In this paper I discuss Kant's theory of conscience. In particular, I explicate the following two claims that Kant makes in the *Metaphysics of Morals*: (1) an erring conscience is an absurdity and (2) if an agent has acted according to his/her conscience, then s/he has done all that can be required of him/her. I argue that (1) is a very specific claim that does not bear on the problem of moral knowledge. I argue that (2) rests on a strongly internalist line of argument.

0 Introduction

Kant makes the following two claims in the second half of the *Metaphysics of Morals*:

1. I shall here pass over the various divisions of conscience and note only that, as follows from what has been said, an *erring* conscience is an absurdity. (6:401)
2. But if someone is aware that he has acted in accordance with his conscience, then as far as guilt or innocence is concerned nothing more can be required of him. (6:401)[1]

The first claim seems to be a denial of the problem of moral knowledge. That is, 1 seems to be saying that, through conscience, agents have an immediate, intuitive awareness of what is permissible and what is not. It is through conscience that agents judge whether they have fulfilled their duties. But because an erring conscience is an absurdity, agents must have an immediate, intuitive awareness of what their duties are by means of conscience.

Based on this reading of 1, 2 is uncontroversial. If an erring conscience is an absurdity, then it follows that if one is acting in accordance with conscience, one is acting permissibly. But if one is acting permissibly, then, as far as guilt or innocence is concerned, nothing more can be required of one. If one is acting permissibly, then one *eo ipso* is doing what morality requires. One can understand this argument more clearly by means of the distinction between objective and

[1] Cf. also, e.g., 6:189, where Kant claims that "more [than acting in accordance with conscience] cannot be required of a human being". See also 27:335 and 355.

subjective rightness.[2]

Philosophers often distinguish between "objective" and "subjective" senses of "rightness". Suppose it is an objective truth that I am obligated to do D. It is possible for me to believe that I am not obligated to do D, and suppose that in fact I believe that I am not. If this is so and if I do not do D, then I have not acted objectively rightly but I have acted subjectively rightly. In other words, accurate beliefs or full information might lead an ideal agent to conclude that action A is the right one (objective rightness), but given what the actual agent in the context knew and thought, it was only to be expected that this agent would think B was the right action (subjective rightness). It is then a separate question, about which there can be disagreement, whether it is culpable not to act rightly objectively when one acts rightly subjectively. But if 1 is read in the way that I suggested above (i.e., if 1 is taken to be a denial of the problem of moral knowledge) then the distinction between objective and subjective rightness implodes; it is not possible for me to believe that I am not obligated to do D if it is an objective truth that I am obligated to do D.

In this paper, I argue that this is a misreading of these two claims. In particular, I argue that 1 is much more subtle than a denial of the problem of moral knowledge, and I argue that 2 rests on an appeal to an internalist line of argument. The paper is divided into three sections. In the first section, I describe the functions of a Kantian faculty of conscience. In the second section, I explicate 1. I argue that the natural reading of 1 is incorrect; 1 is a very specific claim about a particular function of conscience and, I argue, Kant gives a plausible defense of it. In the third section, I explicate 2.

[2]Kantians might want to substitute 'permissibility' for 'rightness' to avoid ambiguity; I shall not be talking about Kant's duties of right, which are more properly discussed in the context of his *Rechtslehre*.

1 Conscience

Kant thinks of conscience as an inner court.[3] He often depicts it as a court with a divine tribunal.[4] Like a judge in a court, conscience "acquits or declares us to deserve punishment" (27:353).

Now Hill argues that, on Kant's view, conscience itself punishes.[5] Wood argues that, on Kant's view, conscience does not and, moreover, cannot punish. According to Wood, a Kantian conscience cannot punish because an agent cannot punish him/herself.[6] The text seems to support Wood's reading over Hill's. In his discussion of criminals, Kant argues that agents cannot punish themselves. Moreover, Kant does not view the painful feeling that comes with the recognition that one has violated the moral law as a punishment. At 6:439n, Kant says that conscience judges one punishable, but he says it leaves to God whether to punish one.

Just as conscience does not punish, conscience does not reward. The only joy that conscience affords is in avoiding bad actions and, thus, avoiding deserving rightful punishment (6:440). However, if conscience is to motivate, as Kant suggests that it does at 6:400, Kant's moral agent must have a motive to want to avoid deserving punishment. Unlike the motive to want to deserve the pain or harm involved in actually being punished, having a special motive, quite distinct from this one, to want to avoid deserving punishment is a significant point in Kant's moral psychology. I touch on these issues briefly at the end of this section and, again, at the end of this paper. Nonetheless, the only pleasure conscience affords is negative (in the avoidance of pain). Thus Kant makes the following claim in the *Metaphysics of Morals*:

> It should be noted that when conscience acquits him [i.e., any agent] it can never decide on a *reward* (*praemium*), something gained that was not his before, but can bring with it only

[3] See, e.g., 5:98; 6:438; 27:295, 351, 353-354 and 572-573.
[4] See, e.g., 6:146n, 439; 27:296, 355 and 574-575.
[5] [5], chapter 11.
[6] [11], 187-189.

> *rejoicing* at having escaped the danger of being found punishable. Hence the blessedness found in the comforting encouragement of one's conscience is not *positive* (joy) but merely *negative* (relief from preceding anxiety)... (6:440)

According to this quotation, a Kantian conscience never rewards or confers happiness; it merely allows one the relief in having avoided a wrong-doing. However, it is not clear whether Kant is consistent on this point. At 27:296-297 Kant claims that conscience conveys an "inner joy" at good actions. One might resolve the textual inconsistency by arguing that the inner joy referred to here is supposed to be merely negative.

A related worry arises when one considers Kant's classification of actions. On Kant's account, actions are forbidden (blamable), permissible, required (blamable not to do) or meritorious. But courts tend either to mete out punishment (blame) or to omit to do so (if the verdict is 'innocent'). Courts do not offer rewards or declare actions meritorious. Kantian conscience, however, if it were the sort of tribunal authorized to make the standard ethical judgments Kantian ethics recognizes for actions, ought to be empowered to declare some actions meritorious and, hence, deserving of a reward (though of course, as in the case of deserving punishment, conscience itself neither rewards nor punishes). One might wonder why Kant does not empower conscience to judge actions meritorious as well as innocent or guilty. Perhaps it is merely an infelicitous feature of his choice of the idea of conscience as a court, for criminal courts decide only between innocence and deserving punishment rather than authorize rewards. Alternatively, merit might be registered in the agent's consciousness without its being a possible verdict of conscience. Conscience might be thought of as having to do with whether one has done (or is proposing to do) wrong or has incurred (or is proposing to incur) guilt but not with whether one's actions are meritorious; if this is correct, then one need not deny that agents might be conscious of merit when their actions possess it. However it is not clear whether any decisive conclusions can be reached here; the texts are silent.

Moreover, any discussion of these topics would require a thorough investigation of exactly how Kant's classification of actions plays out, which is far beyond the scope of the present investigation.

It is (in principle) possible for one to judge another in the way that one's conscience judges oneself (provided that one knows all of the other's relevant beliefs and principles). But a Kantian conscience is involved only in one's judging of oneself. I shall return to this point at the end of this section.

Kant warns against confusing conscience with understanding. He tells us that "it is understanding, not conscience, which judges whether an action is in general right or wrong" (6:186). Kant's particular phraseology here ("in general") is extremely important. For one thing, it indicates that Kant does not think that the general duties he enumerates in the *Metaphysics of Morals* are either universal or provisionally universal. This interpretation is buoyed up by the presence of casuistical questions and by the way the examples are set up in the *Groundwork for a Metaphysics of Morals* (i.e., these examples seem to have agents approaching general duties with the assumption that they are neither universal nor even provisionally universal).[7] For another, it leaves room for individual judgment (and, thus, error) in judging particular cases. Although the issue of error is discussed to some extent later in this paper, the status of Kant's general duties (as universal or provisionally universal) is beyond my present scope.[8] The general idea Kant is expressing in this quotation from 6:186 is that once an action has been subsumed under a principle, it is understanding (rather than conscience) that is responsible for judging whether that action is in general permissible. For ex-

[7]Many of Kant's predecessors and successors ascribe to conscience the functions of resolving cases of conflicting duties and answering casuistical questions. However, Kant does not do this. For one thing, he argues that conflicts of obligation are impossible (6:224). For another, he thinks that figuring out whether a particular action is right involves the faculty of reason rather than of conscience. This is discussed further in the text above.

[8]However, it is perhaps worth noting that Kant distinguishes between right and ethics in this regard; he argues that "[in contrast to the doctrine of right], the doctrine of virtue... cannot refuse some room for exceptions (*latitudinem*)" (6:233). Ethical principles allow for exceptions; principles of right do not.

ample, it is understanding that is responsible for judging whether suicide is generally permissible (or whether making a false promise is generally permissible).

Kant also warns against confusing conscience with reason. He tells us that "conscience does not pass judgment upon actions as cases that stand under the law, for this is what reason does so far as it is subjectively practical" (6:186). Reason, insofar as it is subjectively practical, is responsible for judging whether any given action is permissible; this is not a function of conscience. Thus, for example, it is the function of subjectively practical reason (rather than of conscience) to determine whether some particular suicide or false promise (or whatever) is permissible. Thus, Kant claims that agents must "... have prior knowledge of the good, if conscience is to judge" (27:356; see also 27:576 and 27:617).

Thus far, I have described what conscience does not do on Kant's account. Conscience does not punish, reward, judge others, judge actions as generally permissible or judge whether any particular action is permissible. Now I turn to Kant's account of what conscience does do.

There are two passages that are particularly helpful here. The first occurs in the *Religion within the Boundaries of mere Reason*. It is as follows:

> ... [in conscience] reason judges itself, whether it has actually undertaken, with all diligence, that examination of actions (whether they are right or wrong) and it calls upon the human being to witness *for* or *against* himself whether this has taken place or not. (6:186)[9]

The second passage occurs in the *Metaphysics of Morals*. It is as follows:

> ... conscience is practical reason holding the human being's duty before him for his acquittal or condemnation in every case that comes under a law.... (6:400)

Based on the passage from the *Metaphysics of Morals*, it seems clear that on Kant's account conscience is responsible for judging whether an agent has fulfilled his/her duties in any given

[9] Cf., e.g., 27:614-615.

instance. Conscience holds an agent's duty before him/her for acquittal or condemnation in every case that stands under a law. Once the action has been subsumed under a law and judged to be permissible or impermissible, conscience appraises the agent in having done (or in proposing to do) some action. I shall call this the duty function. Kant never uses this term to describe one of the functions of conscience; I have chosen it because it reflects that which conscience does on Kant's account. Conscience fulfills the duty function insofar as it determines whether an agent has fulfilled his/her duty in any given instance and, thus, whether an agent is blameworthy.

Now before turning to the passage from the *Religion within the Boundaries of mere Reason*, it is useful to address a potential objection. It might be objected that Kant is profoundly pessimistic about agents' abilities to assess whether they are behaving permissibly. But if agents are opaque to themselves in this way, then the practicability of the duty function is highly attenuated. So either I am misinterpreting Kant in assigning the duty function to conscience, or Kant's theory of conscience does not fit well with a view that seems deeply rooted in his philosophy. And because what I am calling the duty function is taken right from the surface of the passage from 6:400, it looks like the conclusion that is unavoidably forced out is that Kant's theory of conscience does not fit well with a view that seems deeply rooted in his philosophy.

However, closer inspection reveals that this objection is based on a misunderstanding of that about which Kant is pessimistic. Kant is not pessimistic about agents' abilities to assess whether they are behaving permissibly. Rather, he is pessimistic about agents' abilities to determine in any given instance what their motives are: whether they are acting from duty or whether they are acting from prudence (or both). In other words, Kant is pessimistic about agents' abilities to determine what their dispositions are, not about their abilities to determine whether, for example, the actions they are about to perform are permissible. This is made clear at the beginning of part II of the

Groundwork for a Metaphysics of Morals, where Kant claims that "although many things are done in conformity with what duty prescribes, it is nevertheless always doubtful whether they are done strictly from duty, so as to have a moral worth" (4:406). So there is no deep tension here between Kant's theory of conscience and his beliefs about the opacity of agents. In other words, the duty function assesses whether one has acted in accordance with duty. But it is possible to do this without determining whether one has acted from duty. So the objection is unfounded. Let us now turn to the passage from the *Religion within the Boundaries of mere Reason* to see what Kant says about the role of conscience there.

The passage from the *Religion within the Boundaries of mere Reason* is more subtle than the passage from the *Metaphysics of Morals*. Kant tells us that conscience determines whether an agent actually has undertaken to examine his/her actions in any given instance. Moreover, it determines whether an agent has done so with "due diligence".

Kant's use of the term 'due diligence' here is puzzling. This is a legal notion, and it usually is involved with judgments about negligence. The term suggests that there is some standard according to which an agent either uses "enough" or "not enough" care in examining his/her action in any given instance. One way to cash this out would be to say that the more "perfect" a duty is, the more careful one must be in determining whether a given action is a permissible exception to the rule.[10]

At times, Kant indicates that the duty to exercise due diligence is quite strict. For example, the following passage can be found in his *Lectures on Ethics*:

> Nobody can take a thing to be right or wrong, even when probability is present, so long as he cannot dismiss the possibility of the opposite. (27:615; see also 6:185-186)

But it is not clear why Kant takes such a strong line and it is not clear that a Kantian ethics would

[10]Kant's distinction between perfect and imperfect duties is notoriously difficult and, as already noted, far beyond the scope of this paper.

need to absorb this part of Kant's ethics. It seems like the line of thought expressed in this quotation would make it very difficult to get anything done, for it is rare that agents are in a position to dismiss the possibility of the opposite of one of their beliefs.[11] I shall take no settled position here. It might be the case that these references to probability are references to the Jesuit doctrine of probabilism, which was well-known during Kant's time through Pascal's harsh critique of it in the *Provincial Letters*.[12]

According to probabilism, in difficult cases, cases in which one is uncertain about what to do, one may follow a doctrine that is *probable*. Now the proponents of probabilism do not spell out precisely what counts as "probable" and it is difficult to know exactly what they had in mind when one gets down to the nitty-gritty. Obviously I cannot try to settle exegetical questions here about the intentions of authors like de Medina. But one thing that is notable about this doctrine is that one is permitted to follow an opinion that is probable *even if the opposite opinion is more probable*. Thus, for example, it seems that according to probabilists it would be "permissible" to leave my umbrella at home even if there is a 75% chance of rain.[13]

It seems unlikely that Kant would subscribe to this position, especially given what he says

[11] Strictly speaking, this is somewhat misleading. It is not rare for agents to be in a position to dismiss some line of action as clearly wrong. For instance, it clearly would be wrong for someone to steal his/her neighbor's recycling bins because s/he has decided to start a collection; it is easy to dismiss the thought that this would be permissible. Even more obviously, it clearly would be wrong to shoot one of my classmates because s/he is sitting in the seat in which I usually sit. Millions of vicious or thoughtless actions I never even consider would be dismissed easily as wrong if I did consider them. But in cases that are likely to trouble one's conscience, it is not rare to remain unsure whether one has done the right thing. Sometimes conscientious reflection leads to certainty, as when one sees clearly that one's temptation to think a certain action is OK is a result of self-deceptive self-love or some other corrupt mental process. But conscience sometimes wrestles with hard decisions and a conscientious agent might never be sure whether this or that course is the right one. Kant, it seems, likes to think that if an agent is sufficiently conscientious, s/he always will arrive at a point where s/he can be morally comfortable with what s/he does, even if it is the hard choice from a prudential standpoint. But I think many of us regard that as a moralist's illusion.

[12] In Kant's discussion of these issues at 6:185-186, he mentions probabilism explicitly. Probabilism is also mentioned explicitly in the *Lectures on Ethics* only a few pages after the passage quoted above (see 27:622).

[13] 'Permissible' is in scare quotes here because this is not a moral question. But hopefully the idea is clear.

in his *Lectures on Ethics*. But the point is that his remarks in the *Lectures on Ethics* might be a reaction to probabilism, perhaps even a reaction inspired by the work of Pascal, who claimed that probabilism leads to moral lassitude. Perhaps in his more reflective moments, Kant would have taken a middle road. For example, he might have adhered to the view that when one is uncertain about what to do, one ought to do that which has a *higher* probability of being right. There is no way to tell, and here I shall follow Kant's prescription insofar as "probability is present": I remain in *aporia* with regard to what Kant's position is about what counts as due in due diligence.

Regardless of what one makes of Kant's discussion of due diligence, it is clear that the function he is ascribing to conscience in this passage is not the same as the duty function. I shall call this function the moral reflexivity function. 'Moral reflexivity function' is not Kant's term; this is my term to describe one of the functions that Kant ascribes to the faculty of conscience. Conscience performs the moral reflexivity function insofar as it determines whether an agent has examined his/her actions to determine whether they are permissible. If an agent has examined his/her actions in this way, the moral reflexivity function answers in the positive; if an agent has not examined his/her actions in this way, the moral reflexivity function answers in the negative. Even factoring in Kant's remarks about due diligence, the moral reflexivity function is still a simple binary.

I shall explain the distinction between the duty function and the moral reflexivity function at greater length in the next section. But before I do that, I need to make one point about the account I already have given more explicit.

I pointed out above that conscience does not judge others. However, both the duty function and the moral reflexivity function, as I describe them, in principle could be exercised on someone else. By eavesdropping on someone else and his/her actions, I could make a judgment whether s/he is doing the thing s/he ought to do (the duty function) or whether s/he has exercised due care

in making the judgment that s/he ought to do it (the moral reflexivity function). But it is essential to Kant's conception of conscience that whatever judgment it makes it makes only about the agent him/herself. This is what Kant means when he says that in conscience "reason judges itself"; he means the individual's (practical) reason regarded as the seat of the individual's moral personality. This is what he is driving at also when he says that conscience is practical reason "holding the human being's duty before him for his acquittal or condemnation". The judgment in question is essentially self-reflexive. Nobody can be another's conscience.

This self-relatedness is closely connected to the fact that conscience involves a feeling (6:400). A feeling for Kant is not cognitive; it is only the subject's awareness of its own subjective state in a way that motivates or tends to motivate. The judgment conscience makes is a judgment about oneself and it involves either a pleasure (although only a negative one, that one is not guilty) or displeasure (that one is guilty) in performing a certain action. It is via these feelings that conscience motivates. I now turn to Kant's discussion of errors of conscience.

2 Errors of conscience

In the introduction of this paper, I distinguished between objective rightness and subjective rightness. The basic idea is that if (and only if) an agent performs an action that is in accordance with his/her (subjective) principles, s/he is acting in a way that is subjectively right; if (and only if) an agent performs an action that is in accordance with principles that are objectively correct, s/he is acting in a way that is objectively right.

As pointed out in the previous section, conscience is not involved directly in determining whether a given action is permissible. Whether a given action is permissible is determined by an

agent's reason and understanding. Conscience answers the duty function and the moral reflexivity function in accordance with an agent's subjective principles; conscience functions in accordance with an agent's prior knowledge of the good.

Because different agents have different principles (the USA, for example, is a pluralistic society), different agents will be acquitted or condemned by their respective consciences for different things. Bertrand Russell had no qualms about living with a partner despite not being married. A member of the religious right might not be able to behave like Bertrand Russell without being condemned by conscience.

Now, as pointed out in the introduction, at 6:401 Kant claims that conscience cannot err. Hill interprets Kant's claim that conscience cannot err as the following claim: agents never make a mistake in determining whether they are guilty of performing acts-as-they-perceive-them that are not in accordance with their moral-principles-as-they-accept them.[14] That is, Hill takes Kant to be saying that agents never err in assessing whether they have acted subjectively rightly (the duty function). Call this Hill's reading.[15] In the introduction, I argued that the most natural reading (which I shall refer to as such) of Kant's claim is as a denial of the problem of moral knowledge. That is, I argued that the most natural reading of Kant's claim is that agents' principles are always objectively correct; there is no gap between subjective and objective rightness.[16] In this section, I shall argue that both Hill's interpretation and the most natural reading of Kant's claim at 6:401 are misinterpretations of the text.

However, it is worth pointing out that if one looks at the *Lectures on Ethics*, one sees that Kant

[14] [5], 348p3.

[15] It should be noted that Hill is not the only commentator who subscribes to what I am calling Hill's reading. For example, Skorupski also subscribes to this reading (see [7], 555p3).

[16] This reading is taken by Hardwig. Hardwig argues, first, that on Kant's account there is no problem of moral knowledge ([2], 285p1) and, second, that Kant is forced to adopt this position on pain of contradiction ([2], 287p3-4).

is not consistent on this point about errors of conscience. At 27:297, Kant argues that conscience cannot err on the grounds that conscience is an instinct and thus must be distinguished from the faculty of speculative judgment because it is not free. Presumably the idea is that only free faculties (such as the faculty of speculative judgment) can make errors. At 27:615, Kant takes the claim that conscience cannot err as a starting point in order to adduce its function. At still other places, Kant admits explicitly that conscience can err (see, e.g., 27:42 and 354-355).[17] I shall take the claims from Kant's published writings, especially those from 6:401 and 8:268, as his genuine position.

At 6:401, Kant is not claiming that conscience cannot act in accordance with bad principles (i.e., that agents cannot perform actions that are subjectively right but not objectively right — the natural reading) or that agents never err in assessing subjective rightness (Hill's reading). On the contrary, Kant's claim, taken in context, is as follows:

> . . . while I can indeed be mistaken at times in my 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 practical reason (here in its role as judge) for such a judgment; for if I could be mistaken in that, I would have made no practical judgment at all, and in that case there would be neither truth nor error. (6:401)

Kant here admits that agents can be mistaken in their judgments as to whether something is a duty. This immediately puts the natural reading to rest.[18] Moreover, a close look at this passage also shows why Hill's interpretation is incorrect. Kant's claim is that agents cannot be mistaken in their judgments as to whether they have submitted something to practical reason at all rather than about

[17] One must remember that the lecture notes were not written by Kant himself. The lecture notes were written by Kant's students. Moreover, they were not written during the lecture course (as students take notes today). Rather, they were written down from memory after the lecture was over. (No doubt, they contain extrapolations where memories were missing.) Given these facts about the *Lectures on Ethics* and given the fact that the various claims about conscience made in the *Lectures on Ethics* are inconsistent, I hope I shall be excused for not looking as closely as one might at these various claims.

[18] Cf also 27:42, where Kant distinguishes between two different kinds of errors of conscience, logical errors and moral errors, and 27:354, where Kant allows for "false *principia* of conscience".

whether agent's can be mistaken in judging that their action is in accordance with their principles. That is, Kant's claim is about the moral reflexivity function rather than the duty function (*pace* Hill).[19] The moral reflexivity function is not the same as the duty function. The moral reflexivity function is like the bell on a microwave: when it goes off, it indicates that practical reason has been employed. Of course, it is not always the case that the moral reflexivity function is employed when practical reason is employed. One might arrive at a practical judgment (e.g., "I ought to X") without having exercised due diligence in doing so but without answering the moral reflexivity function, either. The point is that *if* the moral reflexivity function is exercised, *then* practical reason has been employed. The converse of this is probably false.

Regardless of whether it is true, when Kant claims that conscience cannot err, he is making a very specific claim. This very specific claim is independent of whether conscience can err in the duty function or whether agents can adopt false principles. In order to see this, consider the following. The moral reflexivity function answers the following question: have I submitted action X to my moral principles? The duty function answers the following question: in performing action X, did I (or shall I) be acting in a way that is deserving of punishment? In answering the second question, the agent determines whether his/her action is in accordance with his/her principles — an agent assesses subjective rightness. But these two questions are fundamentally different. Kant's claim is about the first question rather than about the second, so Kant leaves open the possibility of an error of judgment (and thus an error in assessing subjective rightness, the kind of error that Hill argues Kant is trying to rule out). Presumably if an agent can answer the duty function question,

[19] It is perhaps notable that in Hill's discussion of Kant's conception of conscience, he does not mention anything like the moral reflexivity function. He mentions something like what I have called the duty function. He also spends a lot of time on due diligence. However, the moral reflexivity function does not make its way into his discussion (moreover, because due diligence is a proper part of the moral reflexivity function, his discussion of due diligence is somewhat dissatisfying).

then the agent ought to answer the moral reflexivity function question in the affirmative (I come back to this shortly). This does not mean that the agent actually does answer the moral reflexivity function question in the affirmative if s/he can answer the duty function question. The idea is that there is a sort of conceptual priority to the moral reflexivity function. If an agent has not submitted X to moral principles, there is no way for him/her to determine whether X is in accordance with them. But the answer to the duty function question is independent of the answer to the moral reflexivity function question. The point is that the moral reflexivity function is not the same as the duty function, and it is only the moral reflexivity function that Kant declares to be incapable of error.

In order to make Kant's claim stand out as clearly as possible, I would like to spend some time distinguishing some of the different things that might be called an "erring conscience" in addition to Hill's errors of subjective rightness and the natural reading's objectively bad principles. For example, one might have a fanatical conscience, a conscience that strews one's path with "mantraps" (6:409). An agent might be said to have a fanatical conscience if s/he thinks duty constrains him/her at every point, making issues of prudence into issues of morality. Thus, an agent who judges him/herself morally guilty for eating meat for dinner rather than fish despite knowing that fish is much healthier might be said to have a fanatical conscience. This is not to say that there could not be moral reasons for eating fish rather than meat; it is to say that someone who condemns him/herself for eating meat on one occasion and who condemns him/herself on that occasion *because* meat is not as healthy as fish has a fanatical conscience.

Similarly, one might have a morbid conscience. An agent might be said to have a morbid conscience if s/he has a conscience that makes him/her feel (morbidly) guilty all the time. A fanatical conscience can go hand in hand with a morbid conscience if an agent thinks s/he has

duties s/he does not have and then feels guilty for not performing them. But the two are not the same. An agent can feel morbidly guilty without thinking s/he has violated a specific duty; s/he can feel morbidly guilty owing to the sins of his/her ancestors or for simply having sexual desires, even if s/he does not act on them. And an agent with a fanatical conscience might think s/he is constrained by duty at every point but not feel guilty at all if s/he actually performs all his/her imagined duties.

Alternatively, one might be said to have a rigoristic conscience if one sticks strictly and unthinkingly to a set of moral rules without regard for whether there might be mitigating conditions. An agent might be said to have a rigoristic conscience if s/he lives in a way resembling the way that Kant is usually caricatured; e.g., so strict in his walks that his neighbors can set their watches by him.

An agent might be said to have an erring conscience if s/he is very vicious. That is, an agent who never pays any heed to his/her conscience might be said to have an erring conscience. One might think that very vicious agents somehow must have silenced their consciences (how else could they go on like that?), and this kind of erring conscience might be called a silent conscience. One might be said to have an erring conscience if one fails in the kind of diligence described in the previous section. This might be called a negligent conscience, and an example of a negligent conscience would be someone who does not check to make sure that s/he is holding a bee bee gun rather than a rifle and accidentally shoots his/her friend to death in a "prank" on April 1st.[20] An agent might be said to have an ignorant conscience if s/he does not know what to do in a given situation and, without an expert to consult, chooses blindly and wrongly.

The list could go on. All of these things might be thought of as different kinds of errors of

[20] A case like this recently (today's date is 12.01.2010) was reported in the news.

conscience. However, as the text makes clear, Kant is not talking about any of these things in claiming that an erring conscience is an absurdity. Kant does think that some agents strew their paths with man-traps (what I called having a fanatical conscience); he also thinks that some agents feel morbidly guilty all the time (a morbid conscience); he thinks that some agents are inflexible in their application of rules (a rigoristic conscience); and he thinks that it is possible to pay no heed to conscience (a vicious agent); he thinks that it is possible to have what I described as a negligent conscience (Kant illustrates a negligent conscience by describing an inquisitor); and it would be difficult to make sense of Kant's casuistical questions in the *Metaphysics of Morals*, the answers to which do not seem to be intended to be self-evident, if he ruled out the possibility of an ignorant conscience.[21] Kant's claim about an erring conscience is very specific, and the specificity of his claim is brought out even more clearly in his discussion at 8:268. At 8:268 Kant makes the following claim:

> . . . an erring conscience is an absurdity; and, if there is such a thing, then we could never be certain we have acted rightly, since even the judge in the last instance can still be in error. 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, for this judgment — or rather this proposition — merely says that I judge the object in such-and-such a way.(8:268)

In this passage, Kant argues that agents can be mistaken about whether X is permissible, but he argues that they cannot be mistaken about whether they believe X to be permissible. Again, this gives decisive evidence against the natural reading discussed in the introduction; Kant clearly allows for the possibility of subjective rightness not being in agreement with objective rightness. The distinction that Kant is making in this passage is, I think, roughly analogous to what he says in the passage from the *Metaphysics of Morals*. What is perhaps of note in this passage is the

[21] Interestingly, Kant does not think that what I described as a silent conscience is possible. However, as the text makes clear, this is not what he is referring to when he claims that an erring conscience is an absurdity.

apparent inconsistency between the first and the second sentences.

In the first sentence, Kant tells us that if there were such a thing as an erring conscience (there is no such thing; such a thing is an absurdity. But if there were such a thing), then agents never could be certain that they have acted rightly. Agents never could be certain that they have acted rightly because if there were such a thing as an erring conscience, then even the judge in the last instance can be in error.

In the second sentence, Kant tells us that agents can be in error with regard to judgments about whether some action is right. This sort of judgment belongs to the understanding and the understanding is fallible. Indeed, it is precisely because of this that subjective rightness does not need to be in agreement with objective rightness (on Kant's account).

But now it looks like there is a problem, for if agents can err in judgments about whether some action is right, then it is difficult to see how agents ever could be certain that they have acted rightly. Kant seems to be telling us precisely that the judge in the last instance *can be in error*. Kant's argument seems to implode; so far from showing that an erring conscience is an absurdity, the second sentence seems to warrant the conclusion (by *modus tollens*) that an erring conscience is a real possibility.

This inconsistency is of such numbing grossness (to coin a phrase) that it calls for some reconsideration, for Kant's language suggests that he thinks the second sentence supports the first. Thus, he must not take infallibility in judgments of understanding to be required to avoid the absurdity pointed to in the first sentence. Kant's use of the term 'last instance' here is very significant. It shows he is using the court of judgment idea and using features of it (*viz.*, the idea of a court of last instance) that need explanation. I think the issues Kant is raising in the first sentence have to do more with bindingness and authority than they have to do with "error" in the way we normally

think of it. I shall return to this in the next section.

For now, I simply point out that in the second sentence, Kant tells us that agents cannot be mistaken in the exercise of the moral reflexivity function (at least insofar as that function is stripped of consideration of due diligence). Agents cannot be in error in the judgment of whether they have submitted something to practical reason in its role as judge; agents cannot be in error in the judgment about whether they believe (or merely pretend to believe) that something is right. I shall discuss the relevance of this in the next section when I come back to the notion of a conscience as a court of last instance.

3 Acting according to conscience

Given what Kant means by 1 (the claim that an erring conscience is an absurdity), 2 (the claim that agents who act in accordance with conscience have done all that they ought) is all the more striking. That is, given that Kant does not deny the problem of moral knowledge — given that he, in fact, explicitly affirms it in the discussion of an erring conscience — it is striking that Kant adopts such a strong theory of conscience.

In the introduction to this paper, I made the following argument. If an erring conscience is an absurdity, then agents who act according to conscience *eo ipso* act objectively rightly. And if an agent acts objectively rightly, then the agent has done all that morality can require of him/her. But Kant cannot make this argument. As was seen in the previous section, Kant does not think that if an erring conscience is an absurdity, then agents who act according to conscience *eo ipso* act objectively rightly. According to Kant, an agent might act according to conscience but act objectively wrongly if his/her principles are bad. In such a case, the agent has made an error of

understanding. Such errors are possible. The point is that they are not culpable.[22] In the remainder of this section, I shall give two examples to help clarify the issues on the table and I shall discuss what I take to be Kant's argument for 2 (insofar as he has one).

The first example is not one in which the agent has bad principles. Suppose that X goes out to lunch with Y but that X forgets her wallet. X asks Y to spot her ten dollars to pay for lunch and she promises to repay it at lunch next week. Ten dollars is not a huge amount of money, and one might suppose that X and Y are good enough friends that this is not such an extraordinary event. X intends to pay back the money. She even writes the following down in her planner: bring an extra ten dollars to lunch next week to give to Y. But X's memory is pretty bad and her handwriting is even worse, and X forgets the promise. There is nothing nasty here. The example is a one-time event, it is not a chronic or unconscious habit. Nonetheless, some might feel queasy saying that something like forgetting can alter the moral landscape. Presumably X is behaving in accordance with her conscience in this instance; if one ascribes to 2, then X has done nothing wrong.[23]

The second example is one in which the agent has bad principles. Compare, again, Russell and the member of the religious right. The principles of Russell and the member of the religious right are probably diametrically opposed on many issues. I shall not give a specific example of a bad principle. But the basic point to be made is that probably some of Russell's principles are good and some are bad, and the same goes for the member of the religious right. I take it as given that at

[22]Cf. 27:354-355, where Kant makes the following claim: "he who acts according to an errant conscience is acting conscientiously and if he does so, his action might be defective, but [it] cannot be imputed to him as a crime" (clearly Kant is using 'errant conscience' in a different (less technical, less highly theoretical, more popular) sense here. In using the term 'errant conscience', he is referring to the fact that the judgment of understanding that what one does is right might be mistaken. This passage as a whole upholds his consistent view throughout – namely, that an agent cannot be blamed for doing something if s/he acts conscientiously).

[23]This is not to say that X would not be behaving wrongly if, once reminded of the ten dollars by Y, she were to act as though she now owed nothing. It is merely to say that having forgotten about it, X does nothing wrong in not bringing the ten dollars the following week.

least some of either Russell's or the religious right member's principles are bad, for some of their principles are opposed and, thus, some of them must be bad. If one ascribes to 2, then, insofar as they act in accordance with their principles, both Russell and the member of the religious right are behaving permissibly.[24]

I turn now to the argument that comes closest to reaching the claim that if an agent has acted according to his/her conscience, then, as far as morality is concerned, s/he has done all that s/he ought. At 6:402, immediately after making this claim, Kant argues as follows:

> . . . when it comes, or has come, to a deed, conscience speaks involuntarily and unavoidably. Therefore, to act in accordance with conscience cannot itself be a duty; for if it were, there would have to be yet a second conscience in order for one to become aware of the act of the first. (6:402)

It is tempting to read this passage as appealing to the following line of thought: if X is a judgment of conscience but the agent does not follow it, then one must suppose a second conscience that overruled X. But then the first judgment was not a judgment of conscience at all. The idea is not that there could be a second conscience, it is that if something played this role, then the thing called 'conscience' (i.e., the thing that made X, the first judgment of conscience) was not conscience at all. So there is only conscience, no second conscience. In other words, an agent, in exercising conscience, necessarily decides to act in accordance with it. If s/he does not so act, then s/he did not exercise (or not *really* exercise) the function of conscience at all.

The trouble with this argument is that rather than explaining the text, it goes straight against it. Kant admits twice in the passage leading up to this quotation that it is possible for an agent to

[24]This is not to say that the forming of those principles is permissible. It is to say that acting in accordance with the principles once they have been formed is permissible. In any given instance, I might decide that I have accepted such and such principle on the basis of customary morality but now, as judged by conscience, I judge myself (and feel) guilty in following it – which presumably would lead me to repudiate it and follow a different principle even though it is not conscience itself that picks principles. The point is that as long as I followed this principle conscientiously, my behavior was not culpable.

"pay no heed" to the voice of conscience. Thus, the claim that an agent, in exercising conscience, necessarily decides to act in accordance with it is false on Kant's view. So this cannot be Kant's argument.

The idea seems to be, rather, that conscience is ultimately binding in the sense that a judgment of conscience provides an agent with motivation to do something that s/he might not have been motivated to do otherwise. That is, conscience motivates agents to act in accordance with it directly; Kantian conscience does not motivate by appeal to something else, such as the prospect of an afterlife or divine punishment. Of course, an agent might not do that which conscience provides the motivation for the agent to do; agents might not do that which they ought. But conscience speaks unavoidably and unasked, and agents are always motivated to comply (even when they choose not to). Moreover (and this is where the bindingness comes in), in any given instance in which the judgment of conscience weighs in favor of doing X, the agent ought to X; that is constitutive of conscience. There can be no duty to comply with conscience because this would make no sense; it would mean that there would be a duty to do one's duty, which immediately would generate an infinite regress (if it is the case that for all duties X there is a duty to do X, then there is a duty to do X. But then there is a duty to do the duty to do X. And if that is the case, then there is a duty to do the duty to do the duty to do X. Etc.). Let me explain.

On the face of it, if an agent can fail to heed conscience, it looks like there could and should be a duty to heed it, and this could be described as a duty to act in accordance with it. That is, if agents cannot fail to heed conscience (if an agent, in exercising conscience, necessarily decides to act in accordance with it), then, because ought implies can, it is not the case that one ought to heed conscience (by *modus tollens*). But agents can fail to heed conscience, so a duty to heed conscience is not precluded by appeal to "ought implies can". Moreover, if an agent fails to heed

conscience, s/he (*eo ipso*) is doing something that s/he ought not to do. So an agent ought to heed conscience, so there is a duty to heed conscience. What is going on?

What Kant seems to be saying is that there is no duty to heed conscience over and above the duty to do X. That is, if an agent's conscience would condemn him/her for omitting X, then the agent has a duty. The agent has a duty to do X. But the agent does not have a *further* duty to heed conscience and, thus, to do X. There is one duty in this case, and it is the duty to do X. This might be described as a duty to heed conscience insofar as there is a duty to do something and if one does not do that thing, then one will not be heeding conscience. But what Kant is saying is that there is no duty to heed conscience in the sense of a duty to do the duty to do X or a (logically posterior) duty to do that which one (logically prior) ought to do. Plainly, such a duty would generate an infinite regress of the type described, and this might be what Kant is getting at in the passage quoted.

The trouble with this argument is that it goes precisely nowhere in explaining why Kant thinks that an agent who acts according to conscience *eo ipso* has behaved permissibly. I think the explanation of how Kant uses the notion of a court of last instance in the passage explored in the previous section is essential to tying together all of these issues with why Kant thinks some kinds of error (*viz.*, errors of understanding, which might result in a gap between subjective rightness and objective rightness) are irrelevant and why some kinds of error (*viz.*, errors in the moral reflexivity function) are absurd. Clearly in the real world, courts of last instance (e.g., the U.S. Supreme Court) can err. But the legal system must treat their judgments as finally valid, even though independent agents might think that these judgments are (in some instances) monstrous. I think Kant's idea is that conscience serves for a moral agent in some such "last instance" function.

This means that if an agent acts according to his/her conscience, then the agent has done all

that can be expected of him/her in the sense that s/he must be judged by him/herself (and should be judged by others) as doing what s/he ought as far as moral accountability is concerned. I think Kant's view is the following: To suppose that there could be no "court of last instance" in this sense, that I could have made as good a judgment as was possible for me about whether I am guilty or innocent in doing A and yet not have my judgment be binding, so that I honestly might think that I am innocent even though really I am guilty – this would mean that no agent ever could be in a position to be finally responsible or accountable.

Analogously, to assume a judicial system in which there is no court of last instance whose judgment counts as legally valid is to assume a judicial system that is incapable of performing its essential function. In both cases, agents could revisit the question later. In the legal case, they could not reverse the judgment of last instance as far as its effect is concerned (at least on the specific case under review, although they might on later cases, as when the Supreme Court reverses itself); in the moral case, they could not judge the agent blameworthy for having done the thing (even if the action is judged objectively wrong). Even if the agent him/herself later decides that what s/he did was objectively wrong, Kant thinks s/he should not blame him/herself for having done it if s/he was judged not guilty in doing it at the time by a judgment of his/her conscience.

This explains why Kant tells us that agents cannot be mistaken in the exercise of the moral reflexivity function. If agents could be in error about this, then the idea of conscience as a "court of last instance" would make no sense. This idea would make no sense because this just is part of the judgment that has finality in the sense just described.

The basic point is that agents must take the judgments of conscience (as the court of last instance) to be authoritative (even when they do not act accordingly), just as a criminal must take the law as authoritative even when s/he breaks it, and even when those judgments (unbeknownst

to the agent) reflect false principles, just as when the Supreme Court issues a decision with which one disagrees. If they do not do this, then their autonomy is undermined in the sense that they are not capable of being self-governing agents. Thus, if agents did not take conscience to be a court of last instance, they would not be in a position to be finally responsible or accountable.

As a textual argument, this seems to me to be the best one can do. The notion of conscience is thus seen to be central to Kant's theory. However, I suspect that those who are not already sympathetic to Kant's project will not like this argument. I suspect that those who are not already sympathetic to Kant's project will not accept the idea that self-governance and, thus, conscience as a court of last instance is required for being finally accountable; they even might argue that the concept of self-governance itself is an oxymoron.

Perhaps Kant knew this already, and perhaps this is why Kant does not try to argue for 2. 2 is a starting point of Kant's theory rather than an ending point. What is clear is that Kant thinks that agents can adopt false principles of conscience; subjective rightness and objective rightness can come apart; and, according to Kant, if an agent follows his/her conscience, then s/he has done all that s/he ought as far as morality is concerned.[25]

References

[1] Adams, Robert Merrihew. "Involuntary Sins." *Philosophical Review*, 94 (1985), 3-31.

[2] Hardwig, John. "Action from Duty but not in accord with Duty". *Ethics*, Vol. 93, No 2 (Jan., 1983), 283-290.

[25] This, of course, leaves it an open question as to whether such an agent has done all that s/he ought as far as legality is concerned. If, according to Kant, it is not the case that such an agent necessarily has done all that s/he ought as far as legality is concerned, then it seems that Kant thinks that agents can be legally culpable for something even if they are not morally culpable for anything. This recalls concepts such as strict liability and absolute liability. But I obviously cannot discuss this here.

[3] Herman, Barbara. *The Practice of Moral Judgment*. Cambridge, Massachusetts, Harvard University Press, 1993.

[4] Hill, Thomas E., Jr. *Respect, Pluralism, and Justice: Kantian Perspectives*. Oxford, Clarendon Press, 2000.

[5] Hill, Thomas E., Jr. *Human Welfare and Moral Worth: Kantian Perspectives*. Oxford, Clarendon Press, 2002.

[6] O'Neill, Onora. *Constructions of Reason: Explorations of Kant's Practical Philosophy*. New York, NY, Cambridge University Press, 1989.

[7] Skorupski, John. "Conscience". In John Skorupski (ed.), *The Routledge Companion to Ethics*, 550-562. Routledge, 2010.

[8] Timmons, Mark. "Decision Procedures, Moral Criteria, and the Problem of Relevant Descriptions in Kant's Ethics", *Jahrbuch für Recht und Ethik*, 5 (1997), 389-417.

[9] Wood, Allen. *Hegel's Ethical Thought*. Cambridge, Cambridge University Press, 1990.

[10] Wood, Allen. *Kant's Ethical Thought*. Cambridge, Eng.: Cambridge University Press, 1999.

[11] Wood, Allen. *Kantian Ethics*. Cambridge, Eng.: Cambridge University Press, 2008.