Philosophy and
Phenomenological Research

Philosophy and Phenomenological Research
Vol. LXXIX No. 3, November 2009

The Duty of Self-Knowledge

OWEN WARE
University of Toronto

Kant is well known for claiming that we can never really know our true moral disposition. He is less well known for claiming that the injunction "Know Yourself" is the basis of all self-regarding duties. Taken together, these two claims seem contradictory. My aim in this paper is to show how they can be reconciled. I first address the question of whether the duty of self-knowledge is logically coherent (§1). I then examine some of the practical problems surrounding the duty, notably, self-deception (§2). Finding none of Kant's solutions to the problem of self-deception satisfactory, I conclude by defending a Kantian account of self-knowledge based on his theory of conscience (§3).

> The depths of the human heart are unfathomable.
> Kant—*Metaphysics of Morals*

One of the most striking features of Kant's moral epistemology[1] is his claim that one lacks direct cognitive access to the ground or pur-

[1] I will quote and refer to Kant's texts parenthetically using the following translations:
AP: *Anthropologie in pragmatischer Hinsicht* (1798), trans. Robert B. Louden, *Lectures on Pedagogy in Anthropology, History, and Education*, ed. Günter Zöller & Robert B. Louden (Cambridge: Cambridge University Press, 2007).
G: *Grundlegung zur Metaphysik der Sitten* (1785), trans. Mary K. Gregor, *Groundwork of the Metaphysics of Morals* in *Immanuel Kant: Practical Philosophy*, ed. Allen Wood (Cambridge: Cambridge University Press, 2006).
KpV: *Kritik der praktischen Vernunft* (1788), trans. Mary K. Gregor, *Critique of Practical Reason* in *Immanuel Kant: Practical Philosophy*, ed. Allen Wood (Cambridge: Cambridge University Press, 2006).
KrV: *Kritik der reinen Vernunft* ([A] Edition: 1781/ [B] Edition: 1787), trans. Allen Wood & Paul Guyer, *Critique of Pure Reason* (Cambridge: Cambridge University Press, 1998).
LE: *Lectures on Ethics* (1764-1794), trans. Peter Heath, ed. J.B. Schneewind & Peter Heath (Cambridge: Cambridge University Press, 1997).
MS: *Die Metaphysik der Sitten* (1797), trans. Mary K. Gregor, *Metaphysics of Morals* in *Immanuel Kant: Practical Philosophy*, ed. Allen Wood (Cambridge: Cambridge University Press, 2006).

ity of one's disposition.[2] This is what I will call the Opacity Thesis. Versions of the thesis play a crucial role in nearly every aspect of his thought from the early 1770s to the late 1790s. And yet—in his last major contribution to ethics—Kant advances what appears to be an opposing claim: that one is under obligation to *know oneself* (*MS* 6:441).[3] How can we reconcile these apparently inconsistent views? How can Kant affirm that self-knowledge is the first command of all duties to oneself but that direct cognition of the self is, strictly speaking, impossible?

Kant often points out that establishing the logical possibility of a concept is not sufficient for accepting its objective reality, whether it has a practical function in moral life.[4] In the present context we first need to determine whether the concept of self-knowledge is logically acceptable as a duty. This will form the larger part of my task in section 1. The difficulty here is that Kant distinguishes between generic self-knowledge and particular self-knowledge, and it is not clear which of these is essential to the command "Know Yourself." In section 2 my task will be to determine whether the duty of self-knowledge has a practical function in morality, and whether that function is positive or negative. For Kant, the obligation to know myself is a matter of knowing my moral character and progress toward the good. He insists,

---

MpVT: "Über das Mißlingen aller philosophischen Versuche in der Theodicee" (1791), trans. George Di Giovanni, "On the Miscarriage of all Philosophical Trials in Theodicy," in *Immanuel Kant: Religion and Rational Theology*, ed. Allen Wood & George Di Giovanni (Cambridge: Cambridge University Press, 2005).
PR: *Vorlesungen über die philosophische Religionslehre* (1817), trans. Allen Wood, *Lectures on the Philosophical Doctrine of Religion* in *Immanuel Kant: Religion and Rational Theology*, ed. Allen Wood & George Di Giovanni (Cambridge: Cambridge University Press, 2005).
R: *Die Religion innerhalb der Grenzen der bloßen Vernunft* (1793), trans. Allen Wood, *Religion within the Boundaries of Mere Reason* in *Immanuel Kant: Religion and Rational Theology*, ed. Allen Wood & George Di Giovanni (Cambridge: Cambridge University Press, 2005).
Unless otherwise noted, I will cite from what is known as the Academy Edition of *Kant's Gesammelte Schriften* (Berlin: Walter de Gruyter & Co., 1902), included in the margin of the Cambridge University Press translations of Kant.

2 For examples of the Opacity Thesis in Kant's moral theory, see: *G* 4:407; cf. *R* 6:51/71; cf. *R* 6:71/87-88; cf. *MS* 6:447.

3 Kant's formulation of the duty of self-knowledge plays a central role in his early lectures on ethics. The Herder notes show that Kant worked on this theme as early as 1762 (*LE* 27:43). In the Collins notes of 1784-1785, Kant speaks of "self-testing" and "self-examination" as "the primary duty to oneself" and disposing the agent so that "he may be capable of observing all moral duties" (*LE* 27:348). (See my criticism of Denis, note 12).

4 Variations of this claim run throughout all three of Kant's critiques. For one example, see the footnote from section XXVI in the B-edition of the *Critique of Pure Reason*.

however, that I can't have introspective certainty of my character or progress—for two reasons. First, my disposition is beyond the limits of immediate consciousness. I can't simply cognize my underlying character by "observing myself." Second, and more seriously, I am prone to deceive myself. I may be convinced on a psychological level that I acted virtuously, or that I reoriented my life to the good. But my psychological confidence may be entirely mistaken. In truth I may have acted selfishly. And my disposition may be the same as before.

In response to such difficulties Kant tentatively proposes a theory of inferential self-knowledge. If I have genuinely resolved to reorient my character to the good, Kant claims, that resolve should be visible in my life conduct. I should be able to infer the moral status of my disposition, whether restored or unrestored, by way of my *actions*. The problem with this theory is that actions are not judgment-neutral. We still need to evaluate, assess, and appraise our own actions, which forces us back to the problem of self-deception. Who's to say I'm a legitimate judge of my own life conduct? How can I possibly examine my actions sincerely and impartially? I could easily turn a blind eye to my past exploits and thereby construct a false conception of myself. After all, it's within my interest to judge myself in a morally flattering or forgiving light. The question here, then, is whether inferential self-knowledge can ever be free from the threat of self-deception. For the remainder of section 2, I will outline and assess Kant's solutions to this threat. Finding none of his solutions entirely satisfactory, I will present my own account along Kantian lines in section 3.

Preliminaries: Kant's Opacity Thesis

As noted, the limits Kant places on self-knowledge are rather strict and wide-ranging. Not only does he limit the knowledge we can have of others, he also limits the knowledge we can have of ourselves. "Indeed," Kant writes, "even a human being's inner experience of himself does not allow him so to fathom the depths of his heart as to be able to attain, through self-observation, an entirely reliable cognition of the *basis* of the maxims which he professes, and of their *purity* and *stability*" (*R* 6:63–my emphasis). I cannot know, for example, whether my particular actions arise from conformity with the moral law or from some hidden self-interest. As Kant argues in Section II of the *Groundwork of the Metaphysics of Morals*, "it is absolutely impossible by means of experience to make out with complete certainty a single case in which the aim of an action otherwise in conformity with duty rested simply in moral grounds"

(*G* 4:407). For even when I think—and Kant believes we have a tendency for such thoughts—that I am bending my actions to the strict and noble commands of duty, I may, on further reflection, perceive my actions arising from the "dear self," which Kant notes "is always turning up" (*G* 4:407).

The simplest version of the Opacity Thesis is that the ground of my maxims lies beyond the reach of cognition. I am opaque to myself to the extent that I can never know my disposition immediately by way of introspection. What Kant is rejecting here is the idea that I can catch a glimpse of my true character by way of some intuition, as when Hawthorne speaks of "one of those moments—which sometimes occur only at the interval of years—when a man's moral aspect is faithfully revealed to his mind's eye."[5] Kant's discussion from *Groundwork II* also introduces a second type of opacity. I can never be certain of the moral purity of my actions because of my deeply selfish nature. I can never be entirely confident that "no covert impulse of self-love" determined what I thought was my self-sacrificing deed. Kant observes that "we like to flatter ourselves by falsely attributing to ourselves a nobler motive, whereas in fact we can never, even by the most strenuous self-examination, get entirely behind our covert incentives" (*G* 4:407).[6] We are therefore opaque to ourselves in two distinct senses. First, the ground of our maxims is opaque because *unknowable*, i.e., it simply falls outside our epistemic reach. This is what I will call Type-1 opacity.[7] Second, the purity of our maxims is opaque because *covert*,

---

5 Nathaniel Hawthorne, *The Scarlet Letter: A Romance* (Ohio: Ohio University Press, Penguin Books, 2003): p. 150.

6 Kant's idea of the "dear self" suggests an asymmetry between our self-knowledge of virtuous and vicious motives. In general, he believes we reach the limits of introspection with respect to what we imagine is a *good* motive behind our deed, because he believes everyone is prone to inflate the moral value of his or her actions. On the other hand, he does not believe our *bad* motives are opaque to us, implying that one can't possibly act on a covertly virtuous motive. I am grateful to an anonymous reviewer of *PPR* for drawing my attention to this asymmetry more clearly.

7 The Type-1 version of the Opacity Thesis has its origins in Kant's critique of self-knowledge from the Paralogisms of the first *Critique*. In that text he argues against the idealist view that one's internal thoughts provide indubitable certainty of one's status as a thinking substance—e.g., Descartes' "I am." Kant points out that the empirical knowledge we have of ourselves is no more, and no less, reliable than the empirical knowledge we have of objects in the world. Objective self-knowledge is therefore beyond the scope of theoretical reason; and that's because introspection only ever reveals appearances of inner sense, just as sensation only ever reveals appearances of outer sense. See, for instance, *KrV* A371. Unfortunately, a fuller discussion on how Kant's version of the Opacity Thesis in the first *Critique* relates to versions found in his later writings falls outside the scope of the present discussion.

i.e., we can never know for certain whether our motives have been corrupted by other, less praiseworthy, motives. This is Type-2 opacity.[8]

It will soon become apparent that the greater threat to the duty of self-knowledge is not the fact that I'm incapable of perceiving myself directly, whatever that may mean. The greater threat, rather, is self-deception or Type-2 opacity. For even if I avoid all forms of introspection by appraising my actions alone, I still need to judge and draw inferences from my actions. But as long as my authority as *self-judge* is in doubt, the possibility of moral self-knowledge will remain uncertain.[9]

1. The Duty of Self-knowledge

In view of the restrictions Kant places on self-knowledge, how are we to understand his claim that moral self-cognition is "the **First Command** of all Duties to Oneself" (*MS* 6:441)? As he tells us in the *Metaphysics of Morals*:

---

[8] These two types of opacity are related, although I am not committing myself to any further explanation their relation. Roughly, we can say the phenomenon of self-deception is only possible on the condition that I can never have objective knowledge of myself. In this sense, Type-2 opacity is dependent on the Type-1 variety. The nature of this dependency is, of course, a mystery for Kant, since he believes the root of deception (like the root of radical evil more generally) is inscrutable.

[9] The tension between the duty of self-knowledge and Kant's Opacity Thesis often remains a peripheral issue in the philosophical literature. For example, Lara Denis's recent study *Moral Self-Regard: Duties to Oneself in Kant's Moral Theory* (London/New York: Routledge, 2001), devotes only half a page to the duty of self-knowledge. Two notable exceptions are Onora O'Neill's "Kant's Virtues," in *How Should One Live? Essays on the Virtues*, ed. Roger Crisp (Oxford: Clarendon Press, 1996): pp. 77-98; and, more recently, Jeanine Grenberg's *Kant and the Ethics of Humanity: A Story of Dependence, Corruption, and Virtue* (Cambridge: Cambridge University Press, 2005). However, I think O'Neill and Grenberg misrepresent the duty in terms of its "prescriptive efficacy." For Kant, moral self-knowledge does not provide us with guiding knowledge of *how to act*. And so, in this sense, our duty is not to know, even minimally and humbly, the objective correlation between our actions and the principles upon which we think we act. As we will see, our duty is to examine our character, the ground of our maxims, and so to determine whether we have formed our moral judgments with due care.

Other places where expositors of Kant's views note both the duty of self-knowledge and the difficulty of self-knowledge are: Roger Sullivan, *Immanuel Kant's Moral Theory* (Cambridge: Cambridge University Press, 1989): pp. 60-62; Allen Wood, *Kant's Ethical Thought* (Cambridge: Cambridge University Press, 1999): pp. 196-202; Paul Guyer, *Kant on Freedom, Law and Happiness* (Cambridge: Cambridge University Press, 2000): pp. 384-385; Jacobs and Kain (eds.), *Essays on Kant's Anthropology* (Cambridge: Cambridge University Press, 2003), in the essays by Wood (pp. 48-50) and Jacobs (pp. 110-111; 120-129); and Patrick R. Frierson, *Freedom and Anthropology in Kant's Moral Philosophy* (Cambridge: Cambridge University Press, 2003): pp. 100-103.

> This command is "*know* (scrutinize, fathom) *yourself*," not in terms of your natural perfection (your fitness or unfitness for all sorts of discretionary or even commanded ends) but rather in terms of your moral perfection in relation to your duty. That is, know your heart—whether it is good or evil, whether the source of your actions is pure or impure, and what can be imputed to you as belonging originally to the *substance* of a human being or as derived (acquired or developed) and belonging to your moral *condition*. (*MS* 6:441)

Kant speaks of natural self-knowledge, knowledge of my "natural perfection," and moral self-knowledge, knowledge of "my heart." To complicate matters, he divides moral self-knowledge into two kinds: what is imputable to my moral condition either (i) substantially or (ii) derivatively. What Kant may be saying is this: I need to know what is good and evil of myself that is also good and evil of everyone, generically. Substantial self-knowledge would then be knowledge of the good and evil imputable to me qua member of the human species (which I am responsible for, nonetheless). Derived self-knowledge would be, in turn, knowledge of the good and evil imputable to me qua individual, knowledge of my own idiosyncratic habits, propensities, and tendencies.

On top of these divisions Kant makes the dramatic assertion that "only the descent into the hell of self-cognition can pave the way to godliness" (*MS* 6:441). But given the different kinds of self-cognition outlined above, we must ask: Which hell? Must I descend into the hell of cognizing my generic moral condition, the good and evil I have in common with the human species? Or must I descend into the hell of cognizing my idiosyncratic moral condition, the good and evil I have cultivated within myself? The question we have to face here is what *kind* of moral self-cognition is essential to the duty of self-knowledge.

1.1. Generic Self-Knowledge

To begin with, we should ask why Kant rejects knowledge of natural perfections from the duty of self-knowledge.[10] The rejection is instructive, because it tells us not every object of self-knowledge is basic to the first self-regarding duty. To repeat, Kant argues that the duty of

[10] In the *Metaphysics of Morals*, Kant divides all self-regarding duties into *perfect* and *imperfect* duties. Perfect duties are what he calls "*limiting* (negative) duties" (*MS* 6:419), which place specific constraints on our actions, either physically (e.g., against self-mutilation and suicide) or morally (e.g., against lying, avarice). On the other hand, imperfect duties are what Kant calls "*widening* (positive duties to oneself)" (*MS* 6:419), which place us under obligation to adopt choices as ends, specifically, the end of self-perfection. For Kant, the duty of self-knowledge is neither strictly perfect nor imperfect but rather conditions the possibility of *both*, which is why, to repeat, he describes the duty of self-knowledge as "the **First Command** of all Duties to Oneself" (*MS* 6:441).

self-knowledge is not a matter of knowing my "fitness or unfitness for all sorts of discretionary or even commanded ends" (*MS* 6:441). I take it this is because natural perfections, or "gifts of nature" as Kant calls them, lack intrinsic moral worth. I may have a naturally courageous character, but if my heart is corrupt I could employ that gift to vicious ends, say, by performing evil deeds with a steady hand. Kant therefore excludes knowledge of natural perfections from the duty of self-knowledge because I must presumably know my heart (whether it is good or evil) *before* I can successfully fulfill the moral obligation of self-perfection.[11] The alternative is clear: The only suitable object of self-knowledge is the ground of my maxims, my enduring moral character. I need to know my *heart*, whether it is good or evil.[12]

This requirement raises a new set of difficulties. In his *Religion within the Boundaries of Mere Reason*, Kant argues that a human being adopts evil maxims in such a manner that "he expresses at the same time the *character of his species*" (*R* 6:21–my emphasis). And a little later he notes, "by the 'human being' of whom we say that he is good or evil by nature we are entitled to understand not individuals (for otherwise one human being could be assumed to be good, and another evil, by nature) but the whole species" (*R* 6:25). This doesn't mean, as Kant is careful to point out, that we can infer evil from the general concept of humanity (otherwise, evil would be a necessary quality of human nature). Nor is it an attempt to explain the ultimate origins of evil, which for Kant are inscrutable. His point, rather, is that we can "spare ourselves the formal proof that there must be such a corrupt propensity rooted in the human being" in light of the many examples of humanity's evil, both inside and outside the boundaries of the "civilized" world (*R* 6:32). However we decide to empirically survey human

---

11 A possible exception to this rule would be the case of the *good-hearted fool*, someone who is naturally well intentioned but lacks practical judgment. One could argue that such an individual would have an obligation to know his good nature so that he could properly align his intentions with his judgments. Thanks to Steve Engstrom for this suggestion.

12 Denis's passing comment that the duty of self-knowledge is imperfect (*Moral Self-Regard*, p. 115) seems wrong for two reasons. (1) Imperfect duties, such as natural perfections, structurally require the agent's moral self-cognition *first*. (2) If there were such a thing as an imperfect duty to know oneself, we could not specify what about the self one should know. As we will see, however, Kant argues that the duty of self-knowledge does require one to know specific aspects of one's heart, whether it is good or evil, with respect to one's generic and particular self-identity. I believe Kant's claim that the duty of self-knowledge is the first of all self-regarding duties must be taken seriously; it means that without moral self-cognition (i) one could not act out of respect for one's inner humanity, and (ii) one could not ethically pursue one's natural perfections of mind, body, and spirit. Hence, the duty of self-knowledge precedes and conditions the possibility of both perfect and imperfect duties.

nature, experience forces us to the opinion of its corrupt propensity, that it "cannot be judged otherwise" (*R* 6:32). Evil is thereby imputable to the substance of my moral condition, something that everyone brings upon his or herself, without exception.[13] Kant thus reaffirms the words of Sir Robert Walpole: "Every man has his price, for which he sells himself," to which he supplements Romans 3:9: "None is righteous… no, not one" (*R* 6:38).

A curious implication now comes to view. If the notion we have of humanity's corruption arises necessarily in experience, something we perceive "in every human being, even the best" (*R* 6:32), then a duty to know it would be vacuous. This would stand in conflict with Kant's general position that the concept of a duty relates only to what is "*entirely beyond the limits of our experience*" (*MS* 6:444–my emphasis). A duty is intelligible only in terms of what I *should* do, in this case, what I *should* know, as opposed to what I *already* know (or do). For example, Kant rejects the duty of the agent's own happiness because it is an end that "everyone already wants unavoidably" by virtue of his or her sensible nature. "Hence it is self-contradictory," Kant argues, "to say that he is *under obligation* to promote his own happiness with all his powers" (*MS* 6:286; cf. *R* 6:7n).[14] By extension, it is vacuous to place me under obligation to know something that by virtue of experience I know or will know easily enough. The conclusion we can draw here is that the duty of self-knowledge cannot require me to know my *generically evil* heart.

Does the duty of self-knowledge require me to cognize the *good* I share in common with the human species? Kant often expresses the view that while a human being may indeed be corrupt, the humanity *within him* is sublime. No matter how far we stray from the moral law, "there is one thing in our soul which, if we duly fix our eye on it, we cannot cease viewing with the highest wonder… And that is the original moral predisposition in us, as such" (*R* 6:49). In *Groundwork III*, Kant argues that even if

---

13 For those like Rousseau who believe that only civilization makes humans evil, Kant calls attention to the *vices of savagery* among human beings in the "so-called *state of nature*," from the "ritual murders of Tofoa, New Zealand, and the Navigator Islands" to the "perpetual war between the Arathapescaw Indians and the Dog Rib Indians" (*R* 6:33n). And for those who believe we can only cognize the goodness of human nature in its civilized state, Kant calls attention to the *vices of culture* that form the "long melancholy litany of charges against humankind" (*R* 6:32).

14 It is not clear, according to Kant's terminology, why the duty to happiness would be "self-contradictory." To weaken the claim, I will say that the concept of a duty to happiness (or the duty to know one's generically evil nature, for that matter) is simply *vacuous*. This accords with Kant's statement from the second *Critique* that a "command that everyone should seek to make himself happy would be *foolish*, for one never commands of someone what he unavoidably wants already" (*KpV* 5:37–my emphasis). Thanks to Arthur Ripstein for pointing this out to me.

someone is incapable of aligning his empirical will with the moral law, he is still aware of the normative pull the law has on him. The conception we have of our own freedom, as members of the world of understanding, informs our judgments of actions "as being such that they *ought to have been done even though they were not done*" (*G* 4:455). In the *Religion*, Kant adds to this insight the command that "we *ought* to become better human beings," a command, he says, that "resounds unabated in our souls" (*R* 6:45).

We may safely conclude, in light of such comments, that knowledge of my generically good nature is not difficult to attain. In fact, Kant goes so far as to claim that no person "accustomed to the use of reason" can fail to grasp the dignity of the moral law, "not even the most hardened scoundrel… who, when one sets before him examples of honesty of purpose, of steadfastness in following good maxims, of sympathy and general benevolence (even combined with great sacrifices of advantage and comfort), does not wish that he might also be so disposed" (*G* 4:454). The scoundrel perceives in the display of virtuous actions what his corrupt will *ought* to be like. Even *he* is conscious of his moral vocation, if only faintly. And so is anyone else who fulfills his duty on the most trivial level. An individual finds moral support in his dutiful action, Kant maintains, "by the consciousness that he has maintained humanity in its proper dignity in his own person," even if his action consisted of abstaining from telling a harmless lie (*KpV* 5:88). I can therefore attain knowledge of my generically good nature by any number of means, say, by perceiving examples of virtuous action or by refraining from petty transgressions. But clearly I can't be under moral obligation to know what, in Kant's own analysis, is intellectually obvious—even to scoundrels.[15]

There are two other possibilities in which we can make sense of the claim that I have a duty to know my moral vocation. One is that knowledge of the opposite, the evil imputable to the substance of a human being, can easily lead me to loathe humanity. I therefore have a duty to become aware of the noble predisposition to the good within myself, for otherwise I could become overwhelmed by the impression of humanity's evil that experience forces upon me. Kant seems to have this point in mind when he notes:

---

[15] The reason for this is that any "action of integrity done with steadfast soul," Kant writes, "elevates the soul and awakens a wish to be able to act in like manner oneself" (*G* 4:411n). Kant goes so far as to claim that the sublimity of actions performed freely out of duty, without the faintest mixture of non-moral incentives, is so easy to cognize that even children of moderate age are impressed by their unconditional moral worth (*G* 4:411n; cf. *R* 6:48).

> This moral cognition of oneself will, first, dispel *fanatical* contempt for oneself as a human being (for the whole human race), since this contradicts itself.—It is only through the noble predisposition to the good in us, which makes the human being worthy of respect, that one can find one who acts contrary to it contemptible (the human being himself, but not the humanity in him). (*MS* 6:441)

Kant observes with sad irony that "our species, on closer acquaintance, is not particularly lovable," but "*hatred of them* is always *hateful*" (*MS* 6:402). Contempt for the whole human race commits the error of judging the human being solely on the grounds of his corrupt empirical will, ignoring his noble personality, which is predisposed to the good. On this reading, knowing the good I share universally with others can be an effective antidote to what Kant elsewhere calls "another vice, namely that of misanthropy" (*R* 6:34).

The second possibility, which I will only outline briefly, is that self-knowledge of my generic moral vocation could possibly strengthen my feeling for the moral law, which would strengthen my desire for self-improvement. In a famous footnote to Schiller from the *Religion*, Kant remarks that the "majesty of the law… rouses a *feeling of the sublimity* of our own vocation that enraptures us more than any beauty" (*R* 6:23n). Later in the same text, he writes:

> Often to arouse this feeling of the sublimity of our moral vocation is especially praiseworthy as a means of awakening moral dispositions, since it directly counters the innate propensity to pervert the incentives in the maxims of our power of choice. Thus it works, in the unconditional respect for the law which is the highest condition of all the maxims to be adopted… for the restoration to its purity of the predisposition in the human heart to the good. (*R* 6:50)

To avoid confusion, it is important to understand that Kant doesn't say we have a duty to acquire or attain a feeling for the moral law, since he considers moral feeling, along with conscience, love of one's neighbor, and respect for oneself, as "*subjective* conditions of receptiveness to the concept of duty," meaning that such feelings are constitutive of our moral agency as such (*MS* 6:399). One's duty regarding moral feeling is rather to "cultivate" and "sharpen one's attentiveness" to it; hence, one's duty is only indirect (*MS* 6:401). My speculation, in short, is that knowledge of our inner humanity could have something of a looping effect with our moral feeling for the law. Even if moral feeling constitutes our receptiveness to duty, cultivating our understanding of humanity's moral vocation could help incite that receptiveness, thus serving to keep up our spirits while we traverse the path of virtue. Reminding ourselves of our

dignity, and the dignity of others, could thereby enliven our resolve to improve ourselves.[16]

1.2. The Idiosyncratic Self

So far, I have offered a quick sketch of a possible Kantian theory of generic self-knowledge. I concluded in the first case that knowledge of my generically evil nature is vacuous as a duty. Experience itself will supply me with the knowledge that everyone, even the best, has a corrupt propensity. By contrast, self-knowledge of my generically good nature does seem to meet the requirements of duty. It functions negatively to counteract my self-loathing and misanthropic attitudes, and positively to shake off the lull of moral apathy. The more I reflect on my inner humanity, my original predisposition to the good, the more I feel inspired to engage in self-reform.

At least as it stands, the theory of generic self-knowledge remains susceptible to what we might call an existentialist critique. One could level the charge that knowing what is imputable to me qua member of the human species is existentially vacuous. Knowing that I am radically evil or predisposed to the good, for example, involves nothing more than knowing what I share in common with all others; it says nothing about the particulars of my moral condition, such as my motives, habits, traits, propensities, or disposition in general. Moreover, Kant doesn't seem to recognize that thinking of my good or evil qua member of the human species takes away the moral sting of imputability. Yes, I'm evil, and so is everyone else. I have a sublime moral vocation (which sounds like a lot of work), and so does everyone else. The impersonal character of generic self-knowledge, one could argue, weakens its normative pull.

A closer look at Kant's duty of self-knowledge will alleviate these concerns. Knowledge of my moral idiosyncrasies, for Kant, is essential to the duty of self-knowledge. My duty is to cognize not only what belongs "originally to the *substance* of a human being," but also what belongs to my moral *condition*, which is "derived (acquired or developed)" (*MS* 6:441). In the latter case I must ask myself, as one author writes, "why did I act that way? Am I a generally sympathetic person? Stingy or openhanded? Quick to anger or unduly self-effacing? What is most important to me? What attitudes, desires, and beliefs guide the

[16] Kant touches on this idea in a later text, when he tells us that "reason, in representing the morally good by connecting its ideas with intuitions (examples) that have been imputed to them, can produce an enlivening of the will (in spiritual or political speeches to the people, or even in solitary speeches to oneself)" (*AP* 7:254).

over-all structure of my character and actions? Most importantly, have I really placed the pursuit of moral ends above my pursuit of self-love?''[17] Questions like these force me to consider my unique position as a moral agent. I must seek particular knowledge of what kind of person I am, whether *my* heart is good or evil, and what kind of evils stand in the way of my moral restoration.

Though meaningful on an existential level, we must still ask whether the duty of idiosyncratic self-knowledge is vacuous or not. Does it really make sense to say I'm under obligation to know my particular attitudes, desires, beliefs, habits, actions, and intentions? Surely, I know these better than anyone else does. Or do I? Empirical surveys of human nature bring me to the opinion of humanity's evil, of which I recognize exists generically in myself. But with respect to my particular heart, I might be inclined to judge myself more leniently than I do the human species as a whole. One reason for this is that I am just a single person, capable of limited wrongdoing, but the list of humanity's misdeeds is of no comprehensible end. The human species brings to view more examples of evil than I could possibly produce on my own.

Another reason is that I have, or at least think I have, a privileged outlook regarding my own intentions, which I of course lack when judging others. Kant observes that people often produce morally forgiving self-conceptions, which make them less inclined to accept or acknowledge their own shortcomings. It's easy for me to take the sting out of moral imputation, say, by falsely attributing weakness of will to my maxims. For this reason alone, idiosyncratic self-knowledge passes the requirements of duty. I *ought* to know the particulars of my moral self, my own failures, shortcomings, misdeeds, and false self-conceptions—especially since I am less inclined to accept or acknowledge these particulars on my own.

Knowing my particular evils is, for Kant, a preliminary step to ''develop the original predisposition to a good will'' (*MS* 6:441). But this opens up a more difficult issue. While I can easily cognize the generic moral command that I *ought* to become a better person, how can I be certain my disposition has *actually* improved? How can I be certain of a genuine change toward the good in my particular heart? I've changed, I'm a ''new man''—but how can I really know this? The problem here is not the fact that my disposition is inscrutable (Type-1 opacity), but that my resolution to change for the good could be another manifestation of covert selfishness (Type-2 opacity). Kant is well aware that ''one is never more easily deceived than in what promotes a good

[17] Jeanine Grenberg, *Kant and the Ethics of Humility*, p. 226.

opinion of oneself'' (*R* 6:68). And what better opinion can one have than one's restitution to goodness, purity, and righteousness?

Consider the case of Dorian Gray. Gray remained convinced that his act of sparing a girl the shame of social scandal was evidence of his new moral character. Only when he perceived the visible stains of sin on his portrait, elsewhere described in the novel as his ''conscience'' and the ''mirror of his soul,'' did he come to realize the dark truth of the matter:

> Had it been merely vanity that made him do his one good deed? Or the desire for a new sensation, as Lord Henry had hinted, with his mocking laugh? Or that passion to act a part that sometimes makes us do things finer than we are ourselves? Or, perhaps, all of these?... Had there been nothing more in his renunciation than that? There had been something more. At least he thought so. But who could tell?... No. There had been nothing more. Through vanity he had spared her. In hypocrisy he had worn the mask of goodness. For curiosity's sake he had tried the denial of self. He recognized that now.[18]

Dorian Gray was psychologically certain of his moral restoration; he ''felt'' the change, so to speak. In truth, his disposition remained the same, and what he considered his one virtuous deed was nothing more than what Kant would call a bit of ''moral enthusiasm.'' Moral enthusiasts like Gray take up the commands of duty not with an attitude of sober resolve, but with a ''frivolous, high-flown, fantastic cast of mind,'' as if the execution of duty was something meritorious and a call for celebration (*KpV* 5:85).[19] Dorian Gray's disillusion is instructive: it shows we can't rely on introspection to determine the purity of our maxims or of our disposition in general.

The case of Dorian Gray also shows us that the greater threat to particular self-knowledge is not Type-1 opacity. For Kant, perfect self-knowledge is impossible to attain, but that in no way detracts from our duty to pursue moral self-cognition, however imperfect that cognition will be. No doubt, the greater threat is self-deception (Type-2 opacity), because it is essential that the agent appraise her moral worth

---

[18] Oscar Wilde, *The Picture of Dorian Gray* [1897] (New York: Random House, 2004): pp. 252-253.

[19] Moral enthusiasts thus mistake the motive of dutiful action in some empirical feeling rather than in the law itself. Someone who adopts the maxim of sympathy to others, for example, only perceives the moral worth of the maxim itself, and thereby fails to notice the pathology of his disposition. Kant therefore argues: ''Actions of others that are done with great sacrifice and for the sake of duty alone may indeed be praised by calling them noble and sublime deeds, but only insofar as there are traces suggesting that they were done wholly from respect for duty and not from ebullitions of feeling'' (*KpV* 5:85).

and progress sincerely. If self-deception can undermine the sincerity of the agent's inward appraisal, then the duty of particular self-knowledge would be, practically speaking, useless. It would be useless in the sense that the agent could never trust her own self-assessment, effectively destroying the reliability, and possibility, of her self-knowledge. So in light of Kant's assertion that "only the descent into the hell of self-cognition can pave the way to godliness," we must now ask: How can I ever be certain my ascent toward godliness is not, in truth, a plunge into moral enthusiasm?

The final, and most difficult, issue regarding the duty of self-knowledge therefore pertains to its objective reality, whether it has a practical function in moral life. I will organize the remainder of this discussion into a series of alternatives to introspective-based forms of cognition. These alternatives will emphasize two general areas of moral life that, for Kant, are free from the workings of the dear self: my long-term moral conduct and, more importantly, my *conscience*.

2. Self-knowledge in Moral Life

The tension between Kant's commitments to the practical necessity of restoration and to the opacity of the human heart is first visible in his argument for the immortality of the soul from the *Critique of Practical Reason*. The metaphysical end of his solution is to say that, empirically, our progression to the good is infinite, and that we can never rationally hope to achieve holiness of will, or what amounts to the same thing, complete conformity with the moral law (*KpV* 5:122). God, however, who stands outside of time, can intellectually comprehend the restoration of our intelligible character in full. From the standpoint of eternity, then, we have already achieved perfection of will; but within time, this perfection is of an endless duration (*R* 6:67).

The practical necessity of restoration, from which Kant draws his formula "*ought* implies *can*," does not address how I *can be certain* of my particular change of heart. The ought of the moral law only implies that I can formally achieve holiness of will; it does not imply I can immediately cognize the success of my aspiration to holiness. In light of these difficulties, Kant writes: "All that a creature can have with respect to hope for this share [in the highest good] is consciousness of his tried disposition, so that, from the progress he has already made from the worse to the morally better and from the immutable resolution he has thereby come to know, he may hope for a further uninterrupted continuance of this progress" (*KpV* 5:123). In a footnote to this passage, he offers the following caveat: "Conviction of the immutability of one's disposition in progress toward the good *seems,*

nevertheless, to be in itself impossible for a creature'' (*KpV* 5:123n–my emphasis). Once again, Kant's moral theory conflicts with the Opacity Thesis. Morality commands us to restore the original goodness of our heart, to place the law above self-love. But it seems we can never directly cognize this restoration within ourselves.

2.1. The Inferential View

This tension in Kant's thought reaches its peak in the *Religion*. Kant perceives the restored agent's need to have ''assurance of the reality and *constancy* of a disposition that always advances in goodness'' (*R* 6:67). I need to know that my change of heart is authentic and that it will not relapse into evil. But Kant finds the traditional responses to this desired assurance unsatisfactory, and for good reason. On one extreme end, there is the view that a supernatural power will sanctify my resolution if it is genuine (e.g., ''His Spirit gives witness to our spirit''). On the other end, there is the view that I can have no degree of assurance in my changed disposition, so that I must live in a state of ''*fear and trembling*'' (*R* 6:68). The former view contradicts the limits of human understanding and can quickly turn to a form of religious enthusiasm, whereby people claim to perceive the effects of grace in others or in themselves. The fear-and-trembling approach, if taken too far, leads to what Kant calls the ''darkest enthusiasm,'' by which he means a kind of obsessive self-scrutiny, such as we find in the diaries of Pascal and Haller (*AP* 7:133).

The Opacity Thesis no doubt has an intended element of fear and trembling in it, but Kant admits that ''without *any* confidence in the disposition once acquired, perseverance in it would hardly be possible'' (*R* 6:68). Hence, the concept of assurance has a negative, but indispensable, function in moral life. That is to say, it doesn't contribute directly to my restoration (because the concept presupposes such restoration has already taken place), but it gives me confidence that my moral progress is genuine, thereby motivating me to continue on the path of virtue.[20] What is at stake here is not *merely* my psychological

[20] The concept of assurance from the *Religion* is in many respects similar to Kant's discussion of contentment from the *Critique of Practical Reason*. He asks, ''Have we not, however, a word that does not denote enjoyment, as the word happiness does, but that nevertheless indicates a satisfaction with one's existence, an analogue of happiness that must necessarily accompany consciousness of virtue? Yes! This word is *contentment with oneself* [*Selbstzufriedenheit*], which in its strict meaning always designates only a negative satisfaction with one's existence, in which one is conscious of needing nothing'' (*KpV* 5:117). The psychological state of *Selbstzufriedenheit* (literally, ''self-satisfaction'') would be a matter of having what we call in colloquial terms a *clear conscience*, i.e., knowing one has done all that is within one's power. (See the Conclusion).

commitment to self-improvement. For without that commitment my motivation to become a better person would sink below the commands of reason, no matter how loudly those commands resound in my soul. Failing to account for assurance in our moral restoration, which itself gives one "hope of absolution," would therefore have a crippling effect on both reason and morality (*R* 6:76). It would lead us to what Kant calls a "feeling of hopelessness" and "wild despair" (*R* 6:71).

Kant begins to sketch a solution to these difficulties in the above-mentioned footnote from the *Critique of Practical Reason*, where he writes:

> [S]omeone who is aware of having persisted through a long portion of his life up to its end in progress to the better, and this from genuine moral motives, may very well have the comforting hope, *though not certitude*, that even in an existence continuing beyond this life he will persevere in these principles… [I]n this progress which, though it has to do with a goal endlessly postponed, yet holds for God as possession, he can have a prospect of a future *beatitude*. (*KpV* 5:123n–my emphasis)

In the *Religion*, Kant develops the first half of this proposition further, leaving questions of beatitude and grace to the *parerga* of rational religion (*R* 6:53). He argues that we can acquire confidence in our restored disposition "without delivering ourselves to the sweetness or the anxiety of enthusiasm, by comparing our life conduct so far pursued with the resolution we once embraced" (*R* 6:68). The process involves "observing ourselves through actions" (*LE* 27:365), or what I will call the Inferential View. It consists of loosely inferring our disposition, whether it is good or evil, from the moral character of our past actions:

> [Take] a human who, from the time of his adoption of the principles of the good and throughout a sufficiently long life hence-forth, has perceived the efficacy of these principles on what he does, i.e., on the conduct of his life as it steadily improves, and from that has cause to infer, but only by way of conjecture, a fundamental improvement in his disposition. (*R* 6:68)

Kant argues that an assessment of my life conduct from the time I've resolved to change for the good is sufficient to secure my confidence in the authenticity of that change. If the moral curve of my past actions shows steady and uninterrupted progress, I can envision a "boundless future" which is "desirable and happy" (*R* 6:69; cf. *PR* 28:1087). But if my past actions are inconsistently virtuous,

sometimes slipping back into evil, I can only envision a future of misery, one sinking deeper into corruption. Thus the image of a happy future, according to Kant, allows us to infer the authenticity of the agent's restoration. Even here, however, Kant does not for a moment compromise the Opacity Thesis. He quickly adds to his discussion the warning that we "cannot base this confidence upon an immediate consciousness of the immutability of our disposition since we cannot see through to the latter but must at best infer it from the consequence that it has on the conduct of our life" (*R* 6:71). Assurance in our restored disposition can never be a matter of introspective certainty.

By focusing on one's actions and moral conduct the Inferential View avoids the many difficulties we encountered with introspection. Action-based inferences, for example, are not dependent on the agent's often-deluded psychological states or self-conceptions ("I *feel* like a new person"; "I *consider* myself restored"). However, in the course of his argument Kant begins to detect a new set of problems with the Inferential View. The first is that an assessment of the agent's actions does not allow her to infer the *strength* or *stability* of her restoration (*R* 6:71). Someone may have a long history of good deeds behind her, but that does not allow us to conclude that if placed before a morally challenging situation, she would remain steadfast to her virtuous principles. The second problem, which Kant doesn't address, is that it's extremely difficult to say what time period is "sufficiently long" for the agent to legitimately infer a restored disposition from her actions. The example of Dorian Gray shows us that one virtuous deed does not suffice. But where are we to draw the line? Would a month of virtuous deeds suffice? Two years? Half a lifetime?

A related difficulty, which Kant hints at, is that if we do require a long history of "empirical proofs" of the agent's change of heart, say two years, then an agent's impending death would render the production of such proofs impossible. Imagine, for example, that I have undergone a genuine change of heart the same day I discover I will likely die within the week. According to the Inferential View even if I spent the rest of the week performing virtuous, self-sacrificing deeds that expressed the character of my renewed self, I would simply lack the sufficient time-period to substantiate an inference of renewal. The issue of impending death also brings us back to the first difficulty. Elsewhere, Kant addresses the scenario of an evildoer suddenly possessing an "honorable and upright disposition" upon his deathbed. Without certainty in his restored disposition, he won't be able to tell whether he would abandon his dignified attitude if, by some chance, death passed

him by.[21] Kant therefore advises, "a man always has to get to know himself in a gradual fashion" (*LE* 27:365). But then we must ask, what is "gradual"? which brings us back to the second difficulty of the Inferential View.

2.2. The Comparative View

Any one of the above-outlined difficulties threatens to undermine the Inferential View, once again leaving the allegedly restored agent without assurance in her change of heart. Aware of these difficulties Kant modifies his account of inferential self-knowledge in the closing paragraphs of Section One, Book Two of the *Religion*, which I will refer to as the Comparative View.[22] Instead of framing the question of the agent's assurance in terms of the future life she could hope to lead in view of her past actions (boundless happiness and continued improvement, or boundless misery and continued evil), the Comparative View frames the question in terms of the verdict the agent could hope to receive if her *whole life* were placed before a judge:

> [S]ince he can derive no certain and definite concept of his disposition through immediate consciousness but only from the conduct he has actually led in life, he shall not be able to think of any other condition of being delivered to the verdict of a future judge… than that *his whole life* be one day placed before the judge's eyes, and not just a segment of it, perhaps the last and to him still the most advantageous. (*R* 6:77)

---

[21] In general, Kant's views concerning the possibility of a change of heart later in one's life seem inconsistent. On some occasions he argues that such a change is hardly possible if one has led a long life of viciousness. See, for example, his *Über Paedagogik* (1803), trans. Robert B. Louden, *Lectures on Pedagogy* in *Anthropology, History, and Education*, ed. Günter Zöller & Robert B. Louden (Cambridge: Cambridge University Press, 2007): "[T]he human being who has always led a depraved life and wants to be converted in an instant cannot possibly get there, for it would be nothing short of a miracle for him to become in an instant the same as someone who has conducted himself well during his entire life and always thought upright thoughts" (*P* 9:488). On other occasions, Kant argues for what appears to be the opposite view: that only *after* one becomes weary of an unstable life led by instinct does one decide to ground one's character in reason. Kant even states that such revolutionary decisions are rare, even before the age of forty! (*AP* 7:294).

[22] Kant hints at a third alternative in one of the *Religion*'s footnotes, what we might call the Transformative View. If an agent has genuinely restored his disposition, so the argument runs, he will gladly take on any punishments attributable to his old disposition. Thus, the agent *transforms* the meaning of "punishments" into "so many opportunities to test and exercise his disposition for the good" (*R* 6:75n). Here the "proof" of the agent's restored disposition rests in his newfound pro-attitude toward the punitive consequences of his past deeds.

Recall that the Inferential View only takes into consideration the agent's moral conduct *after* her supposed restoration. One might think that assessing the agent's whole life, including her old disposition, would negatively affect the agent's hoped-for verdict. But this assessment can also have a positive import, since appraising the agent's life as a totality allows the hypothetical judge to perceive the actual change that has taken place within her. If I have resolved to change for the good, one way to assess my resolution would be to reflect on the differences between my present and past moral conduct. By considering the agent's old disposition, one could "examine what and how much of this disposition he has cast off, as well as the *quality* (whether pure or still impure) and the *grade* of the supposed new disposition for overcoming the old one and preventing relapse into it" (*R* 6:77). The more perceived difference there is between the two, the more readily I can infer a change of heart that is both authentic and unwavering.

The Comparative View thus overcomes the two general difficulties of the Inferential View. First, it allows us to infer the stability of the agent's restored disposition (by way of comparison with the old one), something we can't infer simply on the basis of improved moral conduct. Second, this view doesn't require a substantial time-period to legitimate the inference of the agent's restoration. While some duration of time is necessary for the agent to exhibit her new moral character, the inference is grounded, not in time-duration, but in the perceived difference between the agent's old and new ways of conduct. So impending death needn't throw one into despair. Or at least it ought not.

Nevertheless, one problem still threatens to undermine the Comparative View, and this concerns the nature of the hypothetical judge. If we can't establish the judge's authority, we have no reason to accept its final verdict. For Kant, there is a deep connection between how we *behave* before this judge and how we *represent* it. He maintains, for example, that representing the judge of the agent's whole life as *another* (i.e., God), "of whom news [of the agent's restoration] will be had through sources of information elsewhere," will have a detrimental effect on the agent's moral conduct. For then the accused

> will have much with which to counter the judge's severity under the pretext of human frailty; he will think he can get around him, whether by forestalling his punishment through remorseful self-inflicted torments that do not, however, originate in any genuine disposition toward improvement or by mollifying him with prayers and entreaties, even with incantations and self-proclaimed professions of faith. (*R* 6:77)

Simply by locating the judge outside the agent's consciousness, Kant believes the accused will attempt to assuage the judge's verdicts through an enthusiastic display of false piety and righteousness. On the other hand, if we represent the judge as *oneself*, what Kant calls the "judge within him," he believes the agent will thereby "pronounce a stern judgment upon himself, *for he cannot bribe his reason*" (*R* 6:77–my emphasis). Kant refers to the "judge within" as one's reason, but it is more precise to say it is one's *conscience*. As he writes in an earlier lecture, "conscience, that judge in us which is not to be bribed, will place before the eyes of each one the whole world of his earthly life and convince himself of the justice of the verdict" (*PR* 28:1087).

Still, introducing the concept of conscientious self-judgment raises more questions than it solves. We might first ask whether the notion of an "inner judge" is even intelligible, for how can I truly condemn myself? If I am responsible for issuing the verdict on my life as a whole, wouldn't I be tempted to deceive myself, to render my life acquitted even if an impartial judge would render me guilty? Intuitively, we often associate a judge's impartiality with his distance (both emotional and physical) from the accused. And yet, at the most crucial point in his argument, when the agent's assurance in her restoration is at stake, Kant seems to have fallen victim to an odd form of optimism. He has effectively entrusted the question of the agent's assurance in her own hands, so that she herself—and no one else—is responsible for judging her life. The Comparative View loses all legitimacy, however, if the agent can delude herself on the level of her own life assessment. Here, the issue we need to resolve is how conscience, or the "judge within," remains incorruptible by any form of deception (type-2 opacity).

3. Conscience: The "Inner Court"

I believe we can find the thread for a solution by moving ahead four years from the *Religion* to the *Metaphysics of Morals*, for only in the latter work does Kant offer a fuller treatment of the *inner judicial court* called conscience.[23] While a complete summary of Kant's discussion falls outside the scope of this paper, I would like to look at (A) the identity of conscience and (B) its judicial functions. In what follows I will suggest

[23] For lack of space, I have refrained from exploring the historical sources of Kant's idea of conscience. It is common to see this idea in a Christian-Lutheran (specifically German Pietist) light, as many commentators of Kant do. However, I think this view can obscure other, equally important, historical sources. To mention a few, one can trace Kant's idea of conscience to Rousseau's *Emile*, and from there to British Sentimentalists such as Butler and Shaftesbury—all three of which draw heavily from Stoic sources (Epictetus' *Discourses* in particular).

that adding Kant's theory of conscience to the Comparative View offers a solution to the problem of possible deception in self-judgment.[24]

(A) In the *Metaphysics of Morals*, Kant addresses the apparent contradiction of identifying the "inner prosecutor" and the accused, for "to think of a human being who is *accused* by his conscience as *one and the same person* as the judge is an absurd way of representing a court, since then the prosecutor would always lose" (*MS* 6:438). Given this passage, I feel it would be a serious mistake to construe Kant's talk of "courtrooms," "prosecutors," "defense counsels," and "final verdicts" as nothing more than metaphors. The distinctions Kant wishes to establish within conscience are normative, not metaphorical. He argues, for instance, than "one constrained by his reason" must necessarily represent the accusations of conscience as the accusations "*of another person*" (*MS* 6:438). He clarifies this idea in a footnote, pointing out that

> A human being who accuses and judges himself in conscience must think of a dual personality [*zwiefache Persönlichkeit*] in himself, a doubled self [*doppelte Selbst*] which, on the one hand, has to stand trembling at the bar of a court that is yet entrusted to him, but which, on the other hand, administers the office of judge that it holds by innate authority. (*MS* 6:438n)

The "doubled self" refers to the human being's twofold empirical and intelligible nature. Now, leaving aside the metaphysical problems that may arise from this distinction, it is important to understand that—for Kant—the agent experiences her conscience, not as her empirical and corrupted will (or who she *is*), but as her free and perfected will (who she *ought* to be). The authority of conscience arises from the fact that we necessarily represent it as our ideal moral self, which is why Kant describes conscience as the "inner judge of all free actions" (*MS* 6:438). I will return to this point briefly.

(B) Conscience plays two judicial roles in Kant's moral theory. The first is a higher-order judgment of whether the agent has properly

---

[24] Kant's theory of conscience has attracted increasing attention within the recent philosophical literature. I have benefited from the excellent accounts presented by Thomas Hill's "Four Conceptions of Conscience" and "Punishment, Conscience, and Moral Worth" in his *Human Welfare and Moral Worth: Kantian Perspectives* (Oxford: Clarendon Press, 2002) and especially from Allen Wood's chapter on "Conscience" in his *Kantian Ethics* (Cambridge: Cambridge University Press, 2008). Other authors who discuss Kant's views on conscience are Felicitas Munzel, *Kant's Conception of Moral Character: The "Critical" Link of Morality, Anthropology, and Reflective Judgment* (Chicago: University of Chicago Press, 1999) and Jason Howard, "Kant and Moral Imputation: Conscience and the Riddle of the Given," *The American Catholic Philosophical Quarterly,* 78:4 (2004): 609-627.

incorporated her moral principles into her actions. In the *Metaphysics of Morals*, for example, Kant outlines the following process of moral deliberation: *Practical understanding* provides me with the "rules" or "principles" of morality that constrain my range of choices. These principles allow me to assess what I ought to and ought not to do. The *faculty of judgment* then determines two things: *generally*, whether my past or projected action has the status of what Kant calls a "deed," an action falling within the jurisdiction of the moral law; and *specifically*, whether my past or projected action properly incorporates the judgments of practical understanding—whether I acted (or will act) on what I judge to be my duty. Finally, *conscience* issues the verdict on my action: immoral ("guilty") or moral ("not guilty"). "All of this takes place," Kant writes, "before a *tribunal*… an *inner court* in the human being" (*MS* 6:438). Notice that I take an *active* role in assessing my duties, in deliberating what actions the moral law, by way of my understanding, compels me to pursue or avoid. But my action (before or after it occurs) is *passive* to the appraisal of conscience (*MS* 6:439; cf. *MpVT* 8:269n). This is one of the senses in which I experience my conscience as *another*, for my conscience condemns me "spontaneously" or "instinctually" if I fail to act on what I judge to be my duty.

Practical understanding is prone to error, however. This leads to conscience's second judicial function. I can assess my actions, and act on what I "objectively" judge to be right, but still fail to properly incorporate the rules of morality into my actions. People make wrong moral judgments all the time. But how are we to make sense of Kant's claim that an "erring conscience is an absurdity" (*MS* 6:401)? If I understand him correctly, Kant's idea is that I cannot fail to believe whether I've submitted my actions to the appraisal of practical understanding. I cannot fail to believe, in other words, whether I have consciously examined my duties. As he writes, "while I can indeed be mistaken at times in objective judgment as to whether something is a duty or not, I cannot be mistaken in my subjective judgment as to whether I have submitted it to my practical reason (here in its role as judge) for such a judgment" (*MS* 6:401). This helps explain Kant's statement from the *Religion* that conscience is "*the moral faculty of judgment, passing judgment upon itself*" (*R* 6:186).[25] By this he means

[25] Thomas Hill offers the important insight that Kant speaks of conscience's "judgment" in two different senses: "Metaphorically speaking, 'judgment$_1$' (one sense of 'judgment') is what is responsible for appraising the act diligently, and 'judgment$_2$' (a second case of 'judgment') on judgment$_1$ as to whether it has fulfilled that responsibility" ("Four Conceptions of Conscience," p. 302).

conscience judges the agent's awareness in having thoroughly appraised her duties. The second judicial function of conscience is thus a higher-order judgment of the care the agent applies (or fails to apply) in the act of examining what action she ought or ought not take.[26] In this case I stand guilty before the inner judge not by failing to act on what I believe to be my duty but by failing to properly scrutinize, under the lights of practical understanding, what my duty is.

While the first function of conscience is to see whether the agent's actions really do line up with the judgments of practical understanding, the second function is to see whether the agent really does perform a self-critical assessment of her judgments (before or after those judgments take effect in action). Conscience thereby displays the capacity we have to judge our own judgments, so that when we take ourselves to be responding to a particular moral demand we can ask ourselves, "Have I *carefully* reflected on how I should act?" which is different from, "Have I *really* acted on what I judge to be my duty?" If it turns out that we've assessed the appropriateness of our action haphazardly—or not at all—then conscience condemns us. So at least pertaining to its higher-order function, Kant is right: an erring conscience *is* an absurdity, for the simple reason that an agent can't critically assess her duties unconsciously. The additional qualification of a "careful" assessment of one's duties is secondary to the basic awareness one has of performing the assessment itself. In this way, we are incapable of disavowing the higher-order verdicts of conscience—whether or not we have reflected on the appropriateness of our moral actions—because we are *conscious* of whether we have or have not done so. Kant describes this type of self-conscious awareness in terms of *truthfulness*, which is the reflexive standpoint we take to own moral judgments, as opposed to *truth*, which is the set of objective facts or features of our

[26] Kant remarks that it is one's responsibility to "enlighten his understanding in the matter of what is or is not duty" but as soon as he acts (or is about to act), conscience "speaks involuntarily and unavoidably" (*MS* 4:401). Conscience is *phenomenologically distinct* from practical understanding in that its judgments are immediate and spontaneous, not deliberative and thoughtful. Kant argues for this point in his early lectures on ethics, when he speaks of the "instinct" of conscience: "Everyone has a faculty of speculative judgment, though that is at our discretion; there is, however, something in us which compels us to pass judgment on our actions. It sets the law before us, and obliges us to appear before the court. It passes sentence on us against our will, and is thus a true judge" (*LE* 27:297). Thus, I am passive to my conscience in two different senses. First, conscience is not developed or acquired (it is constitutive of my moral agency, like moral feeling). Second, as the inner voice of the moral law, I necessarily represent conscience as "another." It is in this sense that I experience my conscience as "external" to myself, i.e., that I experience the reproaches of conscience *as if* issued from another (morally ideal) person.

judgments as they relate to the world or to our deeper (epistemically unavailable) intentions. By drawing this distinction, Kant's point is that while we can never know the objective truth-value of our moral judgments we can still have "immediate consciousness" of holding these judgments to be true (*MpVT* 8:267), and that is precisely the level on which conscience operates.

Here it may appear as though Kant is contradicting his Opacity Thesis. How can I have immediate consciousness of the verdicts of my conscience? How can I claim to know, without doubt, that I stand guilty or not before the inner judicial court? Kant's answer is that conscience accuses me exactly where I am transparent: my sense of truthfulness.[27] But this transparency is not in the order of knowledge. By Kant's definitions, "knowing" is a matter of having *objectively sufficient grounds* to hold something as true, which is why we characterize knowledge in terms of universal agreement (for example, something shared or shareable by every rational person). Only knowledge yields certainty; but certainty is not the only mode of taking something to be true. Kant defines "believing," for instance, as having *subjectively sufficient grounds* to hold something as true in the absence of objective grounds; and while these grounds do not yield "certainty," they do yield what Kant calls "conviction" (*KrV* A822/B850).[28] To say that I am convinced of the verdicts of my conscience does not contradict the Opacity Thesis, for that thesis only constrains the objective grounds of my knowledge-claims. I can, for example, maintain conviction in the honesty of my declaration, even though its objective truth-value falls outside the limits of my understanding. Kant is emphatic on this point: "I can indeed err in the judgment *in which I believe* to be right, for this belongs to the understanding which alone judges objectively (rightly or wrongly); but in the judgment *whether I in fact believe* to be right (or merely pretend it) I absolutely cannot be mistaken" (*MpVT* 8:268). Similarly, I can err in the judgment of what my duty is, but I cannot err in my general conviction of having properly appraised my duties. Again, this is why I cannot avoid accepting the final verdict of my conscience, guilty or not guilty, because I am conscious of holding this verdict within myself. *I know* whether I sincerely examined my actions, or whether I lied to myself or to others. These actions are transparent to me because I cannot rationally deny *my own belief* in

---

27 Kant calls this *formale Gewissenhaftigkeit*, literally "formal conscientiousness" (*MpVT* 8:268).

28 Cf. *The Jäsche Logic*, IX of the Introduction, in *Lectures on Logic*, trans. J. Michael Young (Cambridge: Cambridge University Press, 2007).

having performed them. I therefore have privileged access to my conscience because that access is grounded in a reflective apprehension of my own beliefs, and this is why Kant is at pains to distinguish the first-personal character of conviction from the third-personal character of certainty.[29]

Now in light of (A) the normative identity and (B) the two functions of conscience, I believe we can complete Kant's Comparative View developed in the *Religion*. To summarize, the Comparative View infers the agent's change of heart on the basis of the perceived difference between her old and supposedly new ways of moral conduct. The agent must therefore ask herself what kind of verdict she could hope to receive if her whole life were placed before a judge. Kant argues that if the agent represents this judge as *another*, she will be tempted to deceive it through an empty display of moral righteousness. But if she represents the judge as *herself*, she will appraise herself firmly. As I noted above, the legitimacy of this final verdict remains questionable until we can determine the nature of this judge, its identity and jurisdiction.

What we might call the Comparative-Conscientious View finally answers this question. In the first place, Kant's theory of conscience overcomes the problem that threatened to undermine the Comparative View, which is how I can condemn myself. The normative dualism of conscience explains how I can stand accused before myself, something that is contradictory if the judge of my moral conduct turns out to be the dear self. This also explains why I cannot bribe or even disagree with conscience's final verdict. While the purity of my disposition, whether it is actually good or evil, is impossible to cognize directly, I am immediately conscious of whether I've examined my life conduct with due care. Kant's point, as I discussed above, is

---

[29] Of course, Kant is not denying that I can attempt to ignore the accusations of my conscience, to distract myself from the stirrings of the inner judge. His point is that precisely by attempting to ignore my own guilt, I am testifying against myself, if you will, in favor of conscience. Kant further argues that the privileged access we have to our own beliefs prevents us from immunizing ourselves from the onslaughts of conscience. As he writes: "A human being may use what art he will to paint some unlawful conduct he remembers as an unintentional fault… and to declare himself innocent of it; he nevertheless finds that the advocate who speaks in his favor can by no means reduce to silence the prosecutor within him, if only he is aware that at the time he did this wrong he was in his senses, that is, had the use of his freedom; and while he *explains* his misconduct by certain bad habits… this cannot protect him from the reproach and censure he casts upon himself" (*KpV* 5:98). For Kant, first-personal awareness is necessarily tied to basic beliefs of our own agency. To say, "*I* acted" in such and such way is implicitly to say, "I was *free of constraint*" in so acting. Presumably, the possibility of an "inner prosecutor" arises from our own subjective belief in having acted freely.

that I cannot fail to be aware of my own honesty or dishonesty, whether this applies to the actions I submit to the appraisal of practical understanding, or more simply, to the sincerity of my testimony. Conscience alone does not establish assurance in my restored disposition, but rather allows me to *trust* the final sentence I pass on my life as a whole. While comparative self-knowledge is responsible for assessing the perceived difference between my old and new ways of life, conscience is responsible for condemning or acquitting me in my effort (or lack of effort) to examine this difference diligently. Together, the Comparative-Conscientious View not only supplies me with confidence in the authenticity of my restoration, which is essential for the continual pursuit of virtue; it also gives me confidence in my own self-assessment. Only by adding conscience to the Comparative View can we account for the fundamental *self-trust* needed in our quest for self-knowledge.

Keep in mind that I'm not claiming to have one of Hawthorne's moments, one of those moments "when a man's moral aspect is faithfully revealed to his mind's eye," for that would contradict the Opacity Thesis. Nor am I claiming to infer my disposition and progress toward the good by way of my actions alone, for actions are not judgment-neutral. Of course, I still need to exhibit my moral character through my life-conduct; and to this extent, comparative self-knowledge is still inferential. But the deeper question here is how I can trust the inferences I draw from my conduct as a whole. The advantage of adding conscience to the Comparative View is that we can finally overcome the major threat to the duty of self-knowledge, namely, the dear self. All the dear self can do is obscure the empirical evidence I bring before the "inner court," such as the list of past actions evincing my moral improvement; and it can do this because I will never have objective grounds to assess my moral disposition. But the dear self can't corrupt my awareness of having appraised my life thoroughly. This entails that the truth of the evidence I bring before the inner court of conscience is fallible, because it can always be corrupted by Type-2 opacity, but the truthfulness of my attempt to examine this evidence is beyond corruption. And that's why an erring conscience is an absurdity.

Kant's claim in the *Religion* that one should stand in judgment before oneself makes sense only if we replace "oneself" with "one's conscience," something he alludes to when speaking of the "judge within." This explains why later in the *Metaphysics of Morals* Kant situates conscience, or the "human being's duty to himself as his own innate judge," *before* the duty of self-knowledge. As far as I can tell, this is because the intrinsic authority of conscience, as the access we

have to our own beliefs, is an essential requirement for the duty of self-knowledge. Now one might ask at this point: Why is self-knowledge the *first* command of all duties to oneself? Shouldn't we give primacy to the duty of conscience? Just as I can't properly pursue my natural perfections until I know the purity of my heart, whether it is good or evil, I can't properly judge my heart until I can trust my self-judgments. Isn't self-knowledge only possible on the grounds of conscience, on our capacity to judge ourselves sincerely? Indeed, Kant's answer is affirmative: The duty of self-knowledge *is* grounded in conscience. But, he argues, the concept of a duty to conscience is contradictory for the simple reason that "conscience is not something that can be acquired" (*MS* 6:400). Rather, "every human being, as a moral being, *has* a conscience within him originally" (*MS* 6:400). Conscience, like moral feeling, is thus constitutive of our identity as moral agents. And like moral feeling, our duty to conscience is only indirect. One's obligation here, Kant maintains, is to "cultivate one's conscience, to sharpen one's attentiveness to the voice of the inner judge and to use every means to obtain a hearing for it" (*MS* 6:401). Conscience conditions the possibility of self-knowledge, but self-knowledge is still the first *duty* to oneself.

Conclusion

I argued earlier that the duty to know my particular heart is impractical if I can never have any degree of conviction in my restoration. Kant himself admits that without such conviction one would be led into hopelessness and despair. It would be rather disturbing if generic self-knowledge led me to the insight that I *ought* to become a morally better person but that from the standpoint of my idiosyncratic self I could never *know* whether my attempts to become good were genuine or counterfeit. The duty of self-knowledge clearly requires the agent to apprehend the particular evils that stand in the way of her restoration, but for the sake of continued motivation and perseverance she must also be able to tell, if only slightly, that her self-improvement will not slip back into evil. For Kant, it is not within the power of one's conscience to grant rewards for good behavior, which is why the "comforting encouragement of one's conscience is not *positive* (joy) but merely *negative* (relief from preceding anxiety)" (*MS* 6:440). Perhaps the most we can hope for regarding the assurance of our moral restoration is a clear conscience, the conviction that we have appraised our life-conduct cautiously and with due care. To live with a clear conscience would be to live free from anxiety, free from hopelessness and despair. And that is perhaps the

most we can ask for, given that the depths of the human heart are, after all, unfathomable.[30]

[30] In writing this paper I have benefited from a number of individuals through conversation or written comments. In particular, I am grateful to Steve Engstrom, Paul Franks, Anna Leah Harms, Arthur Ripstein, and Sergio Tenenbaum. Thanks also to the Social Sciences and Humanities Research Council of Canada for offering me financial support during the research and writing of this paper.