# Non-accidental rightness and the guise of the objectively good

**Abstract**

My goal in this paper is to show that two theses that are widely adopted among Kantian ethicists are irreconcilable. The paper is divided into four sections. In the first, I briefly sketch the contours of my own positive view of Kantian ethics, concentrating on the issues relevant to the two theses to be discussed: I argue that agents can perform actions from but not in conformity with duty, and I argue that agents intentionally can perform actions they take to be contrary to duty. In the second, I focus on Barbara Herman's non-accidental rightness condition from "On the Value of Acting from Duty." In the third, I focus on Christine Korsgaard's guise of the objectively good from "Formula of Humanity." In the fourth, I explain why the positions developed by Herman and Korsgaard are irreconcilable and I make a suggestion about how to move forward.

**Keywords**

Non-accidental rightness condition; guise of the objectively good; guise of the good; evil; Kantian ethics; the problem of moral knowledge; Herman; Korsgaard

## Introduction

My goal in this paper is to show that two of the most influential articles by two of the most influential Kantian ethicists of the last 40 years are irreconcilable because the conjunction of their signature theses makes moral evil impossible.

The first article is Herman's "On the Value of Acting from the Motive of Duty," published first in *The Philosophical Review* in 1981 and then again in 1993 as chapter 1 of *The Practice of Moral Judgment*. The thesis on which I shall focus from that article is that action from duty is non-accidentally in conformity with duty, the non-accidental rightness condition. The second article is Korsgaard's "Formula of Humanity," published first in *Kantian Review* in 1986 and then again in 1996 as chapter 4 of *Creating the Kingdom of Ends*. The thesis on which I shall focus from that article is that rational agents necessarily represent their ends as objectively good, the guise of the objectively good thesis.

My paper is divided into four sections. In the first, I briefly sketch the contours of my own positive view of Kantian ethics, concentrating on the issues relevant to Herman's and Korsgaard's articles. In the second, I focus on Herman and, in particular, her non-accidental rightness condition. In the third, I focus on Korsgaard and, in particular, her thesis about how rational agents represent their ends. In the fourth, I explain why the positions developed by Herman and Korsgaard make moral evil impossible, and I make a suggestion about how to move forward: I suggest rejecting both theses, and I briefly chart out some of the puzzles that this rejection raises.

## Section 1. Sketching the contours of my own view

One of Herman's signature theses is the non-accidental rightness condition. This condition, advanced in her 1981 article, "is (among Kantians) uncontroversial" (Baron 1995, 174).[1] As we shall see in section 2 of this paper, this condition states that, if an action is from duty, then it is in conformity with duty. Thus, this condition rules out the possibility of action that is from duty but contrary to duty.

[1] I owe this reference to (Sverdlik 2001, 303).

One of Korsgaard's signature theses is the guise of the objectively good. This thesis, advanced in her 1986 article, has been adopted by other prominent Kant scholars and Kantians, including, for example, Allison (1990, 91); Engstrom (1992, 760); and Wood (1999, 129; 2008, 91). As we shall see in section 3 of this paper, this thesis states that agents always pursue ends that they represent as objectively good. As such, this thesis rules out the possibility of an agent acting on a maxim that is contrary to duty except through a mistake.

In section 4 of this paper, I shall explain not only why I think these two theses are irreconcilable but also why I think that neither of these theses should be ascribed to Kant. In particular, I argue that the conjunction of the non-accidental rightness condition with the guise of the objectively good makes moral evil impossible and, therefore, that the reality of evil requires jettisoning at least one of these theses. I then argue that (a) the textual evidence that is given as grounds for ascribing these theses to Kant does not withstand critical scrutiny and, thus, the evidence for reading Kant as subscribing to either of these theses is weak. Further, I argue that (b) there is textual evidence for ascribing theses to Kant that are inconsistent with the non-accidental rightness condition and the guise of the objectively good and, thus, the evidence *against* reading Kant as subscribing to either the non-accidental rightness thesis or the guise of the objectively good is strong.

The motivation for all of this is threefold: first, I want to set the historical record straight; second, I think both of these theses are false; and third, I think that it is a bad thing for Kant scholarship and Kantian ethics that two of their central theses are irreconcilable. So let me say more about my own positive view.

As I read Kant, agents can act from duty but not in conformity with duty, and agents can adopt ends that they take to be contrary to duty and, thus, not objectively good. Moreover, the reason why agents can act from duty but not in conformity with duty, as I read Kant, entails that agents can adopt ends that they *mistakenly* take to be contrary to duty. From this it may be seen that my rejection of the non-accidental rightness condition and the guise of the objectively good makes possible two kinds of puzzle case: (1) agents who try to act in conformity with duty but fail (agents who act contrary to duty despite their best efforts) and (2) agents who try to act contrary to duty but fail (agents who act in conformity with duty despite their best efforts).

Although I shall not be able to address here all the nuanced ways in which these puzzle cases can be instantiated, I do want to make three general claims about them. First, agents who act contrary to duty only because of a faulty moral judgment, assuming that they are sincere and have done their due diligence, have done all that morality can require of them. Second, agents who act in conformity with duty only because of a faulty moral judgment are behaving impermissibly insofar as they *try* to act contrary to duty. And third, there might be some actions that are contrary to duty but that cannot be performed by a sincere agent who has done her due diligence and is trying to act in conformity with duty. As we shall see in section 4, these three claims all are taken directly from Kant or from the extant student notes from Kant's lectures, and together they form a triad that is inconsistent with both the non-accidental rightness condition and the guise of the objectively good.

## Section 2. Herman's non-accidental rightness condition

Herman's goal in "On the Value of Acting from the Motive of Duty" (henceforth: *Value*) is to explain and defend Kant's conception of moral worth. The motive of duty comes in because, according to Herman, "an act has moral worth if but only if it is done from the motive of duty" (Herman 1993, 1).

In order to accomplish her goal, Herman divides *Value* into 4 sections. In section one, she explains why action in conformity with duty that is also from duty has a special moral worth. In section two, Herman turns to action in conformity with duty in the presence of other motives. She argues that the mere presence of nonmoral motives does not undermine the moral worth of an action provided that the nonmoral motives are not determinants of the action. In section three, Herman discusses the way in which the motive of duty can play a governing role in all actions: although a merely permissible action cannot be done *from* duty, nonetheless the moral motive can play a limiting role in such an action. In section four, Herman then returns to Kant's *Groundwork for a metaphysics of morals* in order to show that her reading of Kant is plausible.

Of prime importance for present purposes is that, according to Herman, the reason the motive of duty is special is that, if an agent acts from the motive of duty, then it is not an accident that her action is in conformity with duty. This is the non-accidental rightness condition.

Herman states this condition in a variety of ways. For example, she says that the motive of duty makes an action's conformity with duty "the nonaccidental effect of the agent's concern" (Herman 1993, 6). By way of contrast, Herman says the following about nonmoral motives:

> [...] nonmoral motives may well lead to dutiful actions, and may do this with any degree of regularity desired. The problem is that the dutiful actions are the product of a fortuitous alignment of motives and circumstances. (Herman 1993, 6)

For example, if an agent acts from the inclination to help others rather than from duty, she might help a thief, murderer, or rapist to carry out her nefarious plans: "the class of action that follows from the inclination to help others is not a subset of the class of right or dutiful actions" (Herman 1993, 5). Moreover, this is true of any motive other than duty; the class of action that follows from any non-moral motive is not a subset of the class of right or dutiful actions. Not so with the motive of duty: the class of action that follows from the motive of duty, according to Herman, *is* a subset of the class of right or dutiful actions. Thus, an action that is done from duty is non-accidentally right.

According to Herman, any model of Kant's account of moral worth must be able to capture its "moral point," namely: "that a right or dutiful action is performed is the non-accidental effect of the agent's moral concern" (Herman 1993, 8). Herman dismisses the moral worth of actions when nonmoral motives are determining on these grounds: in such cases, it is only an accident of circumstance that an action in conformity with duty was performed, for it is merely an accident of circumstance that, for instance, the person one is sympathetically inclined to aid is engaged in a morally innocuous action rather than, say, struggling to carry out some more devious purpose.

Herman also uses the non-accidental rightness condition to explain why the presence of nonmoral motives is consistent with moral worth provided that these motives are *not* determining:[2]

---

[2] Herman, unlike me, prefers to use the word 'motive' to refer to a ground of action that is determining. So, Herman would say that a non-determining motive, the kind of motive referred to in the sentence to which this note is appended, is an oxymoron. The dispute, however, is merely terminological.

> If the agent acts from the motive of duty, he acts because he takes the fact that the action is morally required to be the ground of choice. It does not follow from this that the action's moral worth is compromised by the presence of nonmoral feelings or interests, so long as they are not taken by the agent as grounds of choice [...thus, even in the presence of nonmoral motives one can say] that an agent's doing the right thing is non-accidental because he acted from the motive of duty [...] (Herman 1993, 12)

The reason I am spending so much time with this is that the non-accidental rightness condition brings some (heavy) baggage with it. In particular, actions from duty are a subset of actions in conformity with duty only if the motive of duty cannot attach to an action that is not in conformity with duty. In other words, if the non-accidental rightness condition is correct, then the motive of duty must be infallible in some sense.

One way to defend and elaborate on this infallibility would be to argue that an agent's moral beliefs are infallible. That is, the infallibility of the motive of duty can be grounded in the infallibility of moral beliefs. This strategy, which is suggested by Hardwig, involves a denial of what might be called the problem of moral knowledge. Here is how Hardwig puts it: "I wish to suggest that Kant may very well have held that there is no moral problem of knowledge and hence no actions from duty but not in accord with duty" (Hardwig 1985, 283). Hardwig's suggestion in this quotation, as a defense of Herman's non-accidental rightness condition, can be put more formally as follows: (1) if an agent is acting from duty, then she believes that her action is a duty; (2) if an agent believes that her action is a duty, then her action is a duty; (3) if an action is a duty and is performed from duty, then it has moral worth; therefore, (4) if an agent is acting from duty, her action has moral worth.

An example will help to illustrate what Hardwig has in mind here. Suppose that I am a merchant and that I give a customer correct change from duty. Then, according to this argument, I must believe that giving correct change is, at least in this instance, my duty. Because there is no problem of moral knowledge, this belief is infallible: if I believe that giving correct change is my duty, then it is so. Therefore, in acting from duty, I am performing my duty, and my action has moral worth.

Of course, each part of this argument can be challenged and, indeed, I am going to do just that later on in this paper. But, for present purposes, I merely would like to note three things.

First, if this argument is accepted, then the motive of duty is indeed infallible. Thus, if, as Hardwig suggests, we are prepared to deny the problem of moral knowledge, then there is a path, even if it might be challenged, to Herman's non-accidental rightness condition.

Second, Herman comes close to following this path in *Value*. That is, Herman comes close to grounding the infallibility of the motive of duty in the infallibility of beliefs about duty, as per Hardwig's suggestion, in *Value*. This may be seen in the following passage:

> For an action to *have* moral worth, moral considerations must determine how the agent conceives of his action (he understands his action to be what morality requires), and this conception of his action must then determine what he does […] That is, an action has moral worth if it is required by duty and has as its primary motive the motive of duty. (Herman 1993, 16)

Herman's appeal to an agent's understanding and to the way in which an agent "conceives of his action" in this passage suggests that, perhaps unwittingly, she is meandering toward a denial of the problem of moral knowledge and, thus, an affirmation of premise 2 in my reconstruction of Hardwig's argument (i.e., the argument that there is no problem of moral knowledge and hence no actions from duty but not in accord with duty).

Third and finally, despite the presence of this passage in *Value*, I do not want to saddle Herman with such a view. That is, I do not way to suggest that, according to Herman, our beliefs about our duties are infallible, much less that this infallibility in our beliefs is what grounds the infallibility in our moral motives. I do not want to do so in general, and I do not want to say that she was committed to it at the time of her writing *Value*. Regardless of whether Herman thinks that *beliefs* about duty are infallible, the point is that she thinks that the *motive* of duty is infallible. That much is encapsulated in the non-accidental rightness condition, which, again, as noted at the beginning of section 1 of this paper, is, among Kantians, uncontroversial. In other words, according to Herman and many other Kantians, the motive of duty, whether in its limiting role or its determining role, leads always and non-accidentally to right action: it is impossible to act from duty but not in conformity with duty.

## Section 3. Korsgaard and the representation of ends

Korsgaard's goal in "Formula of Humanity" (henceforth: *Humanity*) is to explain and defend Kant's humanity formulation of the Categorical Imperative. In particular, it is here that Korsgaard first sets out the so-called regress argument that since has come to be a cornerstone of her vision of Kantian ethics.[3]

*Humanity* is divided into six sections. After the first section, which is introductory, in the second section Korsgaard explains what she thinks Kant means by 'humanity' (namely, the capacity to set ends). In the third section, Korsgaard discusses the theory of rational action on which the regress argument is based. She then explicates and defends the regress argument in section four. In section five, Korsgaard gives examples of what it means to treat humanity as an end in itself. And in the sixth section, Korsgaard gives some last textual evidence from the *Critique of the power of judgment* in favor of her reading of Kant.

The most important thing for current purposes is the theory of rational action on which Korsgaard builds the regress argument. In particular, I want to focus on Korsgaard's claim that "[i]nsofar as we are rational agents we will choose what is good—or take what we choose to be chosen as good" (Korsgaard 1996, 115).

Korsgaard takes goodness to go hand-in-hand with having reasons, and she takes reasons to be essentially universal in the sense that they apply to all rational beings.[4] Moreover, she takes having such reasons to be necessary and sufficient for a choice to be justified.[5] Thus, the claim on which I want to focus may be paraphrased in terms of reasons or justification: insofar as we are rational, we choose what we have a reason to choose, what we are justified in choosing—or, at least, we take what we choose to be chosen as reasonable or justified.

---

[3] In its original formulation, Korsgaard took this argument to show that rational nature, or the capacity to choose rationally, is the unconditioned condition of all objective goodness and, therefore, that rational beings are deserving of respect. In her most recent work, Korsgaard has sought to extend this argument to show that there are also duties to nonrational sentient beings (see, e.g., Korsgaard 2004).

[4] Consider: "if an end is deemed good it provides reasons for action that apply to every rational being" (Korsgaard 1996, 115).

[5] This may be seen from the following passage: "Since good is a rational concept, a good end will be one for which there is reason—an end whose existence can be *justified*" (Korsgaard 1996, 116).

It is important not to misunderstand what Korsgaard is saying here. Some philosophers advocate for a related thesis, a thesis that Raz calls the Guise of the Good (GG). According to GG, agents always take their choices to be good in some sense, but they might not take their choices to be good overall: an agent might act "for what he took to be the lesser reason" (Raz 2008, 22n22). Korsgaard's claim, however, is considerably stronger than GG.

As Korsgaard is quick to explain, when she claims that rational agents choose what is good or take what they choose to be chosen as good, she is talking about a very specific kind of goodness: objective goodness. Thus, only one page after first making the claim on which I am focusing, Korsgaard says: "when we act under the direction of reason, we pursue an end that is objectively good" (Korsgaard 1996, 116). To disambiguate Korsgaard's claim from GG, I shall refer to the former as GOG.

According to Korsgaard, X is objectively good if but only if it is unconditionally good or it is conditionally good and its conditions are met.[6] Again, objective goodness goes hand-in-hand, according to Korsgaard, with reasons and with justification. So, it may be seen that GOG is meant to be equivalent to saying that agents always choose in accordance with what they take to be the *strongest* reason or, more simply, that agents always take themselves to be fully justified in their choices. Thus, if GOG is true, then Raz's claim (that an agent might act for the lesser reason) is not.

Note also that Korsgaard takes rationality and morality to be coextensive: duties and permissions track reasons and *vice versa*.[7] So GOG also is meant to entail that, insofar as they are rational, agents always take themselves to be acting in conformity with duty (Korsgaard 1996, 118-119).

Now obviously this is not going to be true on just any understanding of "duty." A cultural relativist might protest that she acts contrary to duty all the time and that she has good reason to do so. But this need not be an objection to Korsgaard: this could be a merely verbal dispute about how to use the word 'duty', not a dispute that I am going to pick up here.

---

[6] "A thing, then, can be said to be objectively good, either if it is unconditionally good or if it is conditionally good and the condition under which it is good is met" (Korsgaard 1996, 118).

[7] It is perhaps worth pointing out that Herman would accept this too. As evidence, consider the title of her dissertation: *Morality as Rationality*.

Moreover, GOG need not be interpreted as a claim about beliefs or conscious, cognitive representations. To be sure, GOG *could* be interpreted as saying that an agent will choose something only if she *believes* it to be objectively good (in the same way that the non-accidental rightness condition *could* be interpreted and defended as a denial of the problem of moral knowledge). But, GOG also can be interpreted in a weaker sense: GOG can be interpreted as ascribing to agents the underlying maxim always to act in accordance with reason. This weaker thesis is strong enough for my purposes, so I shall not engage in dispute about whether Korsgaard is committed to anything more.

However, one might wonder about the condition which Korsgaard prefixes to her claim: "insofar as we are rational agents." If, as I suggested in the previous paragraph, Korsgaard's claim is interpreted in terms of maxim ascription, then perhaps the idea is that this maxim should be ascribed to agents only insofar as they are acting from reason. Perhaps, to continue this hypothesis, GOG is consistent with saying that sometimes agents do not have this maxim: sometimes agents do not take their ends to be objectively good (namely, when they are not acting from reason).

Support for this hypothesis comes from the other way in which Korsgaard prefixes her claim: "*when we act under the direction of reason*, we pursue an end that is objectively good" (Korsgaard 1996, 116, my emphasis). However, when this claim is placed in context, it may be seen that something else is going on:

> […] when we act under the direction of reason, we pursue an end that is objectively good. But human beings, who act on their conception of laws, take themselves to act under the direction of reason. In the argument for the Formula of Humanity, as I understand it, Kant uses the premise that when we act we take ourselves to be acting reasonably and so we suppose that our end is, in his sense, objectively good. (Korsgaard 1996, 116)

Notice what Korsgaard says here. When, as a matter of fact, we act under the direction of reason, then, as a matter of fact, we pursue an end that is objectively good. But, we humans always *take ourselves* to act under the direction of reason, whence it follows that we always *take ourselves* to pursue ends that are objectively good, and *that* is the idea underlying GOG.

Of course, sometimes we fail. In those cases, pretense peels away from reality; our end is not *actually* objectively good. Nonetheless, according to GOG, the pretense is always there: humans always *take themselves* to be acting reasonably. GOG says that, because we are rational beings, we always choose ends that we *suppose* to be objectively good. It is just that sometimes we make mistakes: sometimes, what we take to be objectively good is not so. Indeed, this idea has become a centerpiece of Korsgaard's larger approach to ethics, an ethics built on the concept of internal norms of activity, norms that cannot be disengaged from because they are constitutive of action.[8]

## Section 4. Irreconcilability and moving forward

Now I want to show that Korsgaard and Herman are irreconcilable. But, given the foregoing, that is, I think, easy: if the motive of duty in its determining and limiting role is infallible (à la Herman and the non-accidental rightness condition) and if agents always choose ends that they take to be objectively good (à la Korsgaard and GOG), then there is no room for evil action. Let me explain.

If we accept the non-accidental rightness condition, then evil cannot arise from a cognitive failure or from a general failure to act in conformity with duty *despite* one's best efforts because such an aim is bound (non-accidentally) to hit its mark. But, if we accept GOG, then evil

[8] Consider: "the only way to establish the authority of any purported normative principle is to establish that it is constitutive of something to which the person whom it covers is committed" (Korsgaard 2009, 32).

also cannot arise from a volitional failure or from a general failure to act in conformity with duty *because* of one's best efforts because there are no such aims.[9]

But there is evil action all the same, and any plausible moral theory must be able to account for it. Thus, the non-accidental rightness condition and GOG, centerpieces of Herman's and Korsgaard's work, some of the most influential recent work in Kantian (if not Kant's) ethics, are irreconcilable. And I suggest that the way forward is to take a step back to Kant.

The remainder of this section is divided into three subsections. In the first, I examine the textual evidence for and against Herman's non-accidental rightness condition. In the second, I examine the textual evidence for and against Korsgaard's GOG. In the third, I look at the philosophical ramifications of the position I am ascribing to Kant.

---

[9] Herman, as a matter of fact, generally takes moral error to arise in a failure of will whereas Korsgaard, as a matter of fact, generally takes it to arise in a failure of representation (usually precipitated by an a-rational, pathological process of self-deception).

It might be objected that Korsgaard cannot take moral error to arise in a failure of representation on account of her interpretation of Kant's universalization test. The idea here is that (1) Korsgaard advocates a practical contradiction interpretation of the universalization test; (2) according to the practical contradiction interpretation, a maxim fails the universalization test because it involves a failure to will certain ends; (3) a failure to will ends is not a cognitive failure; (4) a failure of representation is a cognitive failure; therefore, (5) according to Korsgaard's practical contradiction interpretation, moral error is not a failure of representation.

However, this objection does not work. For one thing, (2) is false. According to the practical contradiction interpretation, a maxim fails the universalization tests because, if the maxim were a universal law, then, either the hypothetical imperative on which it is based would be false, or an essential end no longer could be realized. That is, according to the practical contradiction interpretation, there is a contradiction in conception if the maxim's end no longer could be attained through the maxim's means in a world in which the maxim is a universal law, and there is a contradiction in willing if some end that the agent cannot rationally foreswear is unattainable in a world in which the maxim is a universal law (Korsgaard 1996, chapter 3). Thus, according to the practical contradiction interpretation, it is not the case that a maxim fails the universalization test because it involves a failure to will certain ends: (2) is false.

For another thing, (3) is false. As Korsgaard reads Kant, moral error arises from the adoption of impermissible, unjustified ends. And, according to Korsgaard, this occurs by means of a cognitive error, a failure of representation (the agent represents the end as objectively good even though it is not so), for agents necessarily represent their ends as objectively good. On Korsgaard's account, this failure of representation usually is precipitated by an a-rational, pathological process of self-deception. Thus, a failure to will certain ends (or, conversely, the willing of certain ends) can be rooted in cognitive failure or a failure of representation: (3) is false.

## Section 4.1. Kant's disavowal of the non-accidental rightness condition

There is textual evidence for Herman's non-accidental rightness condition, especially in part I of the *Groundwork for a metaphysics of morals*, the section of text with which Herman is most concerned in *Value*. For example, Kant says that he is going to "pass over" all actions that are already recognized as contrary to duty: even though such actions might be useful for some purposes, "with them there is not even once a question whether they might happen *from duty* since they indeed conflict with it" (GMS, AA 04: 397.11-14). In other words, if an action is not in conformity with duty, then it is not from duty. Take the contrapositive of this and we have Herman's non-accidental rightness condition.

Moreover, Kant comes close to a denial of the problem of moral knowledge in some passages. For instance, in the *Critique of practical reason* Kant says: "which form in a maxim renders it suitable for universal lawgiving, and which not, that [is something] the commonest understanding can distinguish without instruction" (KpV, AA 05: 27.21-22). This passage might be read as saying that everyone has an innate, infallible ability to judge whether a maxim is suitable for universal lawgiving and, thus, conforms to the categorical imperative. Similarly, about ten academy pages later Kant says that "what is to be done according to the principle of autonomy of choice is, [even] for the commonest understanding, to be recognized totally easily and without bethought [*Bedenken*]" (KpV, AA 05: 36.28-29).[10] Some might read this as saying that everyone (even the commonest understanding) is able to judge infallibly, easily, and without reflection, what the principle of autonomy and, thus, the categorical imperative, requires. Such a reading is bolstered by a claim Kant makes two pages later: "what duty is, presents itself to everyone of itself" (KpV, AA 05: 38.31). Indeed, it is precisely these kinds of passages that Hardwig has in mind when he claims, as we saw at the end of section 2 above, that, according to Kant, there is no problem of moral knowledge and, hence, no action from duty but not in conformity with duty. Thus, there is some textual basis for both Herman's non-accidental rightness condition and for

[10] The German runs as follows: "*Was nach dem Princip der Autonomie der Willkür zu thun sei, ist für den gemeinsten Verstand ganz leicht und ohne Bedenken einzusehen.*" I find the second clause, after the comma, challenging. "*Ganz leicht*" and "*ohne Bedenken*" seem to be describing the process of "*einzusehen*." But this is most naturally done in english, I think, by changing the active "*einzusehen*" into the passive. Accordingly, I have taken that liberty in my translation. But it is a liberty with which I am not at peace.

Hardwig's suggestion that the rationale for the non-accidental rightness condition is the denial of the problem of moral knowledge.

But, when push comes to shove, Kant not only does not deny the problem of moral knowledge, as per Hardwig's suggestion; he openly admits it. For instance, in his discussion of conscience in the *Metaphysics of morals*, Kant says explicitly that "in the objective judgment, whether something is a duty or not, one can well err from time to time" (MS, AA 06: 401.5-6). If we were tempted, on the basis of the passages cited in the previous paragraph of this paper, to ascribe to Kant the view that moral knowledge is acquired through an infallible faculty of rational intuition, this claim about error at least should give pause. And if an agent can err from time to time in the objective judgment of whether something is a duty, then the non-accidental rightness condition must be jettisoned, for if an agent is mistaken in her judgment of whether an action is a duty--if an agent judges that she ought to D even though D is impermissible--then she might act from duty but not in conformity with duty.

Moreover, a close reading of the passage from the *Groundwork for a metaphysics of morals* from which the evidence for the non-accidental rightness condition is drawn suggests an alternative interpretation. As noted above, Kant says that actions "which already are recognized as contrary to duty" cannot be performed from duty (GMS, AA 04: 397.11-12). But this supports the non-accidental rightness condition only if the agent doing the recognizing is not the same agent as the one performing the action. If, by way of contrast, we read Kant as saying that, if person P recognizes that action A is contrary to duty, then P cannot perform A from duty, then Kant's claim is, to be sure, *consistent* with the non-accidental rightness condition. But it also is consistent with the denial of the non-accidental rightness condition and, in fact, one might think that an implicature (even if not an implication) of this reading is that the non-accidental rightness condition is *false*. However, even if this implicature is denied, I think my case is quite strong: there are good grounds (from Kant's theory of conscience) for thinking that Kant would accept the problem of moral knowledge and deny the non-accidental rightness condition, and the textual evidence in favor of ascribing the denial of the problem of moral knowledge or the acceptance of the non-accidental rightness condition is equivocal at best.

Now it might be thought that, although Kant allows for mistaken assessments of the deontic status of an action and, thereby, allows for the possibility of action from duty but not in

conformity with duty, this allowance was an error, for the values that lie at the core of his ethical system require him to say otherwise. For example, Hardwig thinks that Kant's commitment to universality and autonomy require him to deny the problem of moral knowledge:

> Kant [...] needs to maintain that moral principles and judgments are both autonomous and universally valid. But he can maintain that they are both, only by denying the possibility of mistaken moral judgment. (Hardwig 1983, 288)

That is, according to Hardwig, there is no plausible way for our duties to be universal *and* autonomous if agents can be mistaken in their moral judgments, and Kant maintains that our duties are both universal and autonomous.

The problem with Hardwig's argument is that he misunderstands Kant's standard of universality. Kant's famous universalization formulation of the Categorical Imperative tells agents to act only on maxims that *they* can will *at the same time* as universal laws, and this is different from telling an agent to act only on maxims that *are* universal laws.[11]

Hardwig foresees this problem and answers that it solves the difficulty only if "the principle of universalizability is itself a universal principle" (Hardwig 1983, 289). If the principle of universalizability is not itself a universal principle, then this principle (and its commands) might, in some instances, fail the test of autonomy.

I am going to return to this at the end of this section. For now I point out only that, even if Hardwig is right about this, it is weak enough to allow for mistaken judgments about the deontic status of an action and, therefore, to allow for action that is from duty but not in conformity with duty. Thus, even if Hardwig is right, enough has been conceded to show that the nonaccidental rightness condition is mistaken. I turn now to Korsgaard's GOG.

## Section 4.2. Kant's disavowal of GOG

Just as there are passages in Kant's corpus that speak in favor of Herman's non-accidental rightness condition, so there are passages in Kant's corpus that speak in favor of Korsgaard's GOG.

---

[11] For an extended discussion of this difference and its implications, see (Kahn 2014).

For example, in part II of the *Groundwork for a metaphysics of morals*, Kant says that "practical *good*" is "what determines the will, conveyed by means of representations of reason" (GMS, AA 04: 413.18-21). Some might read this as saying that our will is always determined by what we take ourselves to have most reason to do. In chapter II of the Analytic of Practical Reason in the *Critique of practical reason*, Kant says that "we want under the direction of reason nothing, except insofar as we hold it for good or evil" (KpV, AA 05: 60.6-7). This might be read as an endorsement of the thesis that, insofar as we are rational, we always pursue the objectively good. And in section 4 of the first "Moment" of the Analytic of the Beautiful in the *Critique of the power of judgment*, Kant remarks that "the Good is the object of the will" (KU, AA 05: 209.9).

However, a close inspection of these passages reveals that the support they offer for ascribing GOG to Kant is tenuous. Kant not only thinks that agents sometimes make mistakes about objective goodness and, thus, pursue objectively impermissible ends despite their best efforts; he also thinks that agents sometimes pursue ends that they do not take to be objectively good. This, I believe, becomes evident on close inspection of the texts referred to in the previous paragraph. But it also is on full display, again, in Kant's discussion of conscience, the faculty of "practical reason holding before a man his duty in every case of law" (MS, AA 06: 400.27-28).

In discussing conscience Kant declares that "when one therefore says: this person has no conscience, one means thereby this: he does not turn himself to the demand of it" (MS, AA 06: 400.31-33). And he then clarifies that "lack of conscience [*Gewissenlosigkeit*] is not a deficiency [*Mangel*] of conscience, but rather a propensity not to turn oneself to its judgment" (MS, AA 06: 401.10-11). Given Kant's account of the function of conscience and his account of what it means to say that someone has no conscience, it may be inferred that he thinks that agents sometimes do not act in conformity with what they take to be their duties. Or, in Korsgaard's language but on Kant's account, agents do not always take their ends to be objectively good.

Some might object that I am attributing an overly strong position to Korsgaard and, therefore, that I am committing a straw man fallacy. According to this objection, Korsgaard would admit that Kantian agents do plenty of things, like going to the bathroom, that involve ends that they do not take to be objectively good. So, Korsgaard cannot think that agents always take their ends to be objectively good. Thus, although I might be right that, according to Kant, agents do not always take their ends to be objectively good, this is not a problem for Korsgaard. Rather, it

is a problem for my reading of Korsgaard: I am attributing an overly strong position to Korsgaard and some work should be done to see a more accurate, and plausible, version of her interpretation.

However, this objection does not work. Korsgaard not only explicitly subscribes to GOG, the thesis that agents represent their ends as objectively good, but, more, she has an argument for it: as we saw in the block quotation reproduced in the penultimate paragraph of section 3 of this paper, Korsgaard asserts that (1) to act under the direction of reason is (successfully) to pursue an end that is objectively good, and (2) agents always take themselves to be acting under the direction of reason. The reason the objection does not work is that it misunderstands GOG. To take one's end to be objectively good is not to take one's ends to be morally obligatory; it is, rather, to take one's ends to be morally permissible. Thus, GOG is consistent with doing things like going to the bathroom or reading a book, or any of the other permissible actions that take up most of the day. GOG is a perfectly respectable thesis from moral psychology that traces back at least to Plato's Socrates. The point I am making in this subsection is merely that Kant does not subscribe to it.

## Section 4.3. What now

As we have seen in the previous two subsections of this paper, Kant seems to disavow both the non-accidental rightness condition and GOG. Further evidence for this can be derived from his *Nachlass*. For example, consider the following note written in the mid-1770s

> If the first grounds of morality rest on reason, then it is a question whether departure from the teachings of morality are to be attributed to error or to evilness of the will…False moral judgment is to be attributed to the weakness of reason (against prejudices of self-love); action contrary to these judgments is to be attributed to the powerlessness of reason over the inclinations. (Refl, AA, 19: 133.13-19)

In this passage, Kant suggests, first, that we can be mistaken in our moral judgments, which, as argued above, undermines the non-accidental rightness condition (that is, if the prejudices of self-love influence my moral judgment such that I convince myself that D is my duty, then I might D from duty even though it is actually contrary thereto); and Kant suggests, second, that we can act contrary to moral judgment (when inclination overpowers reason), which, as argued above, undermines GOG.

But, as pointed out above, the rejection of the non-accidental rightness condition and GOG complicates matters considerably. In particular, it opens up the doors to the possibility of an agent who does something that is not objectively good despite the fact that she is aiming at the good. It also opens up the doors to the possibility of an agent who does something objectively good despite the fact that she is not aiming at the good.

Kant was well aware of the first possibility. In the *Metaphysics of morals*, almost immediately after declaring that an agent can make a mistake in his objective judgment about whether something is a duty, Kant says that "when however someone is fully aware of himself as having acted according to conscience, then nothing more can be demanded of him which pertains to guilt or innocence" (MS, AA 06: 401.11-13). Along the same lines, in the *Religion within the boundaries of mere reason* Kant introduces an example in which he asks whether the agent "erred or did wrong with consciousness," thereby indicating that such error is possible, and only a few pages later suggests that acting in accordance with conscience is all that "can be demanded of a person" (RGV, AA 06: 186.27-28 and then 189.15). In both of these passages, Kant claims that an agent is not morally at fault if she does something that is not objectively good despite the fact that she is aiming at the good.

Kant was not, as far as I know, aware of the second possibility--of the possibility of an agent who does something objectively good despite the fact that she is *not* aiming at the good. Or at any rate, I have been unable to find him grappling with this issue explicitly in his texts. Nonetheless, if we trust the student notes from Kant's lectures, then we may infer what Kant would have said about this second possibility.

In the Collins lecture notes on moral philosophy, there is a thought experiment in which an agent is in a place where, if she does not act against her conscience (e.g., by falling down in worship before a picture of an idol she does not recognize), she will cause offense. The lesson

that the students are supposed to learn is that even in such a situation an agent should not act against her conscience, "for this [*viz.*, conscience] must be holy to me" (V-Mo/Collins, AA 27: 335.24). In this context, the claim that an agent's conscience should be "holy" to her means that an agent's conscience should be taken as authoritative. From this it may be inferred that Kant would say that an agent who acts against her conscience, performing a morally right action *despite* her best efforts (despite not aiming at the good), is acting wrongly.

There is something paradoxical about this: if the agent has *failed* to perform an impermissible action, how can it make sense to say that she is acting impermissibly? The answer to this question is that the act is distinct from the intention to do wrong. In conventional cases of intentional wrongdoing, these two go together: a wrong act is performed *because* the agent has the intention to do wrong. But in the nonstandard cases we are imagining right now, the intention and the action come apart, and it is precisely by means of such cases that we can distinguish between our evaluation of an agent's deeper character and our evaluation of the way in which that character is displayed to the world.

It might be objected that there is no way one possibly can know an agent's deeper character, especially if it is not displayed to the world. But, this objection is misguided. In real life, no doubt any evaluation of an agent's deeper character would have to be grounded at least in part on inductive generalizations about the way in which this character manifests on multiple occasions, and no doubt any such generalizations would be uncertain.[12] But that is not a problem for present purposes: this is a thought experiment, one in which the setup (part of which includes the make-up of the deeper character) is stipulated.

Alternatively, it might be objected that this way of explaining the second kind of puzzle case vitiates the attempted response to the first kind of puzzle case. That is, if the evaluation of the agent's intention is kept separate from the evaluation of the agent's action in the second kind of puzzle case, then presumably they should be kept separate in the first kind of puzzle case too. Thus, whereas an agent in the second kind of puzzle case has bad intentions that manifest in a good action, the agent in the first kind of puzzle case has good intentions that manifest in a bad action. In other words, retaining impermissibility in the second kind of puzzle case allows it to creep back in to the first kind of puzzle case.

---

[12] For Kantian literature on motivational opacity, see (Ware, 2009), (Hakim, 2017), or (Berg, 2020).

This objection, however, rests on a misunderstanding. In particular, it runs together two different ways of talking about an agent's action. As may be seen from the foregoing, Kant distinguishes between objective and subjective senses of rightness. Objectively, an agent's duties correspond to what a perfectly rational agent with only good principles would do in her situation. Subjectively, an agent's duties correspond to her best (fallible) judgment about what her duties are. The key here is that, according to Kant, blame and imputability track subjective rightness rather than objective rightness. But that does not change the fact that, objectively speaking, an agent who acts subjectively rightly might be acting (objectively) wrongly.

I would like to say three last things about this.

First, Kant's theory of conscience is not the only part of his practical philosophy that gives support for the idea that, on his view, blame and imputability should track an agent's best judgment about the deontic status of a given action. In addition, one might defend this idea by appeal to Kant's doctrine of autonomy. That is, it might be argued that, if an agent cannot trust her best judgment to be action guiding, then it is hard to see how she can be credited with being an autonomous agent. In other words, in a situation in which an agent's best judgment is not action guiding, the grounds for her action cannot be construed as being in any way internal to her person. This leaves us with an externalist conception of autonomy, which runs the risk of being a contradiction in terms. Given how central autonomy is to Kant's ethics, this gives further grounds for taking blame and imputability to track an agent's best judgment.

Second, it may be seen that the two puzzle cases I have described above actually splinter into six. To see how, note, first, that an agent's judgment might be based on some background values. For example, if an agent is trying to determine whether to lie to a murderer at the door, she might base her judgment on beliefs about the value of human life as compared to the value of telling the truth. But these background values, if based on previously mistaken judgments, can be objectively misguided. Similarly, note, now (second), that an agent's judgment can be mistaken independently of her background values. In the same way that an agent can make a simple arithmetic mistake when calculating a sum or make an invalid inference when solving a logic problem, an agent might fail to judge correctly whether an action is in accordance with her principles.

Putting the possibility of mistaken judgment together with the possibility of mistaken background values we have now six puzzle subcases: (i) an agent with good background values

has faulty judgment and acts in accordance with judgment; (ii) an agent with good background values has faulty judgment and acts contrary to judgment; (iii) an agent with bad background values has good judgment and acts in accordance with judgment; (iv) an agent with bad background values has good judgment and acts contrary to judgment; (v) an agent with bad background judgment has faulty judgment and acts contrary to judgment; and (vi) an agent with bad background values has faulty judgment and acts in accordance with judgment. In cases ii, iv, and vi, the errors interact in such a way that the agent acts objectively rightly; in cases i, iii, and v, the errors are such that the agent acts objectively wrongly.

Two questions immediately arise now. One is: in setting out his theory of conscience (the theory I have drawn upon to show that Kant disavows both the non-accidental rightness condition and GOG), did Kant have fallibility of judgment or fallibility of background values in mind? The other question is: would Kant have to say the same thing about both kinds of fallibility?

Obviously I cannot answer these questions decisively here. But I conjecture that Kant had fallible judgment in mind, not fallible values. And I do not think Kant would have to say the same thing about both kinds of error. For example, concerning bad background values, Kant could say that adopting bad background values corrupts agency in the sense that it undermines an agent's very ability to act autonomously. Bad values would result in figurative holes in personhood. If enough of these holes accumulate, someone could cease to be an agent altogether.

Note that I do not say that Kant (or a Kantian) would *have* to adopt this idea or that he (or anyone) *should* do so. I say merely that Kant (or a Kantian) *could* adopt this idea and, thus, Kant (or a Kantian) would *not* have to extend Kant's remarks about bad judgment (assuming for the sake of argument that my conjecture is correct) to bad background values.

The third and final thing I would like to say has to do with a distinction made in the Collins lecture notes on moral philosophy:

> In regard to his natural obligation / can none be in error; for the natural moral laws could be to none unfamiliar, in that they lie in the reason of everyone; consequently none is innocent there in such error, only in relation to a positive law are errors inculpable, for one can act on the power of a conscientious error as innocent. (V-Mo/Collins, AA 27: 355.5-11)

I am not sure whether the claims in this passage should be taken as representative of Kant's views much less whether they should be endorsed as independently plausible (for the latter, a thorough discussion of how the natural/positive distinction is being used here would be necessary). But I do think that the idea of a distinction between moral views about which there can be culpable error and moral views about which there cannot is *prima facie* plausible and, perhaps, a good way to steer toward a more moderate version of the non-accidental rightness condition. And perhaps something similar can be said about Kant's discussion of radical evil in the *Religion within the boundaries of mere reason* regarding GOG.

This ties back to the discussion of Hardwig from above. Recall that, according to Hardwig, Kant's principle of universalizability itself must be universal. Well, using the natural/positive distinction from the Collins lecture notes, one might argue that inculpable error about the principle of universalizability itself is impossible.

However, some care must be taken in how this position is articulated. First, it must be noted that this does not rule out the possibility of error *sans phrase*. It is merely to say that any such error is culpable. Second, we need to distinguish between know-how and know-that: Kant's Categorical Imperative describes a way of thinking and reasoning about moral principles, and it seems like an agent might know how to reason in accordance with the Categorical Imperative without being able to articulate it and, in fact, while thinking (because of a misunderstanding) that she is not following it.

If Kant's principle of universalizability is going to be understood along the lines of a natural obligation (as Hardwig thinks it must) then I suggest that this be taken along the lines of know-how rather than know-that. On this interpretation, the Categorical Imperative describes how an agent's faculty of reason should function at the most basic level when engaged in practical deliberation The idea is that, if an agent does not engage in reasoning of this kind, it cannot be on account of an inculpable error (in the way, for example, that an inculpable error might lead an agent to break the speed limit (this notwithstanding the fact that traffic violations of this kind are often taken to be strict liability)). Rather, such failure is either culpable error or inculpable and nonmoral error. In the former case the error is moral and the agent is held responsible; in the

latter case the error falls outside the bounds of agency and the individual's actions are evaluated in the way that the actions of wild animals are.[13]

## Conclusion

In this paper I scrutinized Herman's non-accidental rightness condition and Korsgaard's GOG. I argued that these two theses are irreconcilable and, further, that there are good textual grounds for thinking that Kant would disavow both. I argued that the way toward a plausible Kantian ethics is to follow Kant in this disavowal, but I noted that this opens up the possibility of various puzzles having to do with agents who perform good actions with malice aforethought and agents who perform bad actions despite having the best intentions. I sketched some ways of dealing with these cases on the basis of remarks in Kant. But I do not take myself to have given a decisive treatment of them: much work remains to be done.

[13] If this is correct, then a case could be made for saying that the Nazi appeal to the *Führerprinzip* either is disingenuous and therefore culpable or sincere and an admission that they have ceded their agency. In other words, this attempt to abnegate responsibility comes at the price of abnegating agency.

**Bibliography**

Allison, H. (1990), *Kant's Theory of Freedom*, Cambridge University Press.

Baron, M. (1995), *Kantian Ethics Almost Without Apology*, Cornell University Press.

Berg, A. (2020), "Kant on Moral Self-Opacity," *European Journal of Philosophy*, Vol. 28, Issue 3, pp. 567-585.

Engstrom, S. (1992), "The Concept of the Highest Good in Kant's Moral Theory," *Philosophy and Phenomenological Research*, Vol. 52, pp. 747-780.

Hakim, D. (2017), "Kant on Moral Illusion and Appraisal of Others," *Kantian Review*, Vol. 22, No. 3, pp. 421-440.

Hardwig, J. (1983), "Action from Duty But Not in Accord with Duty," *Ethics*, Vol. 93, No. 2, pp. 283-290.

Herman, B. (1990), *Morality as Rationality*, Routledge.

Herman, B. (1993), *The Practice of Moral Judgment*, Harvard University Press.

Kahn, S. (2014), "The Interconnection of Willing and Believing in Kant's and Kantian Ethics," *International Philosophical quarterly*, Vol. 54, No. 2, Issue 2, pp. 143-157.

Korsgaard, C. (1996), *Creating the Kingdom of Ends*, Cambridge University Press.

Korsgaard, C. (2004), "Fellow creatures," *Tanner Lectures on Human Values,* 24, pp. 77-110.

Korsgaard, C. (2009), *Self-Constitution*, Oxford University Press.

Raz, J. (2008), “On the Guise of the Good,” *University of Oxford Legal Research Paper Series,* Paper No 43/2008. Available at SSRN: https://ssrn.com/abstract=1099838 or http://dx.doi.org/10.2139/ssrn.1099838

Sverdlik, S. (2001), “Kant, Nonaccidentalness and the Availability of Moral Worth,” *The Journal of Ethics,* Vol. 5, No. 4, pp. 293-313.

Ware, O. (2009), "The Duty of Self-Knowledge," *Philosophy and Phenomenological Research*, Vol. 79, No. 3, pp. 671-698.

Wood, A. (1999), *Kant's Ethical Thought*, Cambridge University Press.

Wood, A. (2008), *Kantian Ethics*, Cambridge University Press.