Agency and Responsibility

According to Christine Korsgaard, Kantian hypothetical and categorical imperative principles are constitutive principles of agency. By acting in a way that is guided by these imperatives, an individual makes herself into an agent. There is hence, on her theory, an inextricable link between the nature of agency and the practical issue of why we should be rational and moral. The benefits of such an account would be great: in Korsgaard's view, an account that bases morality on the nature of agency is the basis for a refutation of any kind of moral skepticism, providing an indubitable and objective foundation for morality. This may seem too good to be true, and it is. Korsgaard could only succeed at offering a foundation for morality at a great cost. The cost is that Korsgaard gives too restrictive an account of agency. Korsgaard does not present a coherent account of irrational or immoral agency, and the inability to offer an account of such agency implies an inability to offer a proper account of responsibility. Korsgaard's view shares a fundamental flaw with Immanuel Kant's account of morality in the *Groundwork of the Metaphysics of Morals*: Korsgaard cannot give a full, adequate account of individual responsibility. In light of the failure of Kant's and Korsgaard's accounts, Kantians need to provide a better, more comprehensive characterization of agency. Presenting a proper account of agency will require a rejection of a central tenet of traditional Kantian metaethics, but the rejection of this central tenet does not require a full rejection of Kantianism.

1. Korsgaard on Agency

In order to be clear on how Korsgaard's account is supposed to account for morality, it would help to be clear on what she means by agency. Unfortunately, Korsgaard does not offer a single, precise definition of this key notion.

Korsgaard has defined an agent as an active person: "A person is both active and passive, both an agent and a subject of experiences".[i] This definition could be called the active person definition of agency. Korsgaard goes on to say, in the same essay just cited, that "we may regard ourselves as agents, as the thinkers of our thoughts and the originators of our actions."[ii] This latter characterization goes beyond the active person definition. If being an agent requires being the originators of our own actions, then it is not sufficient to just act: we must act in a way that is self-originating. This could be termed the self-originating active person definition.

To make full sense of this, it would help to get clear on the notion of self-origination. Korsgaard's mention of self-origination suggests a kind of libertarian account of freedom of the will, the sort of account offered by Roderick Chisholm[iii], on which the agent or the self is an Aristotelian unmoved mover. Perhaps Korsgaard's notion of agency, of self-originating action, might best be understood in terms of agent-causation: we are agents only insofar as we are active selves, specifically active selves that are not determined to act in the way we act by prior events.

However, this metaphysical, theoretical notion of agent-causation is not what Korsgaard has in mind. Korsgaard invokes a Kantian distinction between theoretical and practical points of view, to suggest that what is relevant to the nature of agency is not the

theoretical, metaphysical fact of whether or not our choices are uncaused causes, but rather the purported fact that, in order to make a choice at all, one must regard herself as an unmoved mover. Korsgaard writes: "We must view ourselves [as agents, as free, as responsible] when we occupy the standpoint of practical reason—that is, when we are deciding what to do. This follows from the fact that we must regard ourselves as the causes—the first causes—of the things that we will".[iv] To properly characterize the self-originating active person conception of agency offered by Korsgaard, one should do so as follows: to be an agent is to be an individual who must see herself as an uncaused cause from the practical point of view.

This characterization of agency, drawing as it does on the distinction between practical and theoretical standpoints, is far from clear. It invokes the notion of a cause, which is a theoretical, metaphysical notion. How exactly one should understand this notion of causation, of first causation, in a non-theoretical fashion is never clearly spelled out by Korsgaard.

Perhaps, though, this characterization suggests that what is required is a positive account of self-origination, of how one supposedly determines action without prior determination from the practical point of view. In *The Sources of Normativity*, Korsgaard reiterates the self-originating active person conception of agency: "Minimally, then, I am not the mere location of a causally effective desire but rather am the *agent* who acts *on* the desires." How is this achieved? Korsgaard claims that it is achieved through consistency: "It is because of this that if I endorse acting a certain way now, I must at the same time endorse acting the same way on every relevantly similar occasion.… For if all my decisions were particular and anomalous, there would be no identifiable difference

between *my acting* and *an assortment of first-order impulses being causally effective in or through my body*."[v] In the same work, she claims that consistency is required for an act to be an act. "This claim to generality, to universality, is essential to an act's being an act of the will."[vi] Hence we have a third conception of agency at work in Korsgaard's discussion, the consistency conception of agency.

The consistency conception of agency draws on Korsgaard's reading of the categorical imperative, a reading that gives emphasis to the formula of the universal law. "The categorical imperative," writes Korsgaard in *The Sources of Normativity*, "as represented by the formula of the Universal Law, tells us only to act on a maxim which we could will to be a law. And *this,* according to Kant, *is* the law of a free will."[vii] This is a fairly accurate representation of Kant's formulation of the categorical imperative, "*act only in accord with that maxim through which you can at the same time will that it become a universal law"*.[viii] The way in which Korsgaard goes on further to spell out the categorical imperative, as she is interpreting it, is a serious departure from Kant. Korsgaard claims that the only "constraint on our choice" presented by the categorical imperative is that our choice "has the form of the law. And nothing determines what the law must be. *All that it has to be is a law*."[ix] This cannot be what Kant has in mind.

For Kant, for a maxim to pass the categorical imperative test, it must both be capable of being a universal law, and it must be a universal law one could will to be a universal law. As Kant notes, in his discussion of indifference to a person in need, it is possible to make it a universal law to ignore the needs of others, "if such a way of thinking were to become a universal law the human race could admittedly well subsist…But although it is possible that a universal law of nature could very well subsist

in accord with such a maxim, it is still impossible to **will** that such a principle hold everywhere as a law of nature."[x] It is this latter test, what one is able to will to be a universal law, that in Kant's view would make it wrong to ignore someone who is in need. One would not want to live in a world where others would ignore her in similar circumstances; hence it is for Kant impossible to genuinely will ignoring others in need into universal law.

Setting this issue in Kant exegesis aside, the account Korsgaard offers of the categorical imperative helps to clarify the consistency conception of agency. It also helps to explain Korsgaard's view that the categorical imperative is a constitutive principle of agency. In order to act, to be an agent, one must act in a way that is guided by universal laws, consistent principles of actions. Insofar as one does not act in such a consistent fashion, she is not an agent.

There are two more points worth noting about Korsgaard's account of agency. Korsgaard alternates between making claims about the constitution of agency, and the constitution of the self. It seems clear, then, that Korsgaard's account of agency is intended as an account of the nature not only of agency but also of the self. Also, Korsgaard thinks that being an agent, or being a self, is in some crucial way necessary for human beings. She claims that it is impossible not to be an agent, not to be a self. This is very important for Korsgaard's justification of morality, and her response to moral skepticism. For if morality and rationality are based on the nature of the agent or the self, and one cannot help but be an agent, then morality and rationality have an indubitable foundation in human nature.

Korsgaard, as noted above, claimed that "if all my decisions were particular and anomalous, there would be no identifiable difference between *my acting* and *an assortment of first-order impulses being causally effective in or through my body*."[xi] To claim this is to assume that there cannot be a way to be both self-controlled and acting, but not acting in a way that would require consistency over time. This is simply not true, it leaves out an alternative: self-controlled action for particular reasons in particular circumstances. Imagine that Mary is in a situation where, by cheating on her tax return, she would be able to pay for an expensive medical treatment for her child. She wants to do so on this occasion and this occasion only: fearing the possibility of being caught for being a tax cheat, she decides to cheat now, but not to cheat on her taxes whenever she is in difficult circumstances. On Korsgaard's account, we would have to consider Mary a sort of wanton, lacking self-control, ruled by the particular impulse on which she acts. This seems a misreading of the facts: why can it not be the case that Mary is fully in control of herself when she makes this decision?

Korsgaard might claim that the example of Mary presents an impossible circumstance. Mary has made the decision to cheat on her taxes just this one time, but does not will the universal principle that one should always cheat on one's taxes when in need. Insofar as Mary is not willing a universal principle, she is not, on Korsgaard's account, acting at all. Mary wants to cheat on her taxes just this one time, while not wanting to cheat again the next time she is in this predicament. There is no consistent universal principle behind her action, so Korsgaard would suggest that Mary's action is simply impossible. Insofar as she does not will a universal principle, Mary is not, by Korsgaard's lights, an agent. It is far from obvious that Korsgaard's denial of Mary's

agency would be correct. Only on a narrow, technical definition of the notion of agency would it be the case that such an action is impossible. If it were the case that one had to be able to will a consistent universal principle each time one acts, then a vast number of human actions would be ruled out as possibilities by this definition. Observation of human behavior reveals that such actions are not just possible but actual and common. One troubling aspect of such a narrow definition of agency is that it would rule out the possibility of irrational or immoral acts, insofar as such agency is not based on universal principles of rationality and morality.

This raises a deeper worry for Korsgaard's account. As stated, the only way to be an agent is to act on rational principles, including the categorical imperative. If this is so, it is difficult to see how immorality or irrationality is even possible.

2. Whither Irrationality and Immorality?

The following is just an obvious, commonsense observation: there are irrational and immoral people in the world. Whatever one's characterization of immorality and irrationality might be, one can regard the person who fails throughout her life to do what is in her own self-interest irrational, and one can consider the parent who abuses her children to be an immoral person. There is no need for any further argument for this claim, for it is an easily observed fact that stands in no need of a defense.

Given that immorality and irrationality exist, it is a constraint on any adequate account of moral agency that it allows for the possibility of immorality and irrationality. This constraint will be termed the failure constraint. It must be the case that we can allow

not only for the existence of successful attempts at rationality and morality; we have to allow for the possibility of failure as well.

Korsgaard's account, on the face of it, does not meet the failure constraint. The reasons for this are as follows: principles of rationality and morality, such as the categorical imperative, are in her views constitutive principles of agency or the self. In order for human being to be an agent, she must act in accord with the categorical imperative principle. However, there are clearly individuals in the world who fail to act in accord with the categorical imperative.

This circumstance presents a dilemma for Korsgaard: either those individuals who fail to act in accord with the categorical imperative are immoral agents, or they are not agents at all. If Korsgaard takes the first horn of the dilemma, she would have to reject her own account to do so. For insofar as an agent is an agent, she must constitute herself by acting in a way that is guided by the categorical imperative. However, an immoral individual, on Korsgaard's account, would be an individual who is not guided by the categorical imperative. On her account, immoral agents would have to, insofar as they are agents, act on the categorical imperative, and insofar as they are immoral, not act on the categorical imperative. This is clearly incoherent.

The second horn of the dilemma presents yet another difficulty. If immoral individuals are not agents at all, then there cannot be immorality. Korsgaard indicates that she may take this horn of the dilemma in her most recent book, *Self-Constitution,* writing that that "the laws of practical reason govern our actions because if we don't follow them, we aren't acting."[xii] Insofar as an individual is not governed by these laws of practical reasoning, she is not an agent. It would only make sense to attribute the property of being

immoral to an agent. However, if being immoral, failing to act on the categorical imperative, is tantamount to not being an agent, then there cannot be immorality for there cannot be immoral agents.

One might consider Mary, in the example presented above, to be performing an immoral act. In her circumstances, she wants to make an exception for herself to the tax laws. This could be seen as a violation of the categorical imperative, for Mary might be willing to make an exception for herself without willing it to be a universal law that others be allowed to cheat on their taxes in difficult circumstances. Based on Korsgaard's definition of agency, one might deny that Mary is an agent at all, given that there is no universal principle behind her action. It would follow from this that it would not be possible to attribute the immorality of the action of cheating on her taxes to her, insofar as she is not an agent. Hence Korsgaard's narrow definition of agency precludes the attribution of immorality.

In her recent essay "Self-Constitution of the Ethics of Plato and Kant," Korsgaard raises this worry about her account. As Korsgaard notes, her neo-Kantian account shares this problem with Kant's own account in the *Groundwork of the Metaphysics of Morals*. "For a well-known problem in the *Groundwork* is that Kant appears to say that only autonomous action, that is action governed by the categorical imperative, is really free action…"[xiii] It is not entirely clear why Korsgaard claims that Kant appears to say this: the account of free action offered by Kant in the *Groundwork* clearly has this implication. For Kant, causation requires a law, and there can only be two kinds of laws: natural laws and the moral law. The moral law, Kant claims, would be the law that governs the free, noumenal will. This account has just the same problematic implication, failure to meet

the failure constraint, as Korsgaard's account: for if individuals are only responsible for free actions, and the only way for an action to be free is if it under the moral law, then there cannot be genuine immorality at all. On Kant's account, the only possible actions a person can do freely, the only actions for which a person can be responsible, are moral actions.

Korsgaard clearly sees that this is an implication of her account and Kant's account in the *Groundwork*. As she writes, "So it looks at first as if *nothing exactly counts as a bad action*."[xiv] Surprisingly, Korsgaard claims that this problem is not a worry at all, it is "our main reason for embracing" this account of the self.[xv] How could this be so? Korsgaard addresses the worry by weakening her account of what is required for agency. The immoral or irrational person is someone who is making a failed attempt to act on principles of rationality such as the categorical imperative: Korsgaard claims that "even the most venal and shoddy person must try to perform a good action, for the simple reason that there is no other way to try to perform an action."[xvi] It is important to note that this revision contradicts her claim that the categorical imperative is a constitutive principle of agency. On her revised view, it is trying to act on the categorical imperative that is constitutive of agency, not acting on the categorical imperative. In *Self-Constitution,* Korsgaard states this view by claiming that the categorical imperative is among the "constitutive principles of action, principles to which we necessarily are trying to conform insofar as we are acting at all."[xvii] On this account, one cannot act without trying to act in conformity with moral principles. What occurs in the case of a bad action, what Korsgaard calls a defective action, is that the attempt to act in this fashion is a failed one. "The kind of practical deliberation that results in bad action is not a different activity

from the kind of practical deliberation that results in good action. *It is the same action, badly done*."[xviii] Morally wrong action is a failed attempt to act on the same principles that lead to morally right action.

Even this revised account does not meet the failure constraint. For it is reasonable to think that people should be given just as much credit for trying to do what is right as for succeeding in doing what is right. The great baseball player Roberto Clemente died in a plane crash while trying to deliver aid packages to the victims of an earthquake in Nicaragua. He was clearly trying to do something good, and is perhaps given even more credit for the tragic circumstances of his death. Given that Clemente tried, but failed, to do something good, he deserves moral praise. Korsgaard, qua Kantian, is committed to the view that one ought to be judged not based on the consequences of actions but on the intention behind the action, hence Korsgaard would think Clemente praiseworthy. As a result, Korsgaard is hoist by her own Kantian petard. If the intent to do what is right is the standard for judging an individual, then we ought to consider all persons who try to conform to the categorical imperative as good persons. The class of good persons would not only include moral saints like Clemente, but every person, including a shoddy or venal person, insofar as every agent is making an attempt to conform to the categorical imperative.

This is not what we would ordinarily say about a venal or shoddy person. However, if, as Korsgaard suggests, venal and shoddy people are always trying to do what is right, then it is not at all clear we should consider them venal or shoddy at all, rather than just unfortunate. If the aim of all people is to act in accord with the categorical imperative, and the categorical imperative is the fundamental principle of morality, then

everyone is aiming to be moral. As a result, on an intention-based morality, everyone is morally good. This revised version of the account, like the original account, violates the failure constraint: it cannot allow for the possibility of immorality.

Furthermore, the suggestion that all agents must, whatever they do, be trying to act in a way that is guided by the categorical imperative, is an ad hoc claim that is given no support whatsoever by Korsgaard. Why should we believe that venal and shoddy people are really trying to act in accord with the categorical imperative? What is the evidence, outside of Korsgaard's theory, that this is true? When a greedy businessman embezzles a pension fund for his own personal gain, why should one think that the businessman has tried, but failed, to be moral? This naïve, early Socratic account of moral psychology, on which all persons must be trying to be rational and good, just does not do justice to the facts.

3. Explaining Agency

In order for an account of moral agency to be adequate, it must allow for the possibility of moral and immoral agency, rational and irrational agency. Failure to allow for these possibilities, as has been illustrated in the discussion above of Kant's view in the *Groundwork* and Korsgaard's theory of agency, is failure to offer an account of agency that can properly characterize the irrational or immoral agent as responsible for irrationality or immorality. Further, as has been illustrated with the example of Mary, who wants to cheat on her taxes once and only once, the consistency conception of agency offered by Korsgaard is too narrow for another reason. It treats anyone who acts for a particular reason on a particular occasion as a non-agent, a non-person, lacking self-

control. The quality of self-control, however, is not plausibly identical to the quality of acting in a fully consistent fashion. Persons can be self-controlled without ideal consistency.

A better account of agency would allow that inconsistent, irrational, and immoral agency is still agency. Better yet, it would be an account that fits a generally naturalistic picture of the world, and does not require the postulation of any kind of ontological oddity such as the Kantian noumenal will. Luckily, progress has been made on matters of this kind in the compatibilist literature on freedom of the will. It is a lacuna in Korsgaard's writing that she does not relate her accounts of agency to the current literature on freedom of the will. She writes that "Agency is almost as mysterious as freedom of the will" in her most recent essay collection.[xix] The great advantage of compatibilist accounts is that they demystify freedom and agency. Further, and most to the point for the purpose of this essay, compatibilists offer an account of freedom that does not violate the failure constraint.

Harry Frankfurt, in his classic essay "Freedom of the Will and the Concept of a Person," offers such an account of agency.[xx] Frankfurt's account can allow for agency that is consistent or inconsistent, rational or irrational, moral or immoral. Frankfurt characterizes freedom in terms of a particular kind of desire, a desire that Frankfurt considers essential to being a person. To act in a way that is free is to act on a desire that one desires to desire. Having free will is a matter of having second-order desires. For instance, if Fritz wants to write a paper, and Fritz wants to have this want to write his paper, then he is writing this paper of his own free will. On the other hand, if Mitch is

compulsive hand-washer, and while Mitch wants to wash his hands, he does not want to have this want, then when Mitch washes his hands he is not acting of his own free will.

Unlike Korsgaard's account, Frankfurt can allow for desires that are particular but self-controlled. If Mary, in the example presented above, both wants to cheat on her taxes just this one time, and wants to want to cheat on her taxes, then she can be self-controlled, freely acting, even if her action is not based on a desire to act in this fashion on every similar occasion. Frankfurt's account also allows for blameworthy instances of irrationality and immorality. If a parent wants her child to be able to go to college, but she does not save any of her money to provide her child with this education, one can say that this parent is irrational. If she desires to spend her money, not save it, and desires to have this desire, then one can rightly say that she has made a free, irrational choice. If the business executive who embezzles the pension fund is not suffering from an uncontrolled compulsion, but instead both desires to embezzle the fund for his own well-being and desires to have this desire, Frankfurt's account, unlike Korsgaard's, allows for the possibility of claiming that the businessman makes a choice that is both free and immoral.

If Kantians accept an account of agency of the sort presented by Frankfurt, based in second-order desires, it seems that they must do so at a significant cost. For it is difficult to see how such an account of agency could serve as a foundation for morality. If it possible for an agent to be rational as well as irrational, moral as well as immoral, then agency cannot serve as any sort of guarantee of morality or rationality. The hope of Kant and Korsgaard that anyone who is free, an agent, a self must be moral is an empty one, one that requires a restricted view of human life.

There are good, more general metaethical reasons to think that the grounding of morality in agency could never succeed. For it is a well-known fact that an ought cannot be derived from an is. Any sort of metaphysical account of the nature of the world, or the nature of persons, does not suffice to answer the questions of what one ought to do and why. Interestingly, I think that Korsgaard's attempts to base morality in agency is based on a similar mistake that is made by the sort of moral realist accounts she rightly opposes: just as the moral realist makes an failed attempt to base morality on the metaphysical structure of reality, seeking a ground for what one ought to do in the nature of what exists, Korsgaard attempts to ground morality in the metaphysical structure of the self. The accounts differ, but the mistake is the same.

For this reason, the search for a metaphysical foundation of morality is a nonstarter. Is this a worry? Should Kantians need to despair for the lack of a foundation of morality in the nature of the self? This is no cause for Kantian moralists to worry. The search for a justificatory foundation of morality is a search quite similar to the epistemic search for a justificatory foundation of knowledge. While there have been many philosophers who have realized that foundationalism about knowledge is a failure, there seems to be a failure to recognize that the search for a metaphysical foundation for morality is flawed for just the same reason: it is an attempt to find a justification, an ought, based on something that is not a justification, something that is an is but not an ought. The Kantian categorical imperative principle, that one should not act on maxims that would lead her to treat persons as mere means, but always at the same time treat persons as ends in themselves, does not need a metaphysical defense. The only way to

justify such a principle is through moral thinking, through considering whether this principle fits well with other considered, moral views.[xxi]

Notes

---

[i] Christine M. Korsgaard, "Personal Identity and the Unity of Agency: A Kantian Response to Parfit" in *Creating the Kingdom of Ends* (Cambridge, England: Cambridge University Press, 1996), p. 363.

[ii] Ibid., p. 377.

[iii] See Roderick M. Chisholm, "Human Freedom and the Self," *The Lindley Lecture*, (1964).

[iv] "Personal Identity and the Unity of Agency: A Kantian Response to Parfit," p. 378.

[v] Christine M. Korsgaard, *The Sources of Normativity* (Cambridge, England: Cambridge University Press, 1996), p. 228.

[vi] Ibid., p. 232.

[vii] Ibid., p. 98.

[viii] Immanuel Kant, *Groundwork of The Metaphysics of Morals* in *Practical Philosophy*, ed. and trans. Mary Gregor, (Cambridge, England: Cambridge University Press, 1996), p. 73.

[ix] The *Sources of Normativity*, p. 98.

[x] *Groundwork of The Metaphysics of Morals*, p. 75.

[xi] *The Sources of Normativity*, p. 228.

[xii] Christine M. Korsgaard *Self-Constitution* (Oxford: Oxford University Press, 2009), p. 32.

[xiii] Christine M. Korsgaard, "Self-Constitution in the Ethics of Plato and Kant" in *The Constitution of Agency* (Oxford: Oxford University Press, 2008), p. 110.

[xiv] Ibid., p. 111.

[xv] Ibid., p. 112.

[xvi] Ibid., p 113.

[xvii] Christine M. Korsgaard, *Self-Constitution*, p. xii.

[xviii] Ibid., p. 132.

[xix] Christine M. Korsgaard, "Introduction" in *The Constitution of Agency* (Oxford: Oxford University Press, 2008), p. 10.

[xx] See Harry Frankfurt, "Freedom of the Will and the Concept of a Person," *The Journal of Philosophy*, Vol. 68, No. 1, (1971).

[xxi] I would like to thank the University Research Committee of Oakland University for a Faculty Research Fellowship that supported this project. I am grateful to Radu Neculau, Christopher Tindale, Phillip Rose, Jeff Noonan, and an audience at the University of Windsor for their comments on an earlier version of this paper. I am also grateful to Rosemary Twomey, David Pereplyotchik, David Enoch, and Keota Fields for discussions of Christine Korsgaard's philosophy that led to my writing this paper. I am especially grateful to two anonymous reviews from *The Journal of Value Inquiry* for their helpful critiques of this paper.