then I argue that the self is constituted by *narrative*. Together these notions combine to provide a rich moral psychology with the resources to produce a more robust version of internalism that fills the gap other internalist theses leave open.

I. The Problem of Moral Judgment and Amoralism

In order for judgment internalism to be reconciled with Amoralism it must provide plausible accounts of both (a) the relationship between judging and motivation, and (b) the conditions for defeasibility. This is because (a) without (b) cannot be reconciled with Amoralism, and (b) without (a) would fail to explain why people are typically motivated to do what they judge that they ought to do. In this section of the paper I will analyze the standard ways in which judgment internalism accounts for (a) and (b), with the aim of motivating my own

[6] Carey (1999) p. 134. It is also interesting to note that many of the Milton scholars have struggled with this same problem, and have provided various arguments on how it is possible that Satan could rebel given his privileged epistemology and apparent assent to God's rightness.

account of moral judgment with associated Defeasibility Condition.

(a) Judgment and Motivation

Consider the Darlene case:

> Suppose that a friend, Darlene, asks you for counsel about a difficult moral decision she has to make. There are two choices, do *A* or do *B*. Suppose that you furnish Darlene with arguments *for* doing *A* and *against* doing *B*, such that based on your arguments Darlene *accepts* and judges that it is right for her to choose to do *A* and *not* to do *B*.

There is a strong (internalist) intuition that on accepting (judging) that it is right to do *A* Darlene will now be non-contingently motivated to do *A*. In fact, for the internalist, it would be surprising if Darlene *actually* did *B* in spite of her judging it right to do *A* and not right to do *B*. This example furnishes strong intuitive reasons to favor the internalist thesis, *prima facie*. Still, although there may be intuitive appeal to accepting internalism, the relationship between judgment and motivation is not entirely clear.

According to the internalist, the work of producing motivation in the judge can be read in two different ways, the metaphysical reading and the psychological reading.[7] In this paper I will focus on the psychological reading. On this reading, motivation arises out of the nature of the psychological *state* of judging an act to be right. In other words, motivation arises out of some psychological feature of judgment itself. Under this reading one can be either a cognitivist or a noncognitivist about moral judgment:

[7] For more on this distinction, see Darwall (1997) pp. 305-312. (For the internalist that accepts the metaphysical reading, motivation arises out of the nature of the moral fact(s) corresponding to moral judgment. So, the *fact* that Φing is right is what tends to motivate people. In other words, the *content* of moral judgments is what does the motivating work. The metaphysical reading is largely ignored because it attributes "queer" properties to moral facts. I will follow this trend. See Mackie (1998) p. 38 and Olson (2010) p. 3. Note: there is some debate over whether Mackie mislocates queerness, and that Mackie should have focused on the moral judgment itself rather than on the moral facts.

Cognitivism: moral judgment so understood is, or is wholly constituted by, ordinary beliefs. [8]

Noncognitivism: moral judgment so understood is, or is wholly constituted by, something other than ordinary beliefs.[9]

Another way of stating the noncognitivist thesis is just to say that to judge some act Φ as being right just is to be motivated to Φ. Given the indefeasible nature of Crude Internalism, there is pressure to accept Noncognitivism. This kind of view nicely situates itself to explain our intuitions in the Darlene case. This is because, assuming we act for reasons, Noncognitivism offers an account of how moral judgments provide reasons to act. [10] To make this clearer, consider the distinction that Michael Smith makes in *The Moral Problem* between *explanatory* and *justificatory* reasons.[11] The former are motivating reasons and the latter are justifying reasons. Having a motivating (explanatory) reason is, Smith claims, the same as having a goal (i.e., they are teleological in nature).[12] And furthermore, Smith says, having a motivating reason is essentially "being in a state [of] desiring."[13] This view of motivating reasons is loosely Humean, appealing to the idea that desires have world to mind directions of fit, in contrast to beliefs, which have mind to world fit.[14] When one gives an explanation of "why" one performed

[8] Note Rosati's (2006) explanation of this under the section on Humean and anti-Humean views. See also Campbell (2007).
[9] It is also possible that one could have a hybrid view that would either give some account of moral judgments consisting either of both ordinary beliefs and something else that is not an ordinary belief, or some belief state that falls outside of "ordinary." Non-cognitivists apparently don't suffer from the same problems of explaining motivation that the cognitivists about morality do, because moral propositions don't have cognitive content, but rather, they are simply expressions of attitudes or preferences. For the sake of brevity, I will not spend time raising objections to these views other than to say that I take it that that moral propositions do have cognitive content (they are either true or false), and therefore I believe that a strictly non-cognitivist view is incorrect.
[10] This is a loosely rationalistic view. For a defense of rationalism, see Smith (1994) Sections 3.1-3.2.
[11] Smith (1994) pp. 94-98. The way that Smith cashes out this distinction is not uncontroversial, for example Nagel (1970) would probably reject the distinction. See Smith (1994) sec. 4.3 for more discussion on this.
[12] Ibid. p. 116.
[13] Ibid. p. 116.
[14] A belief has a world-to-mind direction of fit, whereas a desire has a mind-to-world direction of fit. In other words, a belief is defective insofar as it fails to match up with the *way the world is*. Alternatively, when a person has a

some act Φ, then the explanation would appeal to a motivating reason or desire.

In contrast to motivating reasons, justificatory reasons do not motivate, they simply justify. Thus, if someone were to ask, not what caused you to act, but what justified your act, one would appeal to justificatory reasons. Justificatory reasons are normative reasons. Therefore, it is in virtue of the *explanatory reasons* that one is motivated to Φ, while the *justifying reasons* provide normative justification for Φing.

With this explanatory/justificatory reasons distinction in hand, it becomes clearer that Noncognitivism gives an attractive account of *why* a person would act on her judgment to Φ. It is because her judging it right to Φ *is the same as* having motivation to Φ, which means that there is an explanatory reason for her Φing, because to judge some act Φ as right just is to have a motivating reason to Φ. Thus, Noncognitivism provides for Crude Internalism an account of the relationship between moral judgment and moral motivation.

The problem for Noncognitivism is that it cannot adequately explain Amoralism, because to judge that it is right to Φ, for the non-cognitivist, just is to be motivated to Φ. So, it is difficult to account for what it might be for one to judge it right to Φ and yet fail to be motivated to Φ. This also demonstrates that Crude Internalism, as stated above, is far too strong a claim, and needs to be weakened. This is because Crude Internalism, like Noncognitivism, can have no plausible Defeasibility Conditions, by its very nature.

Because Amoralism presents a devastating objection to the strong (crude) version of

---

desire, the desire functions to *bring about a change in the world* such that the world matches the desire (See Anscombe p. 56 for illustrations). Consider this illustration: Jim believes *and* desires that all children are provided with a minimally decent education before becoming adults. As Jim goes about his life, he begins to see evidence that contradicts his belief that all children are provided with a minimally decent education. Therefore, his beliefs begin to change to match up with the way the world *really* is. But, he continues to desire for the world to be such that all children are provided with a minimally decent education before becoming adults. Therefore, Jim sets out campaigning for minimally decent education for children. Because desire is the direction of fit that is world to mind, it is, for Smith, the direction of fit that motivates. Therefore, desires motivate, while beliefs do not. Therefore, if one accepts the Humean thesis about direction of fit, then she is obliged to recognize that beliefs are not motivating.

internalism, we need an account of the relationship between judgment and motivation which also allows for a plausible account of defeasibility. Crude Internalism simply doesn't have the resources to do this. This highlights the need for (b) conditions of defeasibility because (a) the relationship between judging and motivation alone cannot explain Amoralism.

(b) Defeasibility Conditions

Because of Amoralism, as Stroud puts it, we must navigate "between the Scylla of an extreme internalism about evaluative judgment which would preclude the possibility of weakness of will, and the Charybdis of an extreme externalism which would deny any privileged role to evaluative judgment in practical reasoning or rational action."[15] In order to maintain a broadly internalist stance, there are three ways to navigate these waters: (i) deny the possibility of Amoralism altogether, (ii) concede that there are some unanswerable cases of Amoralism, but argue that Internalism still takes the day, and (iii) confront Amoralism head-on by offering plausible Defeasibility Conditions. I argue that some form of (iii) is the best way.

First off, being able to account for Amoralism can only be a good thing for whatever account of internalism one accepts. David Wiggins makes a similar claim, stating that while he "cannot claim that it is inconceivable that this pretheoretical description of weakness of will should be strictly and literally true of nothing," he does claim that, "he who values his pet theory above the phenomenon, and wants to hold that weakness of will as I have described it as simply an illusion, will need to command some formidable conceptual-cum-explanatory leverage in the philosophy of value and mind-and an Archimedean fulcrum of otherwise inexplicable facts of human conduct."[16] In other words, it's hard to imagine that any philosopher would have the

[15] Stroud (2008) n.p.
[16] Wiggins (1979) p. 251.

cognitive and theoretical resources to deny Amoralism without having dubious assumptions built in. From this I conclude that it would be a virtue of any account of internalism that it be able to explain Amoralism without denying it as a possibility, thus avoiding dogmatism.

Now, consider (i), the denial of Amoralism altogether from the argument from insincerity: although the amoralist claims, for example, that it is right to Φ, she does not *sincerely* judge that it is right to Φ. On this view, the amoralist has the ability to accurately track what is generally taken to be right and wrong in society, but fails to *sincerely believe* that those moral norms *actually* have bearing on her life. Therefore, although she has the appearance of making a moral judgment, she does not in reality make a moral judgment.[17]

I find this response unsatisfying, first because I worry that this kind of objection to Amoralism simply begs the question about the impossibility of amoralism. This is because it can seem to say, Amoralism is impossible, therefore, one cannot be an amoralist. Finally, it seems that it would be an appropriate desiderata for a view of internalism to be able to have a way of giving an explanation of the amoralist problem that doesn't just simply dismiss it by saying to the supposed amoralist, "obviously you are not being sincere, and it doesn't matter that you are self-reporting that you *are* sincere."[18] Therefore, we should reject (i) the denial of Amoralism altogether.

Instead, consider (ii) there are those that argue that internalists can concede some ground

[17] See van Roojen (2010), Lenman (1999), and Brink (1986) for more carefully articulated versions of this argument, along with several other considered arguments against the possibility of the amoralist.

[18] Michael Smith presents a critique of an argument that David Brink makes that is related to the anti-amoralist response under consideration: "[Brinks] puts a prejudicial interpretation on the amoralist's reliable use of moral terms. He assumes that the amoralist's reliable use is evidence of her mastery of those terms; assumes that being suitably motivated under the appropriate conditions is not a condition of mastery of moral terms. But those who accept the [internalist thesis] do not accept the account of what it is to have mastery of moral terms that makes this prejudicial interpretation of the amoralist's use of moral terms appropriate" (Smith 70). Smith correctly, I think, states that this kind of interpretation of the amoralist is a "prejudicial" one. This relates to my concern about this kind of interpretation being question begging. We are both interested in getting an account of moral judgment that doesn't simply beg the question.

to externalism without giving over the whole debate. This kind of view is defended by Simon Blackburn in his book *Ruling Passions*. In the book he basically concedes that there are cases of Amoralism that the internalist cannot account for (Satan and Othello are two specific examples he uses), but he argues that overall internalism "wins the war" regarding moral judgment.[19] This is because, he argues, internalism gives a better overall account of our pretheoretic intuitions about moral motivation, but, he admits, that doesn't mean that it always gets everything right.

I find this kind of view especially problematic, because in principle, if there is even one Amoralist case that cannot be accounted for by an account of internalism, then it acts as an undercutting defeater for internalism—thus, internalism is false. For, if there really is a non-contingent relation between moral judgment and moral motivation, then any case that demonstrates the relation being a contingent one defeats the account. Therefore, we should reject (ii). Furthermore, given that on this view it is conceded that there are indeed difficult cases of Amoralism, there is pressure on the internalist to formulate a defeasible version of internalism, and thus take strategy (iii) to confront Amoralism head-on by offering plausible Defeasibility Conditions.

Therefore, it appears that strategy (iii), taking Amoralism head-on by offering plausible Defeasibility Conditions, is our best choice. There are two virtues of this approach: first, it takes Amoralism, as a problem, seriously à la Wiggins. Second, if we can account for all the cases of Amoralism, then, *ceteris paribus*, we have a more robust view than we have if we are to take either strategy (i) or (ii).

Michael Smith uses strategy (iii) in his account of internalism. He offers an account of

[19] Blackburn (1998) p. 65. He also makes what I consider to be a peculiar argument in which he says that Amoralism is only conceptually coherent against the backdrop of internalism, and thus we should favor internalism. But, it is not clear why the externalist couldn't just as easily give an account of Amoralism that was entirely conceptually coherent.

internalism with, what he considers to be, a plausible Defeasibility Condition that can account for most of the standard cases of Amoralism. On his account "If an agent judges it right to Φ in circumstances C, then either she is motivated to Φ in C, or she is practically irrational."[20] On this account of internalism, practical irrationality is the Defeasibility Condition for moral motivation. This Defeasibility Condition allows Smith's account to reconcile his internalism with most standard cases of Amoralism:[21] for example the depressive, the weak of will, the irrational, and those with overwhelming competing reasons.

But even if Smith's account does in fact allow him to reconcile his account of moral judgment with most cases of Amoralism, it still leaves a gap, one made salient in the case of Satan. Remember that Satan is an angel that lives with God. Satan is privy to *the good* and *the right* in a way that no mortal is. Therefore, the gap remains because it is not clear that Satan is "practically irrational." Indeed, he doesn't seem to be depressed, he is clearly not weak-willed, he doesn't seem straightforwardly practically irrational, and finally, he actually has nearly overwhelming prudential or practical reasons not to rebel.[22] Yet, he does rebel. Therefore, there is at least one case of Amoralism that exists that needs to be explained if internalism is to remain a plausible view.

In this section of the paper I have attempted to make my target clear: I intend to give an account of moral psychology that has the resources to explain the Satan case, with the aim of being able to account for Amoralism generally—while filling in the gap left open by Smith, and

---

[20] Smith (1994) p. 61.

[21] Smith cites Ayer (1945), Frankfurt (1971), Watson (1975), and Stocker (1979) in providing various kinds of Amoralist cases.

[22] The terms 'irrational' and 'rational' are extremely vague. The problem is that there are many different accounts of rationality on offer. To give a full account of rationality would be too much for this paper. Instead I will simply draw from Bernard Williams, according to Michael Smith: An agent is practically rational when (i) the agent has no false beliefs, (ii) the agent must have all relevant true beliefs, and (iii) the agent must deliberate correctly (which, according to Smith, includes some kind of wide reflective equilibrium) (Smith, 156). I personally think that this account of rationality is far too strong. But, I also think that Satan seems to meet this criteria (more or less).

presumably leaving no additional gaps, as the Satan case is the most conceptually difficult case of amoralism.

II. Towards an Account of Moral Judgment

The account (of one part) of moral psychology I offer in this section suggests that our *personal narratives* (our narrative selves) form the content of our *moral understanding*, and that together they provide the basis for a robust internalist view of moral judgment that is able to account for the Satan case and amoralism generally, thus filling the gap left open by Michael Smith.

The section will consist of four subsections. In the first I will draw from some recent literature on the notion of understanding. In the second I will outline the narrative conception of the self. In the third, I will take the notion of understanding and explain how it relates to the narrative self. Finally, in the fourth, I will show how this gives us a robust moral psychology that has the resources to vindicate an internalist account of moral judgment such that it has the plausible Defeasibility Conditions and can be reconciled with Amoralism.

Understanding

Understanding will be an important component of my account of moral psychology and therefore of my account of moral judgment. I assume the following thesis:

> **Understanding:** S understands some proposition or object p just in case S has (i) a malleable mental representation of p and (ii) the ability to manipulate the mental representation of p.[23]

[23] The reader should note that this is a controversial thesis in two ways. First, there are those who register serious doubts that there is some kind of cognitive state altogether different from belief. Second, there is considerable

First, S's understanding some proposition p is in part constituted by S's having a malleable mental representation of p. Linda Zagzebski begins to suggest something like this when she says that understanding "involves seeing how the parts of [a] body of knowledge fit together. . . ."[24] In turn, seeing how some set of, e.g., propositions "fit together" is to have mental representation or mental profile of how the constituent propositions relate to each other. This notion of understanding is further supported by Daniel Wilkenfield, who argues that "In order to understand some object x, a thinker must possess a mental representation of x."[25]

In addition to having a mental representation, one's mental representation must be malleable. This is because a static mental representation is not capable of being manipulated, and thus is unable to adjust to new evidence or upon reflection, etc. Thus, the mental representation is malleable insofar as it not static, but is capable of being changed or adjusted. So conceived, having a malleable mental representation is, in part, constitutive of understanding. Insofar as the mental representation fails to be malleable, it fails to be useful in many ways in which we expect it to be useful, e.g., in using one's understanding of New York City's makeup (layout, pedestrian congestion, etc.) to plot a new and never before used path to a given destination in the city. Therefore, "S understands that p" is not exhausted by S having a mental representation of p. S must also have the ability manipulate the mental representation of p. For example, seeing the implications of p for q. Thus, S's understanding p is constituted by S having both a mental representation of p and the ability to engage with and manipulate the mental representation in various ways.

Furthermore, "understanding … may be achieved in more than one way about the same

---

controversy about how the notion of understanding should be cashed out: is it just a *kind* of knowing, or is it something altogether different, as I am suggesting here? Also, here I am assuming that propositions are not necessarily linguistic in nature, but that propositions are a broader category

[24] Zagzebski (2001) p. 244.

[25] Wilkenfield, (2013) p. 1003. Notice that Wilkenfield uses "objects" rather than "propositions."

portion of reality. More than one alternative theory may give understanding of the same subject matter."[26] In other words, for example, one person can understand some set of propositions differently than another person understands that same set of propositions. Even more strongly stated, two people could both *fully* understand x, but understand it differently. This is because relationships between the objects of understanding can be drawn in different ways, giving each person a potentially unique relational profile, or understanding, of the objects of understanding.

Zagzebski distinguishes between knowledge and understanding as different kinds of cognitive states. Knowledge on the one hand is roughly the cognitive state constituted by one's holding certain things to be the case,[27] while on the other hand understanding is roughly the cognitive state constituted by one's seeing how things fit together.

Consider, further, how understanding is distinct from knowledge or belief. Zagzebski argues, "We can have both understanding and knowledge about the same part of reality. Understanding deepens our cognitive grasp of that which is already known. So a person can know the individual propositions that make up some body of knowledge without understanding them."[28] This is significant in that one can know or believe certain propositions without understanding them. i.e., they can belief some proposition, without having any idea of how it might together with other related propositions or concepts. It also seems that one can understand certain propositions without knowing or believing them. Consider as an illustration two biology teachers. One who is a creationist, and one who is an evolutionist. Perhaps both of these teachers have obtained identical educations, such that they have at their disposal all of the same propositions and facts about the theory of evolution. The difference is that the creationist teacher

[26] Zagzebski (2001) p. 244.
[27] This is clearly a brief gloss on knowledge. Because my interest in this discussion isn't really about knowledge, but rather about understanding, I will be more concerned in how understanding is described. I hope to just get the broad strokes of what constitutes knowledge out of the way.
[28] Ibid. p. 244.

does not believe in the theory, while the evolutionist does. It seems possible that the creationist teacher could teach the theory just as well as the evolutionist. Indeed, while they have different beliefs (and thus different knowledge) about the theory of evolution they may have similar understanding. For they may both satisfy the conditions for Understanding. Likewise, I may believe in the theory of evolution, and yet not have as robust a mental representation, or even more, not have any idea how to manipulate it, thus lacking understanding. Thus, while the creationist teacher does not know or believe the theory, she understands it—and further, while I may not fully understand the theory, I could still believe that it is true (and potentially know it is true or fully accurate).[29] So, knowledge and understanding come apart.

The Narrative Self

Barbara Hardy, a literary theorist and critic, said "We dream in narrative, daydream in narrative, remember, anticipate, hope, despair, believe, doubt, plan, revise, criticize, construct, gossip, learn, hate and love by narrative."[30] I am sympathetic to this statement, and I think that it gives us a clue to seeing how we should conceive of selves. Here I will outline an argument for the narrative self,[31] the self as constituted by narrative, with the aim of providing an account of the content of *moral understanding*.

In his book, *After Virtue,* Alasdair Macintyre argues for a *telos* centered worldview as a worldview that is founded on the narrative self:

[29] This case is reminiscent of another case by Lackey (1999) p. 477. Some readers may be worried that my use of 'know' in parentheses is problematic. The simple point I am trying to make here is that given that knowledge requires belief, and that since belief and understanding clearly come apart, that knowledge and understanding, by implication, also come apart.
[30] Hardy (1968) p. 5.
[31] I should make clear that I am not trying to lay out an argument for identity, but rather I am trying to make plausible a rich conceptual schema for conceiving of the self. I think that whether this is what constitutes the identity of a person in a metaphysical way or not is not in question in the following, but rather how we generally psychologically and phenomenologically conceive of our selves.

> I am what I may justifiably be taken by others to be in the course of living out a story that runs from my birth to my death; I am the *subject* of a history that is my own and no one else's, that has its own peculiar meaning. When someone complains – as do some of those who attempt or commit suicide – that his or her life is meaningless, he or she is often and perhaps characteristically complaining that the narrative of their life has become unintelligible to them, that it lacks and point, any movement towards a climax or a *telos*. Hence the point of doing any one thing rather than another at crucial junctures in their lives seems to such a person to have been lost.[32]

Thus, for Macintyre, the narrative conception of the self consists of a story or history of the self that each person lives which has directionality or a *telos*. Jerome Bruner fills the idea of narrative out: "A narrative is composed of a unique [set] of events, mental states, [and] happenings involving human beings as characters or actors. These are its constituents. But these constituents do not, as it were, have a life or meaning of their own. Their meaning is given by their place in the overall configuration of the sequence as a whole—its plot or *fabula*."[33] I will call these constituent parts narrative facts, or just n-facts. Notice, each n-fact obtains its significance from its position in the narrative as a whole. In other words, the n-facts by themselves do not constitute an intelligible (meaningful) whole.

Drawing from Macintyre, intelligibility has two major components: coherence and purpose. First, consider *coherence*. Intelligibility demands that the n-facts fit together in a coherent way. The narrative self is coherent just in case each of the n-facts relate to each other in

---

[32] Macintyre (1984) p. 202.

[33] Bruner(1990) pp. 43-44. Marya Schechtman argues that the form of the narrative self should be understood as "a conventional, linear narrative" (96). I disagree with Schechtman's interpretation of Bruner's description of narrative. The narrative of one's life doesn't seem to follow a strictly linear path, but is often recursive and/or circular. I see no reason to think that linearity is a necessary condition for intelligibility. Even according to Schechtman's own lights, it is hard to see why we should conceive of the self-narrative as being conventional and linear in order to be intelligible: she points to fiction saying that we can tell when a fictional character has been "well drawn" or not: "Although each of the actions, emotions, beliefs, and so on that are ascribed to her may be unproblematic in itself, we have no sense of how to understand them as coexisting in a single subject—we get no sense of *who* this person is and what the guiding principles of her life are" (97). It is clear that Schechtman shares Macintyre's basic view of intelligibility, but this view doesn't seem to demand that the narrative follow a conventional linear story line. Rather, it must meet the minimum standards of the intelligibility: it must be minimally coherent and have a minimally clear *telos* or purpose. I simply don't think that the view demands such provincial constraints on intelligibility as linearity and conventionality. In other words, we should not add linearity or conventionality to the intelligibility constraints we have already established (coherence and purpose).

consistent ways. Second, consider *purpose*. Intelligibility demands that there be a purpose or *telos* towards which the narrative as a whole is moving. The narrative self must therefore have a sense of directionality, or movement.[34] If one's purpose or *telos* becomes obfuscated, the narrative fails to be intelligible.[35] In sum, in order to be intelligible a narrative self must be both minimally coherent and have a minimally clear purpose or *telos.*

So, the self is constituted by a set of n-facts arranged in a narrative structure that is more or less intelligible depending on its coherence and *telos*. All of this is to say that the self is a narrative. And as I will presently argue, it is the narrative self that forms the content of our moral understanding.

Narrative Understanding

In order to sketch how the notions of *understanding* and *narrative selves* work together, some distinctions need to be made. First, I distinguish between what I will call the base narrative

[34] Bernard Williams suggests that individuals have projects in which they are engaged. I think that these projects might relate to or resemble the notion of *telos* I have in mind. These projects can have greater or lesser value for individuals, depending on their relative importance and priority in the individual's life. For example one could have a high value project in caring for her children, while pursuing her hobby of graphic design has lesser value relative to the priority of being a good mother. When these projects come into conflict she will make decisions on which to pursue based on her prioritization of them as they relate to her telos. Williams critiques utilitarianism, especially of the Benthamite brand, for not being sensitive to the projects of individuals. Williams view seems to give us an insight into intelligibility—when one's high value projects come into conflict, this conflict can fragment the narrative and leave it unintelligible. Thus, the narrative self depends on proper ordering of projects which relates one's intentions and purposes—the telos. This is not to say that any conflict necessarily fragments the narrative to the point of unintelligibility. Rather, each person's narrative has its own relative durability depending on its scope and level of intelligibility. Furthermore, the narrative of a person can be more or less intelligible, and need not be completely intelligible or completely unintelligible (although these states are conceptually possible, they are unlikely). As far as I understand the view, he is not suggesting something like the narrative self, but I see his view as having a clear analogue with my view of the narrative self, thus I think it is profitable to bring his view up. See Williams (1973) pp. 108–117.

[35] It seems that certain kinds of beliefs could obfuscate the *telos.* For example, if someone were to become convinced that they were a worthless human being, perhaps through incessant bullying, the *telos* would be obfuscated, and they would perhaps sink into a depressive state because the narrative would be unintelligible.

and the articulable narrative.[36] The base narrative is the narrative formed by the entire set of relevant n-facts. Therefore, the base narrative is the narrative of the person from a god's eye view. On the other hand, there is an articulable narrative. This is the narrative that informs the decisions that a person makes, insofar as she reflects upon her narrative. The articulable narrative is the narrative formed by bringing some subset of all the available n-facts into a representation.[37] In order to avoid ambiguity, I will call the n-facts that are brought into the articulable narrative n-beliefs for narrative beliefs. [38] The articulable narrative may be more or less accurate depending on both how well it corresponds to the base narrative, i.e., (1) how many of the n-beliefs actually correspond to n-facts, and (2) how many of the n-facts make it into the articulable narrative as n-beliefs. Call these two features of accuracy the *fit* of the articulable narrative.

Second, the narrative has a cognitive and a non-cognitive component. The cognitive component is just one's n-beliefs as brute representations. On the other hand, the non-cognitive component is just one's attitudes, feelings, and so forth about the various n-beliefs. To illustrate this distinction, consider two people that assent to the proposition 'there is a god.' One person might have an apathetic attitude about the proposition, while another person might have a zealous attitude about the proposition. They might each agree about the truth-value of the proposition, they both assent to it, but the latter person's non-cognitive attitudes towards the proposition punctuate the significance of the proposition for her while it may lack any significance entirely for the former person.

---

[36] These correspond loosely with the terms "fabula" and "sujet" in narratology. See Paul Cobley (2005) p. 678.
[37] I should note that it is possible that a person's articulable narrative can be constituted in part by non-n-facts. In other words, one may have narrative beliefs that fail to latch onto the world in the right way entirely.
[38] I don't think that calling them beliefs is misleading, because one's reflection on one's own narrative is always a reflection on beliefs about the phenomenology of events, states of mind, happenings, or states of affairs that makes up the narrative constituent parts. Thus, I think to call these all beliefs is accurate.

These distinctions allow us to see the relationship between understanding and the narrative self. First, one's articulable narrative can be constituted by all true n-beliefs, but one's attitudes towards those beliefs will make them more or less salient, such that one's non-cognitive attitudes towards the various narrative beliefs produces the structure of the articulable narrative. The articulable narrative, then, just is a mental representation of some set of n-beliefs that have various levels of saliency, such that the topography of the representation is determined by the saliency of the constituting n–beliefs. Those n-beliefs that are *minimally* salient are those salient *enough* to enter into one's deliberation when one reflects on her narrative. All of this is just to say that the saliency of the n-beliefs is dependent in large part on our attitude towards them. Those n-beliefs and n-relations that we "feel" to be important, or significant in our deliberation, are what constitute our *deliberation manifold*. And this in turn constitutes one's moral understanding, i.e., our articulable narratives are representations of how we as moral actors relate to the world around us. Notice that there are many n-beliefs that have nothing to do with one's *moral* relationship to others or the world. I am interested in this paper only in those n-beliefs that are important to how we understand ourselves as moral agents. Thus, when a person deliberates, she will use the most salient n-beliefs as guides. In other words, the articulable narrative just is the set of n-beliefs and n-relations after they have been adjusted to match the non-cognitive attitudes S has towards her n-beliefs and n-relations in some context C. And this constitutes one's moral understanding.[39]

The upshot of the foregoing is that the availability of an n-belief, as available for being a part of the deliberation manifold, is largely a matter of non-cognitive significance. So S might

[39] Notice, this may constitute much more than a person's moral understanding. But, on this account, whether there is more that it might constitute, it is at least sufficient to constitute moral understanding.

have all of the relevant n-beliefs but just not find many of them as salient as she should.[40] In other words, while S may have all of the relevant n-beliefs she can have a problematic deliberation manifold. Call this a *distortion*. Thus, she would be working from and making decisions based on a limited subset of her n-beliefs. Distortions in salience like these have central moral significance, since they will profoundly affect our choices and actions. Thus, if our narrative becomes distorted by the *over* or *under salience* of certain beliefs, our moral decision making could be significantly impaired. This is because over or under salient beliefs can function to revise our *telos*, motivating action based on impaired or distorted understanding. For, when the right n-beliefs fail to have the appropriate salience, then I might reduce my ability to manipulate the articulable narrative. Thus, I could have a perfect fitness of n-beliefs and still have a noncognitive distortion of my articulable narrative. The foregoing suggests that I could *understand* myself either *more* or *less,* holding all of my n-beliefs equal. And ultimately, my moral decisions will result from my deliberation manifold in the articulable narrative.[41]

Moral judgment can then be cashed out in a kind of hybrid cognitivism: moral judgment so understood is, or is wholly constituted by, n-beliefs, where our actions are based on the salient set of n-beliefs that we call the deliberation manifold. So, there is indeed a non-contingent relation between judgment and moral motivation. For, the plausible Defeasibility Condition for moral motivation endorsed by this view is constituted by distortions in our understanding. In

[40] Some have worried that I am smuggling in normative language here. This 'should', in my mind, is compatible with any first order normative theory of morality. I do assume that there are some *oughts* and *should*s, but I am not attempting to make any statements on what they are.

[41] The relationship between the narrative self and the decisions made by agents is a reciprocal one. The narrative self both informs and is informed by the decisions agents make. My articulable narrative informs, at least partly, my future decisions. Obviously there are other features that bear on the decisions that we make—physiological needs and motivations, pressure from outside sources, and so forth. In some cases these can be defeaters from the narrative self, overwhelming the narrative self in some ways. But my narrative is reciprocally shaped by my choices (e.g., *I am* the kind of dad who plays with his children when he gets home from work, so I will play with my kids when I get home from work). Thus decisions that I make become normatively salient as they will in part determine the shape that my narrative takes. It is important to note that the narrative formation process that I am describing here is not generally a conscious one, rather the narrative formation process takes place as we make choices, and thus we are morally culpable for our choices.

other words, one can have certain judgments that fail to be minimally salient (i.e., salient enough), such that we never act on them. Thus, we arrive at a view of internalism that has been modified to incorporate the account of moral psychology as set out in the foregoing:

Narrative Internalism: If one judges it right to Φ, where the judgment is minimally salient in the articulable narrative, she will be motivated to Φ.

This has all been rather complicated, so to avoid misunderstanding here is a more formal statement of the components of Narrative Internalism.

P1 Understanding is constituted by having (i) a malleable mental representation and (ii) the ability to manipulate the mental representation.

P2 The self is constituted by a unique narrative that is more or less intelligible, and this corresponds to the base narrative.

P3 The articulable narrative is the narrative formed by bringing some subset of all the available n-facts (true or false) into a representation.

P4 The articulable narrative has both cognitive and non-cognitive components that together provide the motivating and justifying force underlying moral judgment.

P5 Distortions in one's moral understanding are what constitute the plausible Defeasibility Conditions for moral motivation.

Accounting for Satan's Amoralism

Now consider the case of Satan again. We have stipulated that he has all of the right relevant beliefs (including that rebellion is wrong), but still he chooses to rebel. Now, here is a retelling of Satan's story as recast through Narrative Internalism:

Satan was a grand angel, the son of the morning, who had privileged place and

> knowledge. He rightly recognized his own greatness, his own power, and his own glory. And while he also recognized God's greatness, power, and glory, the salience of his various n-beliefs shifted such that his articulable narrative was centrally organized around his own greatness, power, and glory (a sort of gestalt shift catalyzed by his fixation his own greatness, power, and glory). In other words, these features of his narrative became so salient that his understanding became distorted, and his *telos* was revised. Call this kind of distortion in the articulable narrative *Pride*. Thus, he rebelled against God, all the while knowing that it was wrong to rebel, yet finding no motivation not to rebel, because the relevant judgments were marginal in his articulable narrative, and thus not in his deliberation manifold. Therefore, because of Pride Satan rebelled against God.

As the proverb says, "pride came before the fall." Satan allowed Pride, a distortion in his articulable narrative, to direct his choices such that he was willing to rebel against God. Further, because Narrative Internalism can answer the hardest case, the case of Satan, it can generalize across all cases of Amoralism, thus I conclude that Amoralism is reconcilable with Narrative Internalism generally.

Narrative Internalism has explanatory power and fills the gaps other internalists leave open. For, given the platitude that human beings are *feeling* creatures that have various attitudes towards their moral judgments, we can explain why it is that some person might judge that is right to Φ (she has a n-belief that it is right to Φ), yet fail to be motivated to Φ (her n-belief that it is wrong to Φ is not a part of her deliberation manifold). Consider two kinds of cases of Amoralism: first, under my account, strong emotions, like anger, can affect the saliency of various beliefs such that the articulable narrative becomes distorted, which may lead one to act against her better judgment. For example, when a person becomes angry they will often act

against their better judgment.

Second, notice that strong emotions are not at all necessary for one's articulable narrative to become distorted. One could be *lacking* in the appropriate noncognitive attitude she has towards certain n-beliefs, such that certain judgments never rise to the level of being a part of the deliberation manifold. For example, I might judge that it would be wrong for me to beat anteaters while wearing a bear suit. Still, this judgment lacks saliency because I have no significant noncognitive attitudes towards my various beliefs about anteaters or about the wearing of bear suits. Thus, unless I was deeply reflective and introspective, or something else happened to change the saliency of my beliefs about anteaters and bear suits, my judgment that it is wrong for me to beat anteaters while wearing a bear suit will most likely never enter into my deliberation about what I *will* do. It is simply not a part of my deliberation manifold.

Conclusion

I have argued that in order for judgment internalism to be viable in reconciling judgment internalism and Amoralism, it must provide plausible accounts of both (a) the relationship between judging and motivation, and (b) the conditions for defeasibility. Narrative Internalism is such an account. This account has the dual strength that it can account for both why one would typically be motivated to Φ upon judging that it is right to Φ and also the conditions that might obtain such that one would fail to be motivated. This account of moral psychology explains both (a) the relationship between judging and motivation, and (b) the conditions for defeasibility by giving an account of plausible Defeasibility Conditions qua narrative distortion. I conclude that unless there are more plausible accounts of judgment internalism in the offing, which doesn't seem apparent to me, we should adopt Narrative Internalism.

References

Anscombe, Gertrude E. M. *Intention*. Harvard University Press, 2000.

Blackburn, Simon. *Ruling passions*. Oxford: Clarendon Press, 1998.

Brink, David O. "Externalist moral realism." *The Southern Journal of Philosophy.* 24.1 (1986): 23-41.

Bruner, Jerome S. *Acts of Meaning*. Harvard University Press, 1990.

Campbell, Richmond. "What Is Moral Judgment?" *The Journal of Philosophy.* 104.7 (2007): 321-349.

Carey, John. "Milton's Satan." *The Cambridge Companion to Milton.* ED. Dennis Danielson. New York: Cambridge University Press, 1999. 160-74.

Cobley, Paul. "Narratology." *The Johns Hopkins Guide to Literary Theory and Criticism, 2nd ed.* Eds. Michael Groden, et al. Baltimore: John Hopkins University Press, 2005.

Darwall, Stephen. "Reasons, motives, and the demands of morality: An introduction." *Moral Discourse and Practice: Some Philosophical Approaches.* Eds. Stephen Darwall, Allan Gibbard, and Peter Railton. New York: Oxford University Press, 1997. 305-312.

Davidson, Donald. *Essays on Actions and Events*. New York: Oxford University Press, 2001.

Foucault, Michel. "The Subject and Power." *Critical Inquiry.* 8.4 (1982): 777-795.

Frankfurt, Harry G. "Freedom of the Will and the Concept of a Person." The *Journal of Philosophy.* 68.1 (1971): 5-20.

Hardy, Barbara. "Towards a Poetics of Fiction: An Approach through Narrative." In *Novel: A Forum on Fiction*, 2.1 (1968): 5-14.

Lackey, Jennifer. "Testimonial Knowledge and Transmission." *The Philosophical Quarterly.* 49.197 (1999): 471-490.

Lenman, James. "The Externalist and the Amoralist." *Philosophia.* 27.3 (1999): 441-457.

MacIntyre, Alasdair C. *After Virtue*. Notre Dame, IN: University of Notre Dame Press, 1984.

Mackie, J. L. "The Subjectivity of Values." *Essays on Moral Realism.* Ed. Geoffrey Sayre-McCord. New York: Cornell University Press, 1988.

Milton, John. *Paradise Lost*. New York: Barnes & Noble Books, 2004.

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

Olson, Jonas. "In Defense of Moral Error Theory." *New Waves in Metaethics.* Ed. Michael Brady. Palgrave Macmillan, 2010. 62-84.

Rosati, Connie S. "Moral Motivation." *Stanford Encyclopedia of Philosophy*. (2006): n.p.

Schechtman, Marya. *The Constitution of Selves*. New York: Cornell University Press, 2007.

Schroeder, Mark. *Noncognitivism in Ethics*. New York: Routledge, 2010.

Shafer-Landau, Russ. *Moral Realism: A Defense*. Oxford: Clarendon Press, 2003.

Smith, Michael. *The Moral Problem*. Oxford: Blackwell, 1995.

Stocker, Michael. "Desiring the Bad: An Essay in Moral Psychology." *The Journal of Philosophy* 76.12 (1979): 738-753.

Stroud, Sarah. "Weakness of Will." *Stanford Encyclopedia of Philosophy* (2008): n.p.

Van Roojen, Mark. "Moral Rationalism and Rational Amoralism." *Ethics* 120.3 (2010): 495-525.

Velleman, J. David. "Self to Self." *The Philosophical Review* 105.1 (1996): 39-76.

Wiggins, David. "Weakness of Will, Commensurability, and the Objects of Deliberation and Desire." *Proceedings of the Aristotelian Society* 79.The Aristotelian Society; Blackwell Publishing, 1978. 251-277.

Wilkenfeld, Daniel A. "Understanding as Representation Manipulability."*Synthese* 190.6 (2013): 997-1016.

Williams, Bernard. “A Critique of Utilitarianism.” *Utilitarianism: For and Against.* Eds. J. J. C. Smart and Bernard Williams. Cambridge University Press, 1973.

Zagzebski, Linda. "Recovering Understanding." *Knowledge, Truth and Duty. Essays on Epistemic Justification, Responsibility, and Virtue. Ed.* Matthias Steup. Oxford University Press, 2001. 235-252.