Reassessing the foundation of Korsgaard's approach to ethics

Samuel Kahn
15 May 2017

1. Introduction

In a series of well known publications, Christine Korsgaard argues for the claim that an agent acts morally just in case s/he acts autonomously. To act autonomously is to act in accordance with self-given principles. For example, Korsgaard argues that the character Harriet Smith (from Jane Austen's *Emma*) does not act autonomously because, instead of following her own principles, Harriet lets herself be guided by Emma Woodhouse.[1]

Two of Korsgaard's signature arguments for the connection between morality and autonomy are the "argument from spontaneity" and the "regress argument." Both of these arguments have a similar structure: starting from something that all agents do unavoidably, Korsgaard tries to show that autonomy is an internal norm of action. In the argument from spontaneity, Korsgaard maintains that all agents act under the idea of freedom. In the regress argument, Korsgaard maintains that all agents represent their ends as objectively good.

An internal norm for an activity is one that is constitutive of that activity.[2] For example, Korsgaard argues that building a structure that can provide shelter is constitutive of building a house.[3] If this is correct, then it is not possible to try to build a house without trying to build a structure that can provide shelter. In the same way, Korsgaard maintains that because the argument from spontaneity and the regress argument show that autonomy is constitutive of action, to act is to try to act autonomously.

In this paper, I argue that neither the argument from spontaneity nor the regress argument is able to show that an agent would be acting wrongly even if s/he acts in a paradigmatically heteronomous fashion (as Harriet does). The paper is divided into 5 sections. In section 2, I explore the implications of what Korsgaard says about the morality of lying for these two arguments. In sections 3 and 4, I examine, first, the argument from spontaneity and, second, the regress argument in more detail. In section 5, I summarize my results and gesture toward an argument from Korsgaard's *Self-Constitution: Agency, Identity, and Integrity* (SC).

I should note that Korsgaard claims to find both the argument from spontaneity and the regress argument in Kant. I shall not be weighing in on the textual correctness of these arguments. Similarly, I am not arguing that autonomy approaches in ethics are incorrect in general or that Kant does not have the resources to justify such an approach. My goal is simply to show that two of Korsgaard's main arguments cannot provide the foundation for an autonomy approach that she takes them to provide.

2. Korsgaard on the morality of lying

The argument from spontaneity attempts to connect Kant's Formula of Universal Law (FUL) with autonomy and to show that FUL is an internal norm of action. FUL is the first formulation of the Categorical Imperative (CI) introduced in the *Groundwork for a Metaphysics of Morals*. It runs as follows: "Act only according to that maxim by which you can at the same time will that it should become a universal law" (4: 421).[4] The regress argument, by way of contrast, attempts to connect Kant's Formula of Humanity (FH) with autonomy and to show that FH is an internal norm of action. FH, like FUL, is introduced in the *Groundwork for a Metaphysics of Morals*. It runs as follows: "So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means" (4: 429, emphasis omitted). My goal in this section is to show that Korsgaard's position on the morality of lying requires (on pain of inconsistency) that the argument from spontaneity (and therefore the kind of autonomy shown to be normative by this argument) is weaker than the regress argument (and the kind of autonomy it shows to be normative).

Korsgaard's position on lying emerges from her account of the murderer at the door case, which is taken from Kant's *On a supposed right to lie from philanthropy*. The idea is as follows: suppose someone comes to your door and asks to hide in the basement because s/he is being chased by a murderer. You comply, and shortly thereafter the murderer shows up and asks you whether you are hiding the person. To make the case more vivid, you might imagine Nazis asking for the whereabouts of a Jew who is hiding in the basement. The question is whether you may lie.[5]

Korsgaard's investigation of this question, which is in chapter 5 of *Creating the Kingdom of Ends* (CKE), has three main parts. First, she argues that if the murderer is concealing his/her purpose of murdering, then a maxim to lie to the murderer would pass FUL: "the maxim of lying to a deceiver is universalizable."[6] Second, she argues that regardless of whether the murderer is concealing his/her purpose, it would be impermissible to lie to the murderer by the standard of FH because "[FH] says that coercion and deception are the most fundamental forms of wrongdoing."[7] Third and finally, she argues that when strict compliance with the ideal embodied in FH would make them the tools of evil, agents should turn to FUL for guidance.[8]

Now consider the following three theses: (1) FUL is an internal norm for all actions; (2) FH is an

internal norm for all actions; and (3) when FUL and FH give different results, agents should follow FUL.[9] From an intuitive perspective, it is (I think) permissible to lie to the Nazis in the scenario described. But given (1) and (2), it is *prima facie* unclear (from a doctrinal perspective) why FUL should be given more weight than FH. If Korsgaard could show that any duties entailed by FUL are entailed by FH but not the other way around, that might suffice to fix the problem here, for that would suggest that FUL plays a more foundational role in the constitution of action than FH. And that is exactly what she tries to do:

> Suppose that your maxim is in violation of the Formula of Universal Law… What this means is that you are treating the reason *you* have for the action as if it were stronger, had more justifying force, than anyone else's exactly similar reason. You are then acting as if the fact that it was in particular *your* reason, and not just the reason of a human being, gave it special weight and force. This is an obvious violation of the idea that it is your humanity — your power of rational choice — which is the condition of all value and which therefore gives your needs and desires the justifying force of *reasons*. Thus, any violation of the Formula of Universal Law is also a violation of the Formula of Humanity. This argument, of course, only goes in one direction… The Formula of Humanity is stricter than the Formula of Universal Law…[10]

In this passage, Korsgaard argues that any violation of FUL is a violation of FH, but not vice versa: FH is stricter than FUL. But this entails that the premises of the regress argument are stronger than those of the argument from spontaneity: if FH entails all the results of FUL but not the other way around, then the premises of any argument that can be used to prove FH also entail FUL (but not the other way around). As will be seen below (see especially note 19), Korsgaard seems to adopt a similar position in her later work.

3. The argument from spontaneity

As noted above, the argument from spontaneity is supposed to show that Kant's so-called Formula of Universal Law (FUL) is an internal norm of action. My goal in this section is to show that the argument from spontaneity shows at most that agents always do act on maxims they can will as universal laws and, therefore, autonomously (in a weak sense) rather than that they always should act autonomously (in a strong sense).

Korsgaard's first formulation of the argument from spontaneity is in chapter 6 of CKE.[11] The argument begins with the premise that agents must make choices as if they are free.[12] According to Korsgaard, agents must make choices as if they are free even if they do not believe that they are free because even if you believe that you have been preprogrammed to act in a certain way, "[i] n order to *do* anything, you must simply ignore the fact that you are programmed, and decide what to do — just as if you were free."[13]

The second premise of the argument is about what it is to choose freely: to choose freely is to act according to one's own principles (i. e., autonomously).[14] Korsgaard now steps back from the argu-

ment and imagines being asked why a given free will adopts a given principle. For example, one might ask why free agent A adopted principle P1. If A was not acting on a principle in deciding to adopt P1, then P1 was randomly selected — and this is impossible because if A is free, A must act according to A's own principles, and a randomly selected principle is not self-given. But for any principle P2 that A claims to be acting on when choosing P1, one might ask why A adopted P2. If A claims to be acting on some P3 when choosing P2, then one might ask why A adopted P3. A regress looms, and as Korsgaard points out, "to put an end to a regress like this we need a principle about which it is impossible, unnecessary, or incoherent to ask why a free person would have chosen it."[15]

The crucial step in Korsgaard's argument is the claim that FUL plays the role of a regress-ending principle. FUL plays this role because, according to Korsgaard, "[t] his formula merely tells us to choose a law… the free will need do nothing to make the Formula of Universal Law its principle: it is already its principle."[16] Thus, Korsgaard's conclusion from this sub-argument is that to act according to one's own principles is to act according to FUL.

To summarize, Korsgaard's argument is as follows: (1) agents necessarily choose as if they are free; (2) to act freely is to act on self-given principles; (3) to act on self-given principles is to act according to FUL; therefore (4) agents necessarily choose as if according to FUL. Presumably, to choose as if according to FUL is to make FUL an internal norm of action. However, (2) — (3) show that agents act according to FUL if and only if they *are* free, whereas (1) presupposes that agents can act even if they are *not* free. Thus, (1) — (3) entail that whether something counts as an action is independent of whether it is in accordance with FUL, whence it seems to follow (*pace* Korsgaard) that FUL is not an internal norm of action.

It might be objected that if (1) — (3) are true, then Korsgaard nonetheless has shown that agents necessarily try to follow FUL. However, the "as if" in (4) does not connote trying (or the attendant possibility of failure). Korsgaard's argument for (1) reveals her claim to be that for the purposes of deliberation, determinism is moot. Following through to (4), the conclusion seems to be that for the purposes of deliberation, whether a principle can be willed as a universal law is moot. Focusing on the "as if" from (1), Korsgaard seems to face a dilemma: depending on one's interpretation of FUL, either (A) if agents must act as if they are free (for the purposes of deliberation), then agents must will all of their maxims as universal laws (whence it may be inferred that they are able to do so, rendering FUL an empty formalism), or (B) if agents must act as if they are free (for the purposes of deliberation), then whenever an agent acts on a maxim that cannot be willed as a universal law, it is through no fault of his/her own (rendering FUL nonnormative). Taking the focus off of (1) and focusing instead on the equivalences in (2) and (3), another dilemma arises: either (a) agents are free, in which case they will all of their maxims as universal laws (rendering FUL an empty formalism), or (b) agents are not free, in which case they never will any of their maxims as universal laws (rendering FUL nonnormative).

Korsgaard supplements the argument from spontaneity with another argument to show that if agents necessarily act under the idea of freedom, then they really are free.[17] But this does not avoid the

problem because then (3) would entail that agents necessarily act according to FUL: this amounts to choosing (a), the first horn of the second dilemma just described (rendering FUL an empty formalism). Korsgaard ultimately seems to concede something along these lines in *The Sources of Normativity* (SN):

> … the argument I just gave [the argument from spontaneity] doesn't settle the question of the *domain* over which the law of the free will must range. And there are various possibilities here. If the law is the law of acting on the desire of the moment, then the agent will treat each desire as a reason, and her conduct will be that of a wanton. If the law ranges over the agent's whole life, then the agent will be some sort of egoist. It is only if the law ranges over every rational being that the resulting law will be the moral law. Because of this, it has sometimes been claimed that the categorical imperative is an empty formalism. And this has in turn been conflated with another claim, that the moral law is an empty formalism. Now that second claim is false…[18]

In this passage, Korsgaard distinguishes between the categorical imperative, which she identifies as FUL, and the moral law, which she identifies with thinking of oneself as a citizen in a realm of ends.[19] In SN, Korsgaard thinks that the argument from spontaneity "establishes that *the categorical imperative* is the law of a free will. But it does not establish that *the moral law* is the law of a free will."[20] This is important because in the section of text just quoted, Korsgaard claims that the categorical imperative is an empty formalism that everyone, even wantons and egoists, follows.

Passages similar to the one above can be found in Korsgaard's later work.[21] From this it may be seen that (seemingly by Korsgaard's own admission) the argument from spontaneity justifies only the claim that all agents always do act autonomously (in an empty formalistic sense) rather than the claim that all agents always ought to act autonomously (in a strong sense). Evidently the substantive moral duties that Korsgaard derives from FUL in her earlier work are based on a different understanding of this principle.[22] All of this puts a lot of pressure on Korsgaard's regress argument, which is supposed to ground what she is calling the moral law and which I examine in the next section.

4. The regress argument

Korsgaard's first formulation of the regress argument is in chapter 4 of CKE.[23] My goal in this section is to show that the premises of the regress argument do not entail that FH is an internal norm of action.

The first premise of Korsgaard's argument is: (1) agents always take their ends to be objectively good.[24] Korsgaard explains that something is objectively good if and only if (i) it is unconditionally good or (ii) it is conditionally good and its conditions are met.[25] The next few premises contain the regress on conditions: (2) an object of inclination is objectively good only if there is an agent with an inclination that would be satisfied by that object;[26] (3) satisfaction of an inclination is objectively

good only if it contributes to an agent's happiness;[27] and (4) an agent's happiness is objectively good only if the agent has a morally good will.[28] Korsgaard summarizes this part of the argument in chapter 12 of CKE:

> … [2] the things we desire have value because we want and need them, not the reverse. Our desire is a condition of their value. [3] Our wanting them is not enough to make them good, however, for obviously many of the things we want are not good. Even if we want them we will not judge them good unless they are conducive to our happiness… [4] Then we raise the further question… If we say it makes him happy, we ask why it is good that he should be happy… this condition will always be the presence of a good will.[29]

Korsgaard maintains that this argument shows: (5) fully rational choices make their objects objectively good (rather than the other way around);[30] and therefore (6) the power of rational choice (autonomy) is the unconditioned condition (and thus the source) of the objective goodness of anything that has this property.[31]

Each of the first four premises of this argument is contentious: (1) is challenged regularly in disputes about the "guise of the good"; (2) would not be accepted by Mooreans; (3) might be disputed by hedonists; and (4) might be criticized by utilitarians. However, I shall be challenging (5) and (6) on grounds that are independent of whether (1) — (4) are granted: I contend that (1) — (4) are consistent with paradigmatically heteronomous theories of value.

To see why Korsgaard thinks (5) follows from (1) — (4), it will suffice to note that for any given end that is represented as objectively good (from (1)), the condition of its objective goodness is the moral goodness of an agent's will (from (2) — (4)). Because Korsgaard takes moral goodness and full rationality to be the same, it follows that full rationality is the condition of the objective goodness of any end: full rationality "makes" an end objectively good.

The trouble with (5) is in the parenthetical. A morally good will is one that pursues morally good ends. But if moral goodness and full rationality are the same, then a morally good will is one that pursues fully rational ends. This is a problem for Korsgaard because an end is fully rational if and only if it is objectively good, whence it may be concluded that a morally good will is one that pursues only objectively good ends — or, to put this another way, the pursuit of objectively good ends is the condition of a will being morally good, so pursuing objectively good ends "makes" a will morally good and, thereby, fully rational.

The point is easier to see by means of an example. Suppose that I order a pizza for dinner. The pizza is objectively good under the complex set of psychological and physiological conditions that make it necessary and pleasant for me to eat as well as the complex moral and social conditions that include the fact that the food is not stolen.[32] So if my choice is fully rational, then these conditions obtain. But the converse of this also holds, whence it may be concluded that fully rational choice "makes" its object objectively good in the same way that the objective goodness of an object "makes" its pursuit fully rational.

Another way to understand the objection I am raising is to note that what makes an object of choice objectively good (as opposed to merely conditionally good) is that the conditions of its goodness are fulfilled. But that is what makes a choice fully rational. So the objective goodness of an object makes a choice fully rational in the same way that the full rationality of a choice makes its object objectively good.[33] It might be thought that antirealism about value can bolster the move to (5), but I think this is mistaken: antirealism is not compatible with the regress argument, for it would undermine the very conclusion the regress argument is supposed to establish (*viz.* , that rational nature understood as the capacity for rationality is unconditionally good). Along the same lines, it should be noted that Korsgaard must be overstating her case when she says, for example, that objects of inclination "are in themselves neutral" (see note 26 above). Strictly speaking, for Korsgaard's argument to work, this claim must be false. Korsgaard should have admitted that objects of inclination are in themselves conditionally good: in themselves, objects of inclination are not objectively good — but they are also not neutral. However, I am not going to pursue this line of argument here. Even if antirealism can shore up the move to (5) (I do not believe it can, but if I am wrong about this), there is a deeper problem about the move to (6).

The problem with the move to (6) is that even if (5) is granted, what confers value (according to the argument) is the correct exercise of rational choice (i. e., fully rational choice), not the power of rational choice (which, if exercised incorrectly, might result in an irrational choice and, therefore, the realization of an end that is not objectively good).[34] Thus, it may be seen that even if the move from (1) — (4) to (5) is granted, Korsgaard's argument would show at most that a morally good will (rather than the capacity for such a will) is the condition of objectively good choices.[35]

However, Korsgaard might reply to this objection by introducing a further step.[36] Suppose that (1) — (5) show that (6*) what confers value (objective goodness) is the correct exercise of rational choice. Then it might be argued that (7) only by giving agents the opportunity to exercise their powers of rational choice will they be able correctly to exercise rational choice and thereby bring objective goodness into the world, so (8) agents should have the opportunity to exercise their powers rational choice.

But this reply, I think, does not help Korsgaard. In making this reply, Korsgaard in effect will have conceded that the power of rational choice is not unconditionally good: it is objectively good only on condition that it is exercised correctly.[37] And the reason I think this problem is so troublesome is that it reveals that the regress argument underdetermines the content of the morally good will: unless the regress argument can show that the power of rational choice is unconditionally good, it will not give any guidance regarding the standards which determine when one's rational choice is exercised correctly. Indeed, precisely because the value judgment in (8) about enabling agents to exercise their powers of rational choice is derived from the objective goodness of the correct exercise of rational choice, the value judgment is itself conditioned: one should allow agents this opportunity only when doing so is consistent with the correct exercise of one's own powers of rational choice. When will this be the case? Unfortunately, the regress argument yields no prescriptions on this front.

This should not be surprising. (2) — (4) involve a value scheme that distinguishes happiness from worthiness for happiness in a way that many theists would find intuitive: they might argue that God will confer happiness on agents in the afterlife if but only if those agents conducted themselves in this life so as to be worthy of that happiness. And (7) — (8) are analogous to what theists sometimes say in response to the problem of (moral) evil: an omnipotent, omniscient, omnibenevolent God is consistent with the existence of evil because the possibility of evil is a necessary consequence of creating beings with free will, and a world in which there are beings with free will is better than one in which no such beings exist. This is relevant because the point I want to make is that the regress argument is consistent with a religious view according to which what makes a good will good (i. e., worthy of happiness) is following the dictates of God (or, if you prefer, of Emma Woodhouse): the regress argument cannot ground a substantive ethics of autonomy.

5. Conclusion

I am not endorsing Korsgaard's interpretations of FUL and FH as exegetically correct or criticizing anything Korsgaard says on that front. I also am not arguing that there is no way to justify a substantive autonomy approach to ethics using FUL or FH — or that Kant does not do so successfully. I am trying to engage only with Korsgaard's arguments, and I am trying to do so on philosophical (rather than textual) grounds: if the objections in the previous sections work, then neither the argument from spontaneity nor the regress argument is able to show that an agent would be acting wrongly even if s/he acts in a paradigmatically heteronomous fashion (as Harriet does). If the objections in section 3 work, then although the argument from spontaneity succeeds, it does so at the price of normative force; instead of proving that all agents should will their maxims as universal laws (in a strong sense), the argument from spontaneity proves that all agents do will their maxims as universal laws (in an empty formalistic sense). If the objections in section 4 work, then the regress argument cannot ground a substantive ethics of autonomy because it leaves the content of the morally good will open; the regress argument is consistent with a (paradigmatically heteronomous) view that makes moral goodness equivalent to following the dictates of God.

Despite Korsgaard's remarks about the emptiness of FUL (in SN), she continues to use it throughout her work as a principle that tracks moral duties. She also continues to use FH throughout her work as a principle that tracks moral duties despite her claims (in her discussion of the morality of lying) about FH delivering nonbinding results. This suggests that Korsgaard has more than one interpretation of FUL and FH.

It is possible that Korsgaard's alternative interpretations of these principles are derived from other arguments. In chapter 8 of SC, Korsgaard suggests that her internal norm strategy faces a problem:

> [It] … explains why… action must meet the normative standard: *it just isn't action* if it doesn't. But it also seems as if it explains it rather too well, for it seems to imply that only good action really is action, and that there is nothing left for bad action to be.[38]

She then tries to solve this problem by noting the connection between successful action and the unity of an agent:

> The function of an action is to unify its agent, and so to render him the autonomous and efficacious author of his own movements. An unjust or unlawful action therefore fails to unify its agent, and so fails to render him the autonomous and efficacious author of what he does.[39]

This moves away from an argument beginning from a premise about acting under the idea of freedom (like the argument from spontaneity) or from a premise about pursuing ends that are represented as objectively good (like the regress argument); it suggests, instead, an argument that begins from a premise about the unity (and perhaps the constitution) of an agent, an argument that will be familiar to readers of SN. I confess that I am skeptical of this argument: it seems paradoxical to me to say that an agent must unify him/herself, for this separates the agent from whatever is doing the unifying. But in fairness to Korsgaard, she is aware of this paradox, and an investigation of whether her solution to it withstands critical scrutiny will have to await another day.

Bibliography

- Kant, I. (1996), *Practical Philosophy*, trans. by Mary Gregor, Cambridge: Cambridge University Press.
- Korsgaard, C. (1986), "Kant's Formula of Humanity, " Kant-Studien, 77, pp. 183-202.
- Korsgaard, C. (1989), "Morality as Freedom, " in Yovel, Y., ed., *Kant's Practical Philosophy Reconsidered*, Jerusalem: Kluwer Academic Publishers, pp. 23-48.
- Korsgaard, C. (1996A), *Creating the Kingdom of Ends*, Cambridge: Cambridge University Press.
- Korsgaard, C. (1996B), *The Sources of Normativity*, Cambridge: Cambridge University Press.
- Korsgaard, C. (2008), *The Constitution of Agency: Essays on Practical Reason and Moral Psychology*, Oxford: Oxford University Press.
- Korsgaard, C. (2009), *Self-Constitution: Agency, Identity, and Integrity*, Oxford: Oxford University Press.
- Melnick, A. (2002), "Kant's Formulations of the Categorical Imperative, " Kant-Studien, 93, pp. 291-308.
- O'Neill, O. (1989), *Constructions of Reason*, Cambridge: Cambridge University Press.
- Pogge, T. (1998), "The Categorical Imperative, " in Guyer, P., ed., *Kant's Groundwork of the Metaphysics of Morals: Critical Essays*, Lanham, Maryland: Rowman and Littlefield Publishers, pp. 189-213.
- Wood, A. (2006), "The Supreme Principle of Morality, " in Guyer, P., ed., *The Cambridge Companion to Kant and Modern Philosophy*, Cambridge: Cambridge University Press, pp. 342-380.
- Wood, A. (2007), *Kantian Ethics*, Cambridge: Cambridge University Press.

1. Korsgaard uses this example multiple times. See, for example, (Korsgaard, 2008), p. 125. See also (Korsgaard, 2009), pp. 162-3. ↩

2. In her more recent work, Korsgaard argues that "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…a constitutive principle for an inescapable activity is unconditionally binding" ((Korsgaard, 2009), p. 32). ↩

3. See, for example, (Korsgaard, 2008), p. 19. ↩

4. All quotation from Kant in this paper are taken from (Kant, 1996). All citations are given using the Academy Edition pagination which runs in the margins of the Gregor translation. ↩

5. Kant (notoriously) answers in the negative, but Allen Wood argues that this description of the case involves a subtle but important misunderstanding. According to Wood, Kant's case is not about (mere) lying: it is about perjuring. As Kant is imagining the case, when the murderer comes to your door, you are under oath just as you would be when testifying in a court of law. (See (Wood, 2007), chapter 14.) If Wood is correct about this, it changes the contours of the case significantly. It is difficult to imagine why or how you would be under oath when the murderer shows up at your door. But (as Wood points out) it is easy to imagine a more realistic example that seems to capture the essentials of the case in a way that might be simpler to think about: suppose that you are actually in a court of law, and suppose you know that if you tell the truth when you testify, the defendant (in whose absolute innocence you firmly believe) will be put to death. But suppose you also know that if you perjure yourself, the defendant most likely will be let off. The question is whether it is permissible to perjure in these circumstances. You might imagine yourself into Alyosha's shoes in *The Brothers Karamazov*: should he perjure himself to save Dmitri? In the book, Alyosha is not under oath when he takes the witness stand (and the death penalty is not looming). Nonetheless, he does not lie: he tells the truth and, predictably, Dmitri is found guilty. Perhaps Kant's negative answer to his right to lie case is less counterintuitive than a negative answer to the question about whether you may lie about the whereabouts of a Jew when the Nazis confront you (then again, perhaps not: perhaps both should have a positive answer. I am not going to take a stand on that here). ↩

6. (Korsgaard, 1996A), p. 137. ↩

7. *Ibid.*, p. 143. ↩

8. "The Formula of Humanity and its corollary, the vision of a Kingdom of Ends, provide an ideal to live up to in daily life as well as a long-term political and moral goal for humanity. But it is not feasible always to live up to this ideal, and where the attempt to live up to it would make you a tool of evil, you should not do so…even in the worst circumstances, there is always the Formula of Universal Law, telling us what we must not in any case do…The Formula of Universal Law provides the point at which morality becomes uncompromising" (*ibid.*, pp. 154-5). ↩

9. There is active debate about whether the different formulations of the categorical imperative are equivalent. I shall not be entering this debate here: I am neither endorsing nor criticizing Korsgaard's claims on this front. For some recent contributions to the debate about whether *Kant* intended the formulations to be equivalent, see (Wood, 2006) or (Pogge, 1998). For some recent contributions to the debate about whether, regardless of Kant's intentions, the formulations are equivalent, see chapter 7 of (O'Neill, 1989) or (Mel-

nick, 2002). ↩

10. (Korsgaard, 1996A), pp. 143-4. Note that at pp. 126-7, Korsgaard also suggests that if a maxim violates FUL (or, at least, the so-called contradiction in conception test associated with FUL), it violates FH. However, in the passage from 126-7, Korsgaard does not suggest that this goes only in one direction (as she does in the passage to which this footnote is appended). Moreover, the maxim she is considering is to "tell a lie for a certain purpose"—so what Korsgaard says in the passage from 126-7 seems to entail that a maxim of lying to the murderer at the door would fail FUL. These passages come from separate chapters, so the *prima facie* inconsistency between them suggests an evolution in Korsgaard's views. ↩

11. This chapter originally was published in an anthology: (Korsgaard, 1989). ↩

12. (Korsgaard, 1996A), p. 162. ↩

13. *Ibid.*, p. 163. ↩

14. This is actually an intermediate conclusion inferred from other premises: "Anything outside of the will counts as an alien cause, including the desires and inclinations of the person. The free will must be entirely self-determining. Yet, because it is a causality, it must act on some law or other…The free will therefore must have its own law. Alternatively, we may say that since the will is practical reason, it cannot be conceived as acting and choosing for no reason. Since reasons are derived from principles, the free will must have its own principle" (*Ibid.*, p. 163). However, the details of this are less important for my purposes. ↩

15. *Ibid.*, p. 164. ↩

16. *Ibid.*, p. 166. ↩

17. *Ibid.*, p. 175: "(i) We must act under the idea of (at least negative) freedom; (ii) we must therefore act on maxims we regard ourselves as having chosen; (iii) by the Argument from Spontaneity…we are led to the moral law…; (iv) our ability to act on the moral law teaches us that we *are* (negatively) free; (v) if so, we are members of the intelligible world, and have a higher vocation than the satisfaction of our desires; and (vi) this provides us with the incentive to be positively free." In this argument, Korsgaard distinguishes between negative freedom and positive freedom: negative freedom is the ability not to act according to one's inclinations; positive freedom is the ability to act according to a self-given law. ↩

18. *Ibid.*, p. 99. ↩

19. Thinking of oneself as a citizen in a realm of ends is another of Kant's ideas and, importantly for current purposes, it is one that Korsgaard associates with FH in her discussion of the murderer at the door case in CKE. Thus, it may be seen that in SN, Korsgaard still seems to be committed to the view that the premises of an argument for FUL do not need to be as strong as those for an argument for FH. A similar passage can be found on pp. 79-80 of (Korsgaard, 2009). ↩

20. (Korsgaard, 1996B), p. 99. ↩

21. Korsgaard repeats the argument from spontaneity in at least two places in *The Constitution of Agency: Essays on Practical Reason and Moral Psychology*. In each, her conclusion is the same as that in CKE and SN: "[t]he categorical imperative [FUL] is therefore the law of a free will" (p. 110); "the categorical imperative [FUL] *just is* the law of a free will" (p. 320). The only difference is that Korsgaard now supplements

the argument from spontaneity with an argument against particularistic willing. But Korsgaard's argument about particularistic willing is at cross-purposes with the issue I am raising about the argument from spontaneity, and Korsgaard is aware of this. This may be seen from the fact that, after repeating her argument about particularistic willing in section 4.4 of SC, she concludes as follows: "In *The Sources of Normativity*, I distinguished what I called 'the categorical imperative' from what I called 'the moral law.' The categorical imperative is the law of acting only on maxims that you can will to be universal laws. The moral law, as I characterized it there, is the law of acting only on maxims that all rational beings could act on together in a workable system. The arguments I've given above don't—or don't obviously—get us all the way to a commitment to the moral law in that more specific sense…" ((Korsgaard, 2009), p. 80). ↩

22. I would include Korsgaard's discussion of the murderer at the door case with her earlier work in this context because in that article she does seem to think that substantive moral duties can be derived from FUL (although her conclusion, as already noted, is that FUL is not as strict as FH). ↩

23. This chapter originally was published as an article: (Korsgaard, 1986). ↩

24. I think this is actually an intermediate conclusion that Korsgaard infers from two other premises: (1) agents always take themselves to act under the direction of reason, and (2) to act under the direction of reason is to take one's ends to be objectively good. This may be seen from the following passage: "…when we act under the direction of reason, we pursue an end that is objectively good. But human beings…take themselves to act under the direction of reason" ((Korsgaard, 1996A), p. 116). It is also notable that in both SN and *The Constitution of Agency: Essays on Practical Reason and Moral Psychology*, Korsgaard seems to suggest that to act on self-given laws is to regard one's ends as objectively good, thereby connecting this premise of the regress argument with the argument from spontaneity (see, e.g., (Korsgaard, 1996B), p. 107 or (Korsgaard, 2008), pp. 227-30). However, these details are less important for current purposes. ↩

25. (Korsgaard, 1996A), p. 118. ↩

26. *Ibid.*, p. 121: "…the things you want, if they are good at all, are good because you want them…The objects of inclinations are in themselves neutral…" ↩

27. *Ibid.*, p. 121: "…we can easily agree that there are some inclinations of which we want to be free: namely, those whose existence is disruptive of our happiness…" ↩

28. *Ibid.*, p. 121: "…we do not believe that happiness is good in the possession of one who does not have a good will." ↩

29. *Ibid.*, p. 345. See also chapter 9, pp. 260-1. ↩

30. *Ibid.*, p. 261: "Value…does not travel from an end to a means but from a fully rational choice to its object. Value is…'conferred' by choice." ↩

31. *Ibid.*, p. 123: "…the unconditioned condition of the goodness of anything is…the power of rational choice." ↩

32. Korsgaard would agree; the first half of this sentence is almost a quotation (*ibid.*, p. 267). ↩

33. I believe the point still stands (although it might be more complicated to make) even if one concedes, for example, that fully rational choices can be made on the basis of false beliefs. ↩

34. Again, I believe the point still stands (although it might be more complicated to make) even when one begins to grapple with the ways in which false beliefs can infect (ir)rational choices. ↩

35. Korsgaard seems to concede this in a later formulation of the regress argument: "…it turns out to be a good will that is the source of all value" (*ibid.*, p. 241). It is unclear whether she realizes that this is inconsistent with her earlier formulation of the argument. ↩

36. I would like to thank Daniele Bertini for instructive discussion of this point. ↩

37. That Korsgaard would not have conceded this point about the conditioned goodness of the power of rational choice can be seen from the following passage: "The capacity for rational choice…is an unconditional end…as an unconditional end it is the condition of the goodness of all our other ends. If humanity is not regarded and treated as unconditionally good then nothing else can be objectively good" (*ibid.*, p. 125). ↩

38. (Korsgaard, 2009), p. 160. ↩

39. *Ibid.*, p. 161. Note that on page 160, Korsgaard argues that autonomy is "an essential metaphysical property of action" rather than that it is an essential property of *free* action, which seemed to be the starting point of the original argument from spontaneity. Note also that this seems to contradict the second sentence of the quotation to which this footnote is appended: if unlawful *action* fails to render its agent autonomous, then it looks like autonomy is not an essential metaphysical property of action. ↩
