European Journal of Philosophy

DOI: 10.1111/j.1468-0378.2011.00480.x

The Problem of Agency and the Problem of Accountability in Kant's Moral Philosophy

Iuliana Corina Vaida

Abstract: This paper discusses the function and scope of incompatibilist or transcendental freedom in Kant's moral philosophy. The prevailing view among scholars, most notably Allison, is that the function of transcendental freedom is to enable us to articulate a first-person conception of ourselves as rational agents involved in deliberation and choice. Thus, the scope of transcendental freedom is rational agency in general. In order to perform this function, freedom has to be merely conceivable. *Pace* Allison, I argue that our first-person conception is neutral with respect to causal determinism, and that the function of transcendental freedom is to provide the metaphysical conditions of the possibility of genuine moral responsibility and perfect justice, and to get rid of moral luck. In order to perform this function, transcendental freedom has to be not just conceivable, but metaphysically real. My view suggests that we only have reason to attribute freedom to ourselves in situations in which we are aware that the moral law commands us categorically. We do not have a similar reason to believe we are free in purely prudential choices. Thus, the scope of transcendental freedom is not rational agency in general, but only moral agency.

1. Two Distinct Views of the Role of Transcendental Freedom in Kant's Philosophy

In this paper I will assume as a fact that Kant attributes to rational human beings a kind of freedom that other philosophers have found to be impossible or even unintelligible: transcendental or incompatibilist freedom, a freedom that is incompatible with causal determinism and even with a broader conception of natural causality.[1] Given the difficulties attending such a position, it is important that we answer, on Kant's behalf, the following two questions:

> (*Q1*) What is the *function* or *role* that transcendental freedom is supposed to play in our conception of ourselves as rational agents as well as in our moral practices?
> (*Q2*) What is the *scope* of transcendental freedom, that is, what are the kinds of choice that we have reason to believe are free in the transcendental sense?

Regarding (*Q1*), there are two distinct views of the role that freedom plays in Kant's philosophy. They can be understood by appeal to a distinction introduced by Jonathan Bennett, between two Kantian problems involving freedom: the problem of agency and the problem of accountability (Bennett 1974: ch. 10). The problem of agency involves a conflict between the perspective of the agent who, while deliberating, views the future as open, and the perspective of the observer who tries to predict the agent's future action by applying causal laws to his present state, whereas the problem of accountability involves the conflict between moral responsibility and determinism.

According to the first view, the primary role of freedom in Kant's philosophy is to articulate a first-person conception of ourselves as rational agents involved in deliberation and choice. This view regards the problem of agency as fundamental and the problem of accountability as arising from, and being reducible to, the problem of agency. It claims that we regard ourselves as genuine agents only because we are able to engage in a rational process of deliberation and choice 'under the idea of freedom', and we hold ourselves accountable for our actions only because we regard ourselves as genuine agents.

According to the second view, the primary role that freedom plays in Kant's philosophy is to ground genuine moral responsibility, the kind of responsibility without which we couldn't be regarded as justly deserving praise or blame, reward or punishment. On this view, our first-person conception of ourselves as rational agents does not require that we attribute freedom to ourselves, but genuine responsibility does. This gives rise to the problem of accountability: we must show that causal determinism does not rule out the possibility of freedom.[2]

The first view, which asserts the primacy of the problem of agency over the problem of accountability, is present in the writings of such prominent Kant scholars as Henry Allison and Christine Korsgaard.[3] I think it can be regarded as the received view. The received view answers not only (*Q1*), but also (*Q2*): if the primary *role* of freedom is to enable us to articulate a first-person conception of ourselves as rational agents involved in deliberation and choice, it follows that the *scope* of freedom is rational choice or rational agency in general, both moral and non-moral.[4]

In this paper I will argue against the received view, and defend the view according to which the primary role that transcendental freedom plays in Kant's philosophy is to ground moral responsibility. The focus of my discussion will be Henry Allison's version of the received view, as articulated in his 1990 book *Kant's Theory of Freedom*. In Section (2), I will break down Allison's view into two theses, representing his answers to (*Q1*) and (*Q2*), and formulate my own view in opposition to his. Sections (3) and (4) will contain my two-pronged argument against Allison. On one prong, I will argue that our first-person perspective on and understanding of ourselves as rational agents do not require an appeal to transcendental freedom. I will argue that the 'key ingredient' in our conception of ourselves as rational agents is not transcendental or indeterminist freedom, but self-consciousness or reflectivity which, when approached from the first-person perspective, is neutral with respect to causal determinism. On the other prong of

the argument, I will argue that, even if Allison were right and our first-person perspective on ourselves as rational agents did require transcendental freedom or spontaneity, it would require the *mere conceivability* of freedom, not its *metaphysical reality*, whereas moral responsibility requires nothing less than its full blown *metaphysical reality*. It is because freedom has to be real, not just conceivable, that it represents 'the stumbling block of all empiricists' (*CPrR*, 5:7). Thus, while the problem of agency can be dealt with satisfactorily within an empiricist framework, the solution to the problem of accountability requires the appeal to distinctively Kantian presuppositions, such as the doctrine of transcendental idealism. In asserting the primacy for Kant of the problem of accountability, I will try to convey the image of a philosopher preoccupied to banish any form of luck from the realm of morality and to provide the metaphysical conditions of possibility for moral responsibility and perfect justice.

My view of the role of freedom in Kant's philosophy suggests that we only have reason to attribute transcendental freedom to ourselves in moral situations, when we have to choose between morality and self-love. If the only reason why we attribute freedom to ourselves is to establish genuine responsibility in those situations in which the moral law commands us categorically, then we shouldn't attribute freedom to ourselves in non-moral situations. Thus, my answer to (*Q1*) suggests a surprising answer to (*Q*2), an answer that, as far as I know, no one has considered seriously: that the scope of transcendental freedom is not rational agency in general, but only moral agency. Thus, there may be a metaphysical gap between moral and non-moral agency: the presence and exercise, in the former, but not in the latter, of a power that elevates us above the natural world, by extricating us from the web of natural causality.

Since my view presupposes a clear-cut distinction between moral and non-moral agency, in Section (5) I will first briefly discuss the possibility and some of the implications of such a distinction, and then consider some arguments in favour of a metaphysical gap between moral and non-moral agency.

Though the position presented in this paper is meant as a reconstruction of the historical Kant, and so I often defend it by appeal to textual evidence, my interest in it is not merely historical. I believe that Kant's position, as I understand it here, is supported by compelling arguments and that it may be, by-and-large, the correct account of the function and scope of incompatibilist freedom in human agency.[5]

2. Allison's View Articulated and Opposed

A clear illustration of the received view can be found outside Kantian scholarship in Thomas Nagel's 1986 book *The View from Nowhere*. Nagel introduces the problem of agency first, arguing that our ordinary, pre-reflective belief that we are the authors of our own actions conflicts with the external or objective view of actions as events in the natural order. But Nagel also holds that, at a closer look, this belief in our own agency or autonomy is unintelligible: 'I cannot say what would, if it were true, support our sense that our free actions originate with us'

(Nagel 1986: 117). Our impression of ourselves as autonomous fails to justify our belief in our own autonomy first, because, as any impression of something, it does not guarantee the reality of that something, and second, because we are not even able to analyze what the impression is an impression of (Nagel 1986: 117).

After introducing the problem of agency, Nagel moves to the problem of accountability, which he sees as arising from the former as follows:

> The same external view that poses a threat to my own autonomy also threatens my sense of the autonomy of others, and this in turn makes them come to seem inappropriate objects of admiration and contempt, resentment and gratitude, blame and praise. (Nagel 1986: 112)

Nagel thinks that the problem of accountability is reducible to the problem of agency because he believes that all we have to do in order to formulate a judgement of responsibility is to 'enter into the defendant's point of view as an agent', attaining 'a vicarious occupation of his point of view and evaluation of his action from within it' (Nagel 1986: 120).

Since the problem of accountability arises as the projection on others of the first-person problem of agency, it inevitably partakes in the latter's unintelligibility: 'the judge's sense of the defendant's alternative is revealed as an illusion which derives from the judge's projection of his own illusory—indeed unintelligible—sense of autonomy into the defendant' (Nagel 1986: 123). Nagel concludes that both problems are unintelligible (or involve unintelligible concepts), and thus impossible to solve.

In the context of Kantian scholarship, Christine Korsgaard indicates that she believes in the primacy of the problem of agency when she writes: 'it is primarily your own freedom that you are licensed to believe in, and, as a consequence, it is primarily yourself that you hold imputable' (Korsgaard 1996a: 174), and 'at the moment of decision, you must regard yourself as the author of your action, and so you inevitably hold yourself responsible for what you do' (Korsgaard 1996b: 206). Holding others responsible is 'an inevitable concomitant of holding ourselves so' (Korsgaard 1996b: 209), since we cannot enter into genuinely reciprocal relations with people unless we regard them as we view ourselves, namely as free agents, authors of their own actions.

In what follows, I will focus on Henry Allison's version of the received view—since his is one of the clearest and most explicit accounts of the problem of agency in Kant's moral philosophy.

According to Allison, the locus where Kant formulates and attempts to solve the problem of agency is the third antinomy of pure reason. Allison proposes that we view the third antinomy not merely as a conflict between two cosmological models, one involving a finite cause/effect chain proceeding from a first cause or unmoved mover outside of nature, the other involving an infinite cause/effect chain of natural events, but also as a conflict between two models of rational agency: a first-person, incompatibilist, spontaneity-based model, and a third-person, compatibilist, belief-desire model (Allison 1990: 11). According to the

incompatibilist conception, the agent's decision to act in a certain way is not the causal consequence of any antecedent condition (including the agent's psychological state), but the expression of the agent's capacity of self-determination independently of the causality of nature. We adopt the incompatibilist model when we try to understand ourselves as deliberating agents. According to the compatibilist conception, the agent's decision to act in a certain way is determined by her antecedent psychological state. The compatibilist conception is the familiar Humean belief-desire model which 'leaves ample ''elbow room'' for freedom of the familiar compatibilist sort' (Allison 1990: 5). We adopt the compatibilist model when we try to explain and predict human behaviour.

Allison regards the problem of agency as fundamental and problem of accountability as reducible to the problem of agency because he believes that all we have to do in order to formulate a judgement of praise or blame is to take the perspective of the deliberating agent:

> This [spontaneity-based] model is operative both in the context of deliberation, where it characterizes how one takes oneself qua engaged in a deliberative process, and in the context of appraisal or imputation, where it grounds judgments of praise and blame . . . (Allison 1990: 38–9)

Assuming the centrality of the problem of agency, Allison argues that the role of transcendental freedom or spontaneity in Kant's philosophy is to provide us with a first-person conception of ourselves as rational agents. He also believes that Kant's spontaneity-based model correctly captures our first-person conception of ourselves as rational agents. He writes that 'it is a condition of the possibility of taking oneself as a rational agent that one attribute such spontaneity to oneself' (Allison 1990: 45), and that 'I cannot coherently think of myself as a rational agent without also attributing to myself such spontaneity' (Allison 1996a: 127). This view of the role of freedom leads Allison to conclude that the scope of freedom is rational choice or rational agency in general.

I believe that the core of Allison's view is correctly captured by the following two theses, which represent his answers to our questions (*Q1*) and (*Q2*) respectively:

> *Thesis I*: (a) The *function* of transcendental freedom is that of 'key ingredient' in an incompatibilist conception of agency that captures our first-person conception of ourselves as rational agents engaged in deliberation and choice. (b) The *Critique of Pure Reason* conception of *arbitrium liberum* is Kant's main vehicle for the formulation of his incompatibilist conception of rational agency.
> *Thesis II*: The exercise of transcendental freedom is not restricted to moral agency, but underlies our entire rational agency.

Part (a) of *Thesis I* states the view, attributed by Allison to Kant and endorsed by Allison himself, that our first person conception of ourselves as rational agents presupposes a moment of spontaneity. Part (b) of *Thesis I* identifies the main textual evidence for (a) in the first *Critique*. In fact, Allison believes that Kant's

clearest formulation of the spontaneity-based conception of ourselves as rational agents occurs much later, in *Religion within the Limits of Reason Alone*, in the following passage:

> Freedom of the will is of a wholly unique nature in that an incentive can determine the will to an action *only insofar as the individual has incorporated it into his maxim* (has made it into the general rule in accordance with which he will conduct himself . . .). (*RWLRA*, 6:24)

Allison explains the content of this passage, which he refers to as the Incorporation Thesis, as follows: an inclination or desire does not of itself, by its mere occurrence, constitute a sufficient reason to act, but does so only insofar as the agent takes it up or incorporates it into a maxim. Moreover, the act of 'taking up' or 'incorporating' is regarded from the first-person perspective not as the causal consequence of the desire, but as an 'act of spontaneity on the part of the agent' (Allison 1990: 40–1; 1996a: 109). However, the passage from *Religion within the Limits of Reason Alone* is hardly enough to make a case for Kant's alleged spontaneity-based conception of ourselves as rational agents, so Allison argues that the Incorporation Thesis is equivalent to the first *Critique* conception of *arbitrium liberum*. This is defined as the conception of a will that, though affected by sensuous desires, is not necessitated by these desires, but has a power of self-determination 'through motives which are represented only by reason' (A802/B830).

A clarification may be needed here regarding *Thesis I*: the moment of spontaneity involved in the first-person conception of ourselves as rational agents is not intended to capture an alleged inner experience of transcendental freedom or spontaneity of choice. Kant and Allison claim there is no such experience. Rather, the first-person incompatibilist conception that Allison attributes to Kant is a purely intellectual conception we are compelled to adopt towards ourselves every time we deliberate.[6]

Allison's two theses are related in that *Thesis I* provides support for *Thesis II* in the following ways. First, if the function of transcendental freedom is to capture our first-person conception of ourselves as rational agents, then we have reason to believe that its scope is rational, not just moral, agency, since our first-person conception of ourselves is the same in moral and non-moral situations. Second, the *arbitrium liberum*, which purportedly expresses this spontaneity-based conception of agency, clearly represents a conception of rational, not just moral, agency.

Against Allison's two theses, I will argue for the following:

> *Anti-thesis I*: (a) The *function* of transcendental freedom or spontaneity is to provide the metaphysically indispensable condition for a conception of moral responsibility and moral worth which renders them impervious to any form of luck. (b) Kant's conception of *arbitrium liberum*, rather than expressing an indeterminist conception of agency, is neutral with respect to the problem of determinism.
> *Anti-thesis II*: The exercise of transcendental freedom is restricted to moral agency.

Anti-thesis I embodies my disagreement with Allison over the function of freedom in Kant's philosophy. This disagreement is set against the backdrop of the more general disagreement I have with the received view, over the way in which the two Kantian problems involving freedom, the problem of agency and the problem of accountability, are related.

My argument for *Anti-thesis I* and against Allison's *Thesis I* is two-pronged. On one prong, in Section (3), I will argue that our first-person perspective on and understanding of ourselves as rational agents do not involve an appeal to transcendental freedom. The 'key ingredient' in our conception of ourselves as rational agents is not transcendental freedom, but *self-consciousness* or *reflectivity* which, as far as it is restricted to the first-person perspective, is neutral with respect to causal determinism. Though in my discussion I will focus on the *arbitrium liberum* formulation of Kant's conception of rational agency, my comments equally apply to the Incorporation Thesis.

While I do not object to Allison's characterization of the *arbitrium liberum* as capturing our first-person perspective on deliberation and choice, I will argue, *pace* Allison, that the *arbitrium liberum* is equivalent to a belief-desire model of agency which has been adequately enriched to incorporate the feature of *reflectivity* distinctive of human consciousness, and that this first-person model of agency is neutral with respect to causal determinism.

On the other prong of my argument, in Section (4), I will argue that even if Allison were right and our first person perspective on ourselves as rational agents required transcendental freedom or spontaneity, it would require the mere *conceivability* of transcendental freedom, whereas genuine moral responsibility requires nothing less than its full-blown *metaphysical reality*. Thus, I will argue that the problem of agency can be dealt with satisfactorily within an empiricist framework, whereas the solution to the problem of accountability requires the appeal to distinctively Kantian theses, such as the doctrine of transcendental idealism, or at least his transcendental theory of experience.[7] My conclusion will be that Allison gets the order of priority wrong: for Kant, the problem of accountability is more fundamental than, and not reducible to, the problem of agency. My position is further supported by Kant's preoccupation with banishing any form of luck from the realm of morality, for the purpose of providing the metaphysical conditions of the possibility of genuine moral responsibility and perfect justice.

Just as Allison's *Thesis I* supports *Thesis II*, my *Anti-thesis I* supports *Anti-thesis II*: if our first-person conception of ourselves as rational agents is neutral with respect to causal determinism and we attribute transcendental freedom to ourselves only because genuine moral responsibility requires us to, then prima facie we have no reason to attribute transcendental freedom to ourselves in non-moral situations. Of course, one reason why we would want to extend transcendental freedom to our entire rational agency is to have a unified account of agency and avoid a metaphysical gap between moral and non-moral agency, a gap that is not supported by introspective evidence. While I have nothing to say against the general appeal of theoretical unification, in Section (5) I will offer some considerations in favour of the metaphysical gap.

3. The *Arbitrium Liberum*, or the First-Person Conception of Ourselves as Rational Agents, as Neutral with Respect to Causal Determinism

The *arbitrium liberum* conception appears twice in the *Critique of Pure Reason*, first in the *Dialectic*, where Kant defines it as the conception of a will that is affected, but not necessitated, by sensuous motives (A534/B562), and then in the *Canon*, where he specifies that the *arbitrium liberum* can be determined 'through motives which are represented only by reason', motives that are 'representations of what, in a more indirect manner, is useful or injurious' (A802/B830). This is the concept of a will which, though affected by sensuous desires, is not determined by these desires, but has a power of self-determination through representations 'of what, in a more indirect manner, is useful or injurious' (A802/B830). The *arbitrium liberum* is contrasted with the *arbitrium brutum*, which is a will determined to respond to the strongest sensuous desire (A534/B562, A802/B830).

Allison claims that the *arbitrium liberum* captures our first-person conception of ourselves as rational agents. We cannot engage in deliberation and choice unless we view ourselves as capable to frame ends for ourselves[8] and to take or reject inclinations or desires as sufficient reasons for action in the light of these ends. Allison believes that attributing this capacity to ourselves amounts to attributing transcendental freedom or spontaneity to ourselves (Allison 1990: 41).

Though I do not object to Allison's general characterization of the *arbitrium liberum*, I am among those who fail to see why the capacity of self-determination as described above requires spontaneity.[9] I believe Kant's distinction between *arbitrium liberum* and *arbitrium brutum* is purported to capture the distinction between human and animal psychology, and this distinction can be ascertained empirically. The criteria for the attribution of *arbitrium liberum* to human beings are entirely empirical: first-person introspective evidence and third-person behavioural signs (just like the criteria for the attribution of mental states in general). By contrast, the attribution of transcendental freedom or spontaneity is not an empirical matter.[10]

In the effort to represent the *arbitrium liberum* as an incompatibilist conception of agency, Allison contrasts it with what he regards as 'the standard compatibilist belief-desire model': according to the former, it is the value that the agent places on a desire, in an act of spontaneity, that gives it its status as a sufficient reason to act, whereas according to the latter, desires come with pre-assigned weights and the agent is determined to act in a quasi-mechanistic fashion by the strongest desire (Allison 1990: 39–41; 1996a: 113; 1996b: 130–1).

However, I think that, in his attempt to contrast the two models of agency, Allison is guilty of oversimplifying and ultimately misrepresenting the belief-desire model: any belief-desire model worthy of consideration, even one devised solely for the purpose of explaining and predicting human behaviour, has to share with our first-person conception of ourselves an essential feature of our psychological make-up, namely *self-consciousness* or *reflectivity*. Any belief-desire model which assumes that our desires come with pre-assigned weights is bound to be immediately falsified by experience—not only by introspection, but also by

the attempt to explain and predict human behaviour. To have a chance to successfully explain and predict human behaviour, a belief-desire model has to take into account the fact that, at least on some occasions, when we are confronted with desires pulling us in different directions, we do not simply act on the strongest one, but step back and weigh our desires in the light of various considerations. These considerations often end up modifying the initial strengths of our desires. Allison's 'standard belief-desire model', which he contrasts with the *arbitrium liberum*, fails to take this fact into account. Consequently, the principle of charity should prevent us from attributing it to anybody, including Hume.

Before going any further, let us take a closer look at this essential feature of the human mind, *self-consciousness* or *reflectivity*, which must be included in any model of rational agency, whether it purports to explain and predict human behaviour or to express our first-person conception of ourselves as deliberating agents.

Christine Korsgaard gives an illuminating account of how we should understand reflectivity. Reflectivity frees the mind from the direct control by our inclinations or impulses and in this way we no longer act on desires, but on reasons. This is how Korsgaard describes it:

> The human mind *is* self-conscious in the sense that it is essentially reflective . . . We human animals turn our attention on to our perceptions and desires themselves, on to our own mental activities, and we are conscious *of* them. That is why we can think *about* them.
>
> And this sets us a problem no other animal has. . . . For our capacity to turn our attention on to our own mental activities is also a capacity to distance ourselves from them, and to call them into question. I perceive, and I find myself with a powerful impulse to believe. But I back up and bring that impulse into view and then I have a certain distance. Now the impulse doesn't dominate me and I have a problem. Shall I believe? Is this perception really a *reason* to believe? I desire and I find myself with a powerful impulse to act. But I back up and bring that impulse into view and then I have a certain distance. Now the impulse doesn't dominate me and now I have a problem. Shall I act? Is this desire really a *reason* to act? The reflective mind cannot settle for perception and desire, not just as such. It needs a reason. (Korsgaard 1996d: 92–3)

Is the belief-desire model, just in virtue of its structure, bound to the assumption that desires come with pre-assigned weights, or can it be conceptually enriched in order to include the feature of reflectivity? I believe it can, along the lines proposed by Harry Frankfurt in his influential 1997 paper 'Freedom of the Will and the Concept of a Person'. Frankfurt introduces the concepts of *first-order desire* and *second-* and *higher-order desire* and *volition* in order to capture our capacity to reflect upon the desirability of our first-order desires and decide what first-order desires we want to identify ourselves with. According to Frankfurt,

the essential feature that makes a person a person is that, rather than acting on his first-order desires, he has, and can act on, a second-order volition, that is, a second-order desire that a particular first-order desire be effective or move him 'all the way to action' (Frankfurt 1997: 171–6).

But as soon as we include the feature of reflectivity as higher-order desire into the belief-desire model, the contrast between the belief-desire model and the *arbitrium liberum* disappears. I don't see any reason not to equate the *arbitrium liberum* with such a sophisticated belief-desire model: the motives, operative in the *arbitrium liberum*, which are 'represented only by reason', in the light of which the agent judges a desire as being 'useful or injurious' (A802/B830), are the equivalent of Frankfurt's second-order desires. Passages from the *Critique of Practical Reason* also suggest that Kant holds a conception of *desire* which is broad enough to accommodate a distinction between first- and second-order desires. Though of course Kant does not explicitly draw such a distinction, he draws an arguably similar distinction between determining grounds of action that are instinctive and determining grounds of action that are 'thought by reason' (*CPrR*, 5:96–7).

However, there might still be a way in which Allison's contention that the *arbitrium liberum* is an incompatibilist model of agency can be vindicated: namely, if it can be shown that Korsgaard's notion of self-consciousness, or Frankfurt's notion of higher-order desire, requires transcendental freedom or spontaneity. We have to make sure that, by including the feature of self-consciousness or higher-order desire into the simple belief-desire model, we are not surreptitiously introducing an element that is qualitatively different from the other 'empirically accessible causal factors' (Allison 1990: 39): the spontaneity of the agent or the act of self-determination. To be clear, the question is not if transcendental freedom or spontaneity can be added to a belief-desire model, but rather if it is inevitably added when we get from the simple belief-desire model to the sophisticated one.[11]

The answer, I think, is no. Let us first look at what Frankfurt and Korsgaard have to say. Frankfurt explicitly states that his conception of agency based on the idea of higher order desires is neutral with respect to the problem of determinism (Frankfurt 1997: 182). The freedom of the will which is specific to persons and which his model emphasizes is the freedom to 'have the will [one] wants' or to be able to secure 'the conformity of [one's] will to [one's] second-order volitions'. This is compatible with the possibility that it is causally determined what second-order volitions a person actually has, and also that it is causally determined that a person has the will he wants to have (Frankfurt 1997: 179–82).

Korsgaard too does not equate reflectivity and the freedom that comes with it with the incompatibilist or transcendental freedom involved in Kant's third antinomy. As she puts it, she does not believe that 'our experience of our freedom is scientifically inexplicable', but that it can be explained 'in terms of the structure of reflective consciousness' (Korsgaard 1996d: 96–7). Just like simple animal consciousness, self-consciousness is a natural fact.[12] Thus, both Frankfurt and Korsgaard reject the idea that second-order volition or self-consciousness involves indeterminist freedom.

In what follows, I propose that our first-person conception of ourselves as rational agents, as expressed by Kant's conception of *arbitrium liberum*, is neutral with respect to the issue of causal determinism. In my view, this is simply because our first-person understanding of ourselves is not penetrating enough to have a stand on the issue of determinism, and because we do not need to take a stand in order to engage in deliberation and choice.

However, it is to be expected that such a neutral first-person model of rational agency will be perceived as incomplete and will be subjected to pressures from various philosophical or intellectual quarters. To put this model to any useful work beyond expressing our first-person conception of ourselves as rational agents, we have to add to it either causal determinism or incompatibilist freedom. It follows that any determinist or indeterminist conception of agency is a theoretical construct which includes assumptions that go beyond our first-person understanding of ourselves as rational agents, and are motivated by broader concerns, such as the attempt to include the human *psyche* into the naturalistic worldview or, alternatively, to establish its mysterious moral accountability.

Thus, on the one hand, we may approach the *arbitrium liberum* with the scientific attitude and aim to provide explanations and predictions of human behaviour. In this case, insofar as causal determinism is an indispensable condition of scientific knowledge, we have to assume that the sophisticated belief-desire model is subsumed under causal determinism. The assumption that causal laws govern our decision-making processes does not conflict with our first-person conception of ourselves as rational agents.[13] The addition of this assumption to our first-person conception turns it into a model of agency which goes beyond what is accessible from the first-person perspective and which can be used as an instrument for scientific explanation and prediction.

On the other hand, we may approach the *arbitrium liberum* driven by moral concerns, adding to it assumptions that justify the attribution of responsibility and constitute metaphysical conditions of the possibility of ideal justice. I think Kant is right when he claims that genuine moral responsibility requires freedom or spontaneity.[14] Thus, we do not need to attribute spontaneity to ourselves as part of our self-conception, but we must attribute it to ourselves if we are to hold ourselves morally responsible.

Let me pause here in order to underscore a disagreement between my view and both Allison's and Nagel's views. I do not believe that the scientist's compatibilist conception of agency conflicts with the first-person perspective, with our admittedly obscure sense of 'being the authors of our own actions'. Unlike Nagel and Allison, it seem to me that regarding my actions as the causal consequence of my past does not conflict with considering myself the author of the same actions (though it conflicts with considering myself the author of my own self). One may even argue that it is the causal deterministic conception, not the spontaneity-based conception, that thoroughly anchors the agent's choice in *what* or *who he is*, namely in his values, goals, and commitments. Even if these values and goals are causally determined, they, together with the capacity to reflect upon and modify them, constitute *what* or *who I am*.

I think we too often overlook the fact that acting is as much a matter of creating something new, as it is a matter of discovering something that already exists. The idea that, when I decide, I view the future as open does not necessarily mean that I attribute to myself the power to choose something that does not follow from who I am and the circumstance I find myself in. The future may appear open to me not because I haven't yet performed the spontaneous act that will bring about my decision, but simply because I have not yet gone through the process of deliberation—a process which is not entirely transparent.[15] What I am trying to discover while I deliberate is not merely what to do, but who I am, what I really desire, 'what I want from life'. In doing so, I am working with and exploring a material that has already been given to me—genetically, by the environment, and by education. All these things may determine my choice. But they are not something alien, they are or constitute *my self* (or, at least, my non-moral self).

There is a famous passage in the *Canon*, which has puzzled many commentators, that supports my claim that Kant regarded his conception of the *arbitrium liberum* as neutral with regard to causal determinism. Referring to the *arbitrium liberum*, Kant explicitly states it is an open question if the freedom involved in the causality of reason is, at bottom, part of natural causality:

> Whether reason is not, in the actions through which it prescribes laws, itself again determined by other influences, and whether that which, in relation to sensuous impulses, is entitled freedom, may not, in relation to higher and more remote operating causes, be nature again, is a question which in the practical field does not concern us, since we are demanding of reason nothing but the rule of conduct; it is a merely speculative question, which we can leave aside so long as we are considering what ought or ought not to be done. (A803/B831)

In the above passage, Kant neither claims nor denies that the freedom involved in the *arbitrium liberum* is governed by natural laws. What he claims is that, from the first-person perspective, the perspective we endorse when we are considering 'what ought or ought not to be done', we do not know and do not need to know whether we are free or not in the transcendental sense. Self-consciousness or reflectivity makes us rational agents, but the reality of transcendental freedom remains a theoretical or speculative problem.

Moreover, in the second *Critique*, when Kant contrasts the 'determining grounds [of action] thought by reason' with 'determining grounds that are instinctive', he makes it clear that the possession of these 'higher-order desires' is not equivalent with the possession of incompatibilist freedom or, as Kant refers to it, 'that freedom which must be put at the basis of all moral laws and the imputation appropriate to them' (*CPrR*, 5:96).

To sum up my main claims in this section, I have argued, against Allison's *Thesis I*, that the essential ingredient in our conception of ourselves as rational agents, as expressed by Kant's conception of *arbitrium liberum*, is self-consciousness or reflectivity. By the contrast between *arbitrium liberum* and

arbitrium brutum Kant tries to capture an empirical, observable difference between simple animal consciousness and human self-consciousness. This difference does not consist in some non-empirical or metaphysical feature of human psychological make-up, but in a feature accessible both to introspection and to third-person observation: the freedom to take a step back and reflect on our desires, so that we act only on those which we endorse. Our conception of ourselves as rational agents, which incorporates this freedom or reflectivity, is neutral with respect to causal determinism.

4. The Primacy of the Problem of Accountability

On the second prong of my argument against Allison's *Thesis I,* let me for the sake of argument grant Allison that our first-person conception of ourselves as rational agents requires freedom or spontaneity. It turns out that what it requires, even according to Allison, is the *mere conceivability* of freedom, not its *metaphysical reality.* On Allison's view, the solution to the problem of agency turns on the claim that the dependence of our conception of ourselves as rational agents on freedom or spontaneity is conceptual rather than ontological:

> Kant is ... claiming merely that it is necessary to appeal to the transcendental idea of freedom in order to conceive of ourselves as rational (practically free) agents, not that we must actually be free in the transcendental sense in order to be free in the practical sense. (Allison 1990: 57)[16]

This passage suggests that, since the idea of freedom is necessary for our conception of ourselves as rational agents, all we need to establish is that freedom is conceivable, which is different from establishing that freedom is metaphysically possible. A conception of ourselves as free imposes on us a decision-making frame of mind which turns us into genuine rational agents, whether or not we are actually free.[17]

It follows on Allison's interpretation that the problem expressed by Kant in the third antinomy and solved by appeal to the doctrine of transcendental idealism is the problem of the conceivability of freedom. Though this interpretation finds some support in Kant's claim that one has to take into account the distinction between appearances and things in themselves (the doctrine of transcendental idealism) in order to render the representation of freedom 'at least not self-contradictory' (Bxxviii), I believe that the overall evidence is against it, and that the problem expressed by the third antinomy, the problem transcendental idealism purports to solve, is the problem of the metaphysical possibility of freedom, not of its conceivability.

The conceivability of freedom is established by Kant when he shows how pure reason arrives at the transcendental ideas by applying the concept of *the unconditioned* to the pure concepts of the understanding (the categories). In particular, reason arrives at the transcendental idea of freedom as follows: given

any event in the sensible world, it must have a cause; but since this cause must itself have a cause and so on, the only way to attain the totality of conditions or the unconditioned in the causal chain is to postulate a cause which begins a series of events spontaneously or absolutely.

Moreover, in the resolution of the third antinomy, Kant appeals to the doctrine of transcendental idealism not in order to show that the idea of freedom is not contradictory, but in order to show that the thesis and the anti-thesis 'may *both* alike be *true*' (A532/B560).[18] It is this solution that is not available to the empiricist, though arguably the empiricist can acknowledge the 'bare conceivability of freedom', at least in the negative sense, as independence from the causality of nature, while maintaining that it is a concept that cannot be instantiated in experience, 'an empty thought-entity' (A446/B474).[19]

Here are some of the reasons why I believe that Kant's interest was not in the problem of agency and the conceivability of indeterminist freedom, but rather in the problem of accountability and the metaphysical reality of freedom. Thus, if Kant had been interested in the mere conceivability of indeterminist freedom:

(1) he wouldn't have deployed the metaphysical arsenal of transcendental idealism in order to solve the third antinomy, since we don't need transcendental idealism to show that the conceivability of freedom is compatible with causal determinism;
(2) he wouldn't have regarded freedom as the 'stumbling block for all *empiricists*' (*CPrR*, 5:7) since arguably empiricists can accept the conceivability of freedom, or the meaningfulness of the concept of freedom, at least in the negative sense, as independence from natural causality; and
(3) he wouldn't have insisted on a proof of the reality of freedom in the *Groundwor*k and in the second *Critique*, in contrast with the proof of its possibility in the first *Critique*.

If the third antinomy is to be viewed as a conflict between two models of agency, it should be viewed as a conflict that arises from an objective, third-person perspective, between transcendental freedom as a condition of the possibility of morality and determinism as a condition of the possibility of science, between what metaphysically must be the case in order for an all knowing judge to pursue the ideal of justice and what metaphysically must be the case in order for a scientist to pursue the goal of knowledge. If the third antinomy is viewed in this way, Kant has to show nothing less than that the metaphysical possibility of freedom is compatible with the necessity and universality of causal determinism.

In the remainder of this section I will defend my *Anti-thesis I* by arguing that Kant's primary reason to uphold freedom is to provide the metaphysical conditions of the possibility of genuine moral responsibility, the kind of responsibility without which an ideally just judge could not regard us as deserving praise or blame, reward or punishment. This view of the role of freedom in Kant's philosophy fits in a bigger

picture according to which Kant's moral philosophy is one of grandest attempts in Western philosophy to banish any form of luck from the realm of morality.

There are various forms of moral luck[20] that Kant is interested in ruling out, and they undermine the foundation of morality in various ways. First, there is luck in the consequences of one's actions, which Kant rules out with his well known rejection of consequentialism and the adoption of an ethics of motive. It is well known that, according to Kant, the moral assessment of an action is based on its maxim, not on its consequences or the effects achieved in the actual world, which are beyond the agent's control and subject to luck. As he puts it in a famous passage:

> A good will is not good because of what it effects or accomplishes, because of its fitness to attain some proposed end, but only because its volition . . . Even if, by a special disfavor of fortune or by the niggardly provision of a stepmotherly nature, this will should wholly lack the capacity to carry out its purpose—if with its greatest efforts it should yet achieve nothing and only the good will were left . . . then, like a jewel, it would still shine by itself, as something that has its full worth in itself.
>
> Usefulness or fruitlessness can neither add anything to this worth nor take anything away from it. (*GMM*, 4:394)

Second, there is luck in one's circumstances, in 'the kind of problems and situations one faces' (Nagel 1979: 28). For example, take two individuals, *X*, who never lied in his life and never would have lied, regardless of the circumstances, and *Y*, who never lied but would have lied if he could have gained from his lie without endangering his reputation. For Kant there is no doubt that *X* and *Y* differ in their moral worth, and this is how he explains it. The maxim of Y's action follows the hypothetical imperative 'I ought not to lie if I want to keep my reputation' whereas *X*'s maxim follows the categorical imperative 'I ought not to lie even when it would not bring me any discredit'. But an action has genuine moral worth only when it is done from duty, and so *X*'s behavior has genuine moral worth, whereas *Y*'s doesn't.

If circumstantial luck affects our judgements of moral worth, it is only because our knowledge is limited to observing how people act in actual circumstances. Even in our own case, where we are in a better position to know the maxims on which we act, we cannot with absolute certainty establish the moral worth of our actions. As Kant points out, even under the 'keenest self-examination' it is possible for us to have an incorrect idea about our own value since 'when moral worth is at issue, what counts is not actions, which one sees, but those inner principles of action that one does not see' (*GMM*, 4:407). However, from the ideal standpoint of an omniscient judge, people's genuine moral worth and the possibility of perfect justice are not affected by circumstantial luck.[21]

Third, and last, there is constitutive luck, luck 'in how one is determined by antecedent circumstances', including the kind of person that you are by nature and by education, 'your inclinations, capacities, and temperament' (Nagel 1979:

28). If causal determinism were true, then constitutive luck together with other factors would determine the moral choices that we make throughout our life, and *even omniscience could not make perfect justice possible*. Even divine, omniscient justice would be flawed if we were punished for things we did or failed to do because of traits of character which were not in our control. Thus, constitutive luck makes perfect justice impossible. The possibility of genuine responsibility, moral worth, and perfect justice, depends on the elimination of constitutive luck.

According to Kant, all moral agents are equally endowed with two things: the knowledge of moral duty and the power to act from duty regardless of inclinations, temperament, education, and other accidents of personal history. Freedom ensures that, in a situation in which someone knows he ought to do *A*, he can do *A*, by spontaneously beginning a new series of events, regardless of his causal history, and so he is genuinely responsible for his action and the proper object of just reward/punishment, praise/blame.[22]

Bernard Williams powerfully conveys the trend against moral luck in Kant's moral philosophy:

> The successful moral life, removed from considerations of birth, lucky upbringing . . . is presented as a career open not merely to talents, but to a talent which all rational beings necessarily possess in the same degree. Such a conception has an ultimate form of justice at its heart, and this is its allure. Kantianism is only superficially repulsive—despite appearances, it offers an inducement, solace to a sense of the world's unfairness. (Williams 1981: 28)

> [The purity of morality] expresses an ideal, presented by Kant . . . in a form that is the most unqualified and also one of the most moving: the ideal that human existence can be ultimately just. . . . The ideal of morality is a value, moral value, that transcends luck. It must therefore lie beyond any empirical determination. It must lie not only in trying rather than succeeding, since success depends partly on luck, but in *a kind of trying that lies beyond the level at which the capacity to try can itself be a matter of luck*. The value must, further, be supreme. It will be no good if moral value is merely a consolation prize you get if you are not in worldly terms happy or talented or good-humoured or loved. It has to be what ultimately matters. (Williams 1985: 195, my emphasis)

To sum up, on my view, Kant was preoccupied to answer the following questions: what is the metaphysical requirement for genuine moral responsibility and moral worth? What is the condition of the possibility of ideal justice? If there is a God, and if this God punishes us for our wrong deeds and rewards us for our good deeds, what metaphysical property must we have in order for God's punishments and rewards to be just? To all these questions, I take Kant's answer to be, correctly: transcendental freedom or spontaneity.

5. Some Remarks in Defence of the Metaphysical Gap Between Moral and Non-Moral Agency

Allison holds that the scope of transcendental freedom or spontaneity is not restricted to moral agency, but underlies the entire rational agency (*Thesis II*). This follows from the claim that the role of transcendental freedom is to capture our first-person conception of ourselves as rational agents engaged in deliberation and choice (*Thesis I*), insofar as our first-person conception of ourselves as agents is allegedly the same in moral and non-moral situations. Moreover, Allison argues that morality alone could not introduce incompatibilist freedom—which Kant obviously endorses—in a conception of rational agency if it is not already there. He writes that 'if the general argument from the presuppositions of rational agency fails to provide adequate support for an incompatibilist conception of freedom, then no subsequent appeals to specifically moral requirements or ''facts'' could conceivably do the job', since any naturalistic account of our capacity to act on the basis of principles, frame ends or incorporate incentives could be easily extended to our capacity to act from duty alone.[23]

By contrast, on my view, the reasons for attributing incompatibilist freedom to ourselves are strictly related to morality (*Anti-Thesis I*). To clarify, my view presupposes a clear-cut distinction between moral and non-moral agency, according to which 'moral agency' refers to decisions made in those situations in which the agent is aware that he has a moral duty to do *A*, whereas 'non-moral agency' refers to decisions made in those situations in which, justifiably or not, the agent is not aware of having a moral duty to act one way or another. Thus, it is only in the case of moral agency that our awareness of the categorical nature of the moral law forces us to conclude that we have the power to obey it unconditionally: we conclude that we can do *A*—in the absolute, unconditional sense—simply because we know that we ought to do it. This constitutes the ground for attributing incompatibilist freedom to ourselves, as well as for holding ourselves (and others) morally responsible. (As I will argue later in this section, no similar argument can be made for the non-moral case.)

It follows on my view that the distinction between moral and non-moral agency is drawn from the agent's point of view and is based on his perception of his situation as requiring moral attention or not. But how does the agent know when his situation requires moral attention? Is he supposed to raise the question of moral permissibility for any action he plans to perform?[24] Here I will follow Barbara Herman who, in her 1993 book *The Practice of Moral Judgment*, points out that 'normal moral agents do not question the moral permissibility of everything they propose to do' and that we 'expect moral agents to have acquired knowledge of the sorts of actions that it is generally not permissible to do and the sorts of actions that, in the normal course of things, have no moral import' (Herman 1993: 76). Herman points out that, since one and the same action can be described in indefinitely many different ways, the agent 'who came to the [Categorical Imperative] procedure with no knowledge of the moral characteristics of actions would be very unlikely to describe his action in a morally

appropriate way' (Herman 1993: 75). Thus, we have to assume that all competent moral agents possess certain moral knowledge, which Herman calls 'knowledge of rules of moral salience' (RMS) and describes as follows:

> Acquired as elements in a moral education, [RMS] structure an agent's perception of his situation so that what he perceives is a world with moral features. They enable him to pick out those elements of his circumstances or of his proposed actions that require moral attention. (Herman 1993: 77)

Two consequences of my view deserve a brief mention. First, since the distinction between moral and non-moral agency is based on the agent's knowledge of RMS and his perception of his circumstances, there will be situations in which the agent, failing to realize that his circumstances require moral attention, and unaware of his duty to do *A*, chooses to do *B* instead for prudential reasons. Since his is a non-moral choice, we cannot ascribe freedom to him, and therefore we cannot hold him morally accountable, though everyone agrees that what he did was wrong. I regard this consequence of my view as entirely acceptable. For example, I do not believe that, as long as nobody knew that second-hand smoking causes cancer, one was morally responsible and blamable for smoking in the presence of one's young children. Only since we have become aware of the link between second-hand smoking and cancer, have we become morally responsible. By the same token, I do not believe that someone who still does not know that second-hand smoking causes cancer is morally responsible and blameable for smoking in the presence of his young children. It is true that we are reluctant to absolve people of moral responsibility in situations in which we think they should have known that *B* was morally wrong. However, I would argue that our reluctance is due to a justifiable doubt regarding their sincerity—especially when the wrongness of *B* in the given circumstances is common knowledge. Since we are not omniscient, we have to hold other people accountable by appeal to some standard of what the competent moral agent is expected to know. If we could only *know* that the agent was not aware that doing *B* was wrong, then I think we would not hold him morally accountable (just as he should not hold himself morally accountable, though he may be distraught by the consequences of his action).[25]

Another consequence of my view is that, as our empirical knowledge and/or 'knowledge of rules of moral salience' expand, so does our moral agency, in the sense that we will regard more and more of our decisions as moral decisions—decisions in which we exercise incompatibilist freedom and are morally responsible. I think this consequence of my view is also correct. For example, reading about the suffering of farm-raised animals at some point in the past made me realize that eating meat is morally wrong. Since that realization, every time I choose what to eat (at least as long as I am still tempted by a juicy steak) I am making a moral choice, where I was previously making a non-moral choice.[26]

Though there are many naturalist compatibilist accounts to choose from in order to explain our agency and responsibility in non-moral or prudential situations, one may wonder if we really have any good reasons to believe that we possess two distinct kinds of agency and two distinct kinds of responsibility, requiring two distinct theoretical accounts, incompatibilist and compatibilist. Thus, it can be argued that, though the grounds for ascribing incompatibilist freedom to ourselves come from our awareness of the moral law, once we ascribe freedom to ourselves we should hold that it is exercised across the board in all instances of rational agency.[27] In response, in the remainder of this section I will try to make several points to the effect that we don't have any good reasons to believe that we exercise spontaneity in our non-moral choices; the reasons we have to attribute incompatibilist freedom to ourselves are restricted to moral agency (*Anti-thesis II*). And while theoretical unification may be a goal worth pursuing, I do not believe that it can stand on its own as the only reason for endorsing a unified account of rational agency.

I will begin by pointing out that, regarding the problem of luck, we do acknowledge a distinction between moral and non-moral situations. In a non-moral situation, where we have to make a purely prudential choice, there is no need to rule out luck in order to establish genuine responsibility and the possibility of perfect justice. Moreover, nobody attempts to entirely rule out luck from human affairs, and the fact that we claim responsibility for our prudential decisions does not necessarily mean that prudential responsibility is of the same kind as moral responsibility. In fact, there may be a phenomenological distinction between two kinds of responsibility, moral and prudential, a distinction alluded to by Kant when he writes:

> He who has *lost* at play can indeed be *chagrined* with himself and his imprudence; but if he is conscious of having *cheated* at play (although he has gained by it) he must *despise* himself as soon as he compares himself with the moral law. (*CPrR*, 5:37)

Prudential responsibility and the feelings associated with it (shame, regret, disappointment, annoyance) are future-oriented, whereas moral responsibility and the feelings associated with it (guilt, contempt, remorse) are characterized by a certain fixation on the past. Kant accurately describes the peculiar feeling that accompanies the belief in our own moral responsibility, and admits that this feeling of repentance and its fixation on a deed that cannot be undone would be absurd[28] if we did not connect it with our power to act from duty, which is not influenced by our causal history and, in this sense, can be considered timeless:

> This [view of rational beings as free] is also the ground of repentance for a deed long past at every recollection of it, a painful feeling aroused by the moral disposition, which is empty in a practical way to the extent that it cannot serve to undo what has been done and it would even be absurd . . . but repentance, as pain, is still quite legitimate because reason, when it is a question of the law of our intelligible existence (the moral law),

> recognizes no distinction of time and asks only whether the event belongs to me as a deed and, if it does, then always connects the same feeling with it morally, whether it was done just now or long ago. (*CPrR*, 5:98)

Related to the distinction between prudential and moral responsibility, there is also Kant's well-known contrast between, on the one hand, features and activities which we value for reasons unrelated to morality (humour, strength, great natural talents, and activity 'proportioned to them') and the feelings they give rise to (love, fear, admiration), and, on the other hand, moral worth and the feelings it gives rise to (respect, reverence).[29]

An attractive but ultimately too strong argument for my *Anti-thesis II* and against the attribution of spontaneity to purely prudential choices can be made to the effect that the attribution of spontaneity in prudential choices amounts to the introduction of an element of arbitrariness or lawlessness in our rational agency. The idea is that, while in moral situation we have to choose between the principle of self-love and the principle of morality, in purely prudential situations our choice stands entirely under the principle of self-love. It makes sense to claim that spontaneity is involved in the choice between the principle of self-love and the principle of morality: placed by our sensuous nature and as members of the natural world under the principle of self-love (passive role), we have the power to re-place ourselves under a different law, the moral law (active role, spontaneity).[30] But claiming that spontaneity is involved in a prudential choice, where we do not have two distinct principles of action to choose from, amounts to arbitrarily disconnecting the agent's will from his own desires and inclinations. But there is no room for arbitrariness either in a first-person or a third-person conception of agency: when someone appears to act without any basis in his inclinations or desires, we assume he has formed a desire to prove to himself that he is free, a desire to get out of his own skin, or 'to act crazy'. This desire keeps the agent caught in the web of natural causality even as he is trying to escape it.

The above argument attempts to prove that spontaneity cannot intervene in purely prudential choices because it would make them arbitrary. However, there are prudential choices which are very similar to moral choices, in that we have to choose between a reflectively endorsed conception of our own long-term self-interest or happiness and a first-order desire or inclination, the satisfaction of which conflicts with that conception. Consider the following situation: I have to choose between continuing to play basketball with my friends, which is the thing I have the strongest inclination to do, and leaving right now in order to be on time for an important meeting, which I know is the thing I have the strongest prudential reason to do. In this case, one may argue, to attribute spontaneity to myself is to attribute to myself the power to act in accordance with the strongest prudential reason, regardless of my immediate inclination.

This prudential situation clearly parallels the moral situations in which I attribute to myself the power to act in accordance with the moral law, regardless

of all inclinations. If it makes sense and it is non-arbitrary to exercise transcendental freedom or spontaneity in moral situations, then it must be similarly non-arbitrary to exercise spontaneity in (at least some) prudential situations. Thus, I will not attempt to argue that we have a good reason to believe that we do not exercise spontaneity in our non-moral choices. However, I will argue that we don't have any good reason to believe that we exercise spontaneity in our non-moral choices.

What is our reason to believe that we exercise spontaneity in our moral choices? If my discussion in Sections (3) and (4) is on the right track, then the reason is *not* that the feature of reflectivity frees our mind from the direct control by inclinations, and enables us to choose between determining grounds of action that are instinctive and determining grounds of action that are 'thought by reason'. The reason why we believe that we exercise spontaneity in our moral choices is that *the moral law commands categorically* and so, if we fail to do the right thing, we know we could have done otherwise, in the exact same circumstances. Because the moral law commands categorically, we come to believe that we have the power to obey it unconditionally.

It may be thought that the same reasoning applies to the above-mentioned prudential situation. Let us assume that, though I have a clear understanding of the consequences of my action, I follow the strongest inclination and continue to play basketball with my friends instead of leaving in order to be on time for an important meeting. I may even say to myself 'I should leave right now (no matter what!)' and, failing to do so, I will later think that *I should have done otherwise* and, therefore, that *I could have done otherwise* (in the exact same circumstances). Doesn't this mean we have a reason to believe that we can exercise spontaneity in our prudential choices? Let me explain why I don't think this is the case.

Unlike the moral law, which commands categorically, prudential imperatives command only hypothetically, and so the conclusion of my deliberation over continuing to play vs. leaving for the meeting is not 'I should leave right now (no matter what!)', but 'I should leave right now if I want to do the prudent thing'. If I fail to do the prudent thing, I must conclude not that *I could have done otherwise* (in the exact same circumstances), but that *I could have done otherwise if I had wanted to do the prudent thing*. However—someone may press further—*I should have wanted to do the prudent thing* (no matter what!), therefore *I could have done otherwise*.

What we have here is an attempt to establish that there is a categorical prudential imperative, commanding us that we should always be prudent or act according to our long-term best interest, and I think it is bound to fail. Prudence is not a goal prescribed universally and necessarily to all human beings, but one among many goals that someone may have, according to her conception of what makes life worth living. Some people may sacrifice (what most of us would regard as) prudence for the sake of other values, such as following one's heart, living dangerously, or just exercising the right to mess up one's life in one's own unique way. If an action is criticized as a failure of prudence, this is not necessarily a final criticism, since prudential rationality is only one among many

features in terms of which human beings define and value themselves. By contrast, criticizing an act as morally wrong is, if justified, a final criticism.

To sum up, from the fact that *I should have done otherwise*, namely, that I should have acted according to my long-term best interest, I cannot conclude that *I could have done otherwise* in the absolute, categorical sense in which I mean it in the moral context. The conclusion is rather that *I could have done otherwise if I had wanted to do the prudent thing*, and so ultimately I have no reason to believe that I exercise spontaneity in my prudential choices.

But then what is the significance of my insisting now that *I should have left right then*, no matter what, no ifs and buts? What is the force of this claim, if in fact *I couldn't have done otherwise*? It is a future-oriented force, one that will affect my priorities and the strengths of my desires from now on, not a backward-looking one, like in moral imputation. What matters, if indeed I conclude that *I should have done otherwise*, is not that *I could have done otherwise*, but that in the future *I will do otherwise*. This contrasts with the moral case, in which what matters is that *I could have done otherwise*, since this is the ground of the action's imputation, the condition of the possibility of just blame or punishment.

Let me close my argument with a famous passage in which Kant discusses the different meanings of 'I could have done otherwise' in prudential and moral contexts:

> Suppose someone asserts of his lustful inclination that, when the desired object and the opportunity are present, it is quite irresistible to him; ask him whether, if a gallows were erected in front of the house where he finds this opportunity and he would be hanged on it immediately after gratifying his lust, he would not then control his inclination. One need not conjecture very long what he would reply. But ask him whether, if his prince demanded, on pain of the same immediate execution, that he give false testimony against an honorable man whom the prince would like to destroy under a plausible pretext, he would consider it possible to overcome his love of life, however great it may be. He may perhaps not venture to assert whether he would do it or not, but he must admit without hesitation that it would be possible for him. He judges, therefore, that he can do something because he is aware that he ought to do it and cognizes freedom within him, which, without the moral law, would have remained unknown to him. (*CPrR*, 5:30)

The way I read it, the passage emphasizes the following difference between prudential and moral choices. In the prudential situation, when the lustful man claims he couldn't have done otherwise, we can show him that he could have done otherwise by appealing to an alternative scenario in which his desire for immediate gratification is outweighed by the desire to stay alive. *In the prudential situation, in order to show someone that he could have acted prudently we have to point to different circumstances in which he would have acted prudently.* But then the conclusion is not that he could have done otherwise, but that he would have done

otherwise if the circumstances had been different. We cannot show him that he could have done otherwise in the exact same circumstances.[31] By contrast, in the moral situation, regardless of what he actually does, a person knows that he can do his moral duty just because the moral law commands categorically. Even as he fails to do his duty, he knows he could have done otherwise, not by contemplating different circumstances in which his desires would have been different, but because of the categorical demand that the moral law makes on him.

Though the idea of a metaphysical gap between moral and non-moral agency is prima facie implausible, on closer inspection we notice that a metaphysical gap between nature and morality is present in our ordinary thinking: we view life as fundamentally unfair, yet we attempt to pursue justice and fairness, we acknowledge that individuals are fundamentally unequal, endowed with different talents and facing different challenges, yet we claim that morally we are all equal.

In this paper I have argued that the primary role of transcendental freedom in Kant's philosophy is to rule out moral luck, in order to make possible genuine moral responsibility, moral worth, and perfect justice. For this, Kant has to show not only that transcendental freedom is conceivable, but that it is metaphysically real. I have argued against Allison's claim that transcendental freedom is the defining feature of our conception of ourselves as rational agents. Instead, I have claimed that our conception of ourselves as rational agents, expressed by Kant in his conception of the *arbitrium liberum*, is neutral with respect to causal determinism, and that the only reason we have to believe that we are free in the transcendental sense is that the moral law commands us categorically. It follows that we do not have any reason to believe that we are free in purely prudential choices, because prudential imperatives command only hypothetically.

I do not think it can be shown that we are free in our moral choices and not free in our prudential choices. I have argued instead that we have a reason to believe that we are free in our moral choices and have no similar reason to believe we are free in our prudential choices.[32]

Iuliana Corina Vaida
Department of Philosophy
The College of William and Mary
USA
icvaid@wm.edu

NOTES

[1] In the *Critique of Pure Reason*, in the third antinomy, Kant defines transcendental freedom as 'a power of absolutely beginning a state', 'an *absolute spontaneity* of the cause', and 'freedom (independence) from the laws of nature' (A444–7/B472–5). Transcendental freedom is only part of the content of the concept of freedom which is relevant to morality. However, it is the part which, because of its explicit incompatibility with causal

determinism, constitutes 'the real stumbling block' for philosophy (A448/B476). In all parenthetical references, both in the main text and in the notes, '*CPrR*' refers to Kant 1997a, '*GMM*' to Kant 1997b, '*A*/*B*' to Kant 1965, and '*RWLRA*' to Kant 1960.

[2] It could then be argued that the problem of accountability gives rise to a problem of agency: we have to make sure that the attribution of transcendental freedom to ourselves does not conflict with our first-person perspective of ourselves as rational agents.

[3] See especially Allison 1990 and Korsgaard 1996c. This view is also held outside Kantian scholarship, among the relatively few contemporary philosophers who are not satisfied with compatibilist accounts of agency and accountability. See for example Nagel 1986.

[4] In this paper, I will use the term 'moral' in the broad sense in which it is contrasted with 'non-moral' or 'morally neutral' or 'prudential', rather than 'immoral' or 'morally wrong'. Thus, I take 'moral agency' to be referring to decisions made in those situations in which we are aware that we have a moral duty to do *A* (or to refrain from doing *A*). By contrast, 'non-moral agency' refers to decisions made in those situations in which, justifiably or not, we are not aware of having a moral duty to act one way or another. I will say more about this distinction at the beginning of Section (5).

[5] However, a thorough defence of Kantian incompatibilism against competing compatibilist conceptions of agency is beyond the scope of this paper.

[6] Allison states that the Incorporation Thesis is not an empirical claim, and is not 'based on introspection of what goes on in our heads when we decide; so we cannot, as it were, catch ourselves in the act of incorporating' (Allison 1996a: 118).

[7] Vaida 2009 argues that, *pace* Kant, the metaphysical doctrine of transcendental idealism is not essential to a Kantian solution to the third antinomy and to the free will problem. The solution proposed—relying on a realist interpretation of Kant's transcendental theory of experience and the thesis of epistemic modesty it suggests—holds that, while a causally determined event cannot be free, the necessity and universality of our deterministic claims should not be understood as absolute, objective features of reality, but as relative, or restricted to the objects of possible experience. Therefore, they do not entail that free choices cannot exist, but only that they cannot constitute objects of possible experience. This Kantian solution to the free will problem has the advantage of being compatible with the robust form of scientific realism according to which the entities posited by our most advanced scientific theories, in particular spatio-temporal entities governed by causal deterministic laws, exist *independently of the human mind and its forms of cognition*.

[8] These ends don't have to be moral ends: they may be ends we set for ourselves in accordance with what we think would make us happy or what we take to be in our self-interest. Allison argues that the fact that self-interest is an end set through nature rather than through free choice, so that we cannot help but pursue it, is compatible with the conception of rational agent as a spontaneous framer of ends, because 'it is we who determine how this [self-interest] is to be defined' (Allison 1996a: 111–14).

[9] See, for example, Ameriks 1992 and Guyer 1992. In what follows, I will, like Guyer, question Allison's claim that 'the incorporation thesis implies the spontaneity thesis, that is, that action on principles and never merely *per stimulus* alone requires transcendental freedom' (Guyer, 1992: 102).

[10] It is true that in the *Dialectic* Kant claims that the *arbitrium liberum* is 'freedom in the practical sense' or practical freedom, for short (A533/B562–A534/B563), and that practical freedom is based on the transcendental idea of freedom, so that '[t]he denial of transcendental freedom must . . . involve the elimination of all practical freedom' (A534/

B562). These claims lend some support to Allison's view that the *arbitrium liberum* is based on transcendental freedom or spontaneity, and thus it is an incompatibilist conception of agency. However, the relation between practical and transcendental freedom in the *Critique of Pure Reason* is notoriously problematic, since Kant says conflicting things in the *Dialectic* and in the *Canon*. As far as I know, there is no interpretation of the *Dialectic* and *Canon* passages, Allison's included, which does not hold that Kant is contradicting himself at some point, or is at least guilty of sloppy usage of his own terminology.

[11] We could have a sophisticated belief-desire model of agency according to which, instead of reasons or second-order desires coming with 'pre-assigned weights', the agent bestows weights upon his reasons in the very act of (spontaneous) choice. For such a model, see for example Nozick 1995: 101–14.

[12] Korsgaard claims that her account of obligation, based (in part) on the fact of reflective consciousness, is a form of naturalism and is consistent with the Scientific World View (Korsgaard 1996d: 160).

[13] The assumption is roughly that our desires, including higher order desires, and their relative strengths, have a deterministic causal history, and that deliberation, the endorsement and rejection of desires in the light of higher order desires, takes place in accordance with natural causal laws.

[14] Though there are difficulties with this definition, here I take that an agent is genuinely responsible for an action only if in exactly the same circumstances he could have chosen otherwise. This definition follows Kant's famous *dictum* that 'ought' implies an absolute, unconditional 'can', and in particular that, if the agent ought to have done otherwise, he could have done otherwise, even if the circumstances immediately preceding his action, both psychological and external, had been exactly the same as they actually were. The various naturalist compatibilist views of responsibility fail to capture the commonsense conception of responsibility which, in our moral practices, is closely tied to the ideas of justice, just praise/blame, reward/punishment, and thus reveals itself to be an incompatibilist conception.

[15] I'm not sure if all people share the same experience of making decisions, but my first-person experience, as well as my understanding of myself as an agent engaged in deliberation and choice, are that, after gathering as much information as I can, after considering all the relevant desires, goals, or values, their practical compatibility or incompatibility, *I wait for the decision to occur to me*. In order to see myself as the (rational) author of my decision, I don't have to regard myself as arriving at it by means of an argument, because an argument would require me to know the relative strengths of my desires and values, which I don't. Moreover, in order to see myself as the author of my decision, I don't have to regard myself as performing a spontaneous act of endorsement of a certain desire: it is enough that the decision follows causally from *desires and values which I recognize as truly mine*, according to their relative strengths, which characterize me even though I am not fully aware of them.

[16] Allison 1996a makes the same point, when he writes that the Incorporation Thesis is not 'a straightforward metaphysical thesis about what is really going on in the noumenal world', but 'rather a conceptual claim' (Allison 1996a: 118).

[17] However, Allison 1996a remarks that if in fact we were not free but determined, this would amount to our being 'merely complex mechanisms rather than *genuine* (spontaneous) rational agents' (Allison 1996a: 124, my emphasis). In other words, if freedom were in fact an illusion, rational agency in general and moral agency in particular would also be illusory. But then the dependence between the two would no longer be conceptual, but ontological, wouldn't it? I'm at a loss here trying to reconcile Allison's apparently conflicting remarks.

[18] It is important here to remember that, in Kant's view, the resolutions of the third and fourth antinomies are quite different from the resolutions of the first and second antinomies. For example, regarding the first antinomy, Kant argues that the transcendental idea of *the world as a spatio-temporal whole* is a contradictory idea or concept, and therefore both thesis and anti-thesis are false, because they have a contradictory concept as their subject. But in the case of the third antinomy, Kant does not limit himself to pointing out that the concept of transcendental freedom is not contradictory, but argues that both the thesis and the anti-thesis may be true.

[19] One could argue that a strict empiricist like Hume would reject even the 'bare conceivability' of transcendental freedom and that in fact Hume explicitly rejected it when he argued that the liberty of indifference originates in a false sensation or experience (Hume 1955: 103, note 7). However, it is not clear that empiricism can (or should) uphold Hume's strict standard of meaningfulness. If *God* is conceivable, then *transcendental freedom* is too—if only negatively, as independence from the causality of nature. The stumbling block for the empiricist is not the conceivability of freedom, but its reality.

[20] My classification roughly follows the one proposed in Nagel 1979.

[21] Of course, in the Christian tradition it is God who plays the role of the omniscient judge. It is only from a being who sees all, knows all, that we have any chance to receive perfect or ideal justice. But even if one were agnostic or did not believe in God, it would still be interesting to inquire into the conditions of the possibility of perfect justice.

[22] Ruling out constitutive luck also requires that morality is based on reason rather than on moral sentiment. If morality were based on the moral sentiment, as Hutcheson and Hume held, our capacity to obey the moral law would depend on our natural inclinations and we couldn't be held responsible for it. At some point in the *Groundwork*, Kant is asking us to compare a person whose soul is 'so sympathetically attuned that, without any other motive of vanity or self-interest they find an inner satisfaction in spreading joy around them' with someone in whose heart 'nature had put little sympathy', who is 'by temperament cold and indifferent to the sufferings of others' (*GMM*, 4:398). If morality were based on moral feeling, the latter person would be off a very unfair start. And one may wonder if it would be fair or just to hold him to the same standards of moral behaviour as the first one. But if morality is based or reason, then the indifferent man has as much access to the requirements of morality as the sympathetically inclined man. Moreover, freedom makes it possible for him to reject his nature and 'find within himself a source from which to give himself a far higher worth than what a mere good-nature temperament might have' (*GMM*, 4:398). There are other passages in the *Groundwork* and in the second *Critique* in which Kant argues against a morality based on feeling. For example, he argues, that feelings 'which by nature differ infinitely from one another in degree', cannot 'furnish a uniform standard of good and evil, and one cannot judge validly for others by means of one's feelings' (*GMM*, 4:442). I take Kant to mean that, even if a uniform standard could be furnished, it would not be fair, whereas a fair standard would not be uniform.

[23] Allison 1996a: 125. According to Allison, what distinguishes moral from prudential choices is not spontaneity, but autonomy. By autonomy Allison understands the capacity to determine oneself independently of one's needs as a sensuous being—as when an agent performs an action against all his sensuous inclinations, solely out of respect for the moral law.

[24] I thank the anonymous referee of this paper for urging me to clarify my view on the distinction between moral and non-moral agency. The referee has also raised some

legitimate concerns regarding the consequences of my view—concerns that I do my best to address in what follows.

[25] A critic may argue that this is not quite the end of the story: even when an agent is not blameable for doing *B*—since he was not aware that it was wrong—he may still be blameable for his ignorance. My account can acknowledge that an agent may be blameable for his ignorance—though only if at some point in the past the agent made a conscious decision to forgo information that came his way or to turn down opportunities to expand his knowledge, while knowing this was wrong (since we have a moral duty to work on expanding our knowledge). If this is unlikely, then the agent may not be blameable for his ignorance. An agent truly lacking in the relevant knowledge (including knowledge of RMS) may not be regarded as morally responsible in a specific situation; in some cases, he may not even be regarded as a fully developed moral agent (and therefore responsible for his actions).

[26] Though it may seem desirable to work on ever expanding our moral agency, I do not think we should go as far as placing a moral requirement on all our actions, thus incurring the criticism that Bernard Williams famously mounted against Utilitarianism—that it is too demanding and threatens our integrity as persons by not respecting our attachments to morally neutral projects. I think it is a valuable trait of Kant's ethics that it can be interpreted, along the lines proposed by Barbara Herman, as a 'morality of limits', which merely places categorical limiting conditions on the choices that we make in the pursuit of happiness. By contrast, Kantian ethics as a 'morality without limits', as put forward by Christine Korsgaard, is, I believe, neither a correct account of the role of morality in our lives nor an account of what is desirable for us to pursue.

[27] I thank the referee of this paper for pressing this point.

[28] The strategy of the compatibilist is precisely to reject such feelings as absurd or irrational.

[29] See the *Groundwork* (4:393–4) and the *Critique of Practical Reason* (5:76–8).

[30] According to Kant, each cause must have a law of its causality: freedom, which is initially defined as 'absolute spontaneity' and independence from natural causality, is later re-defined as causality in accordance with the moral law.

[31] But what is the point of contemplating this alternative scenario (in which one would be hanged immediately after gratifying one's lust), if in the actual situation the person indeed could not have done otherwise? Possibly, the point is to show him how he can control his 'lust' in the future: by vividly picturing (or even exaggerating) the unpleasant consequences of satisfying his inclination.

[32] One reason to extend the scope of transcendental freedom or spontaneity from moral to non-moral agency is to have a unified account of rational agency. However, I do not believe that unity or simplicity by itself, unsupported by other considerations, such as evidence or explanatory power, constitutes a good enough reason to prefer one account over another.

REFERENCES

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

—— (1996a), 'Kant on Freedom: A Reply to My Critics', in *Idealism and Freedom: Essays on Kant's Theoretical and Practical Philosophy*. Cambridge: Cambridge University Press.

—— (1996b), 'Autonomy and Spontaneity in Kant's Conception of Self', in *Idealism and Freedom: Essays on Kant's Theoretical and Practical Philosophy*. Cambridge: Cambridge University Press.

Ameriks, K. (1992), 'Kant and Hegel on Freedom: Two New Interpretations', *Inquiry*, 35: 219–32.
Bennett, J. (1974), *Kant's Dialectic*. Cambridge: Cambridge University Press.
Frankfurt, H. (1997), 'Freedom of the Will and the Concept of a Person', in D. Pereboom (ed.), *Free Will*. Indianapolis/Cambridge: Hackett Publishing Company.
Guyer, P. (1992), 'Book Review', *Journal of Philosophy*, 89: 99–110.
Herman, B. (1993), 'The Practice of Moral Judgment', in *The Practice of Moral Judgment*. Cambridge, MA: Harvard University Press.
Hume, D. (1955), *An Inquiry Concerning Human Understanding*, ed. C. W. Hendel. Indianapolis and New York: The Bobbs-Merrill Company.
Kant, I. (1960), *Religion within the Limits of Reason Alone*, trans. T. M. Greene and H. H. Hudson. New York: Harper & Row.
—— (1965), *Critique of Pure Reason*, trans. N. K. Smith. New York: St. Martin's Press.
—— (1997a), *Critique of Practical Reason*, trans. M. Gregor. Cambridge: Cambridge University Press.
—— (1997b), *Groundwork of the Metaphysics of Morals*, trans. M. Gregor, Cambridge: Cambridge University Press.
Korsgaard, C. (1996a), 'Morality as Freedom', in *Creating the Kingdom of Ends*. Cambridge: Cambridge University Press.
—— (1996b), 'Creating the Kingdom of Ends: Reciprocity and Responsibility in Personal Relations', in *Creating the Kingdom of Ends*. Cambridge: Cambridge University Press.
—— (1996c), *Creating the Kingdom of Ends*. Cambridge: Cambridge University Press.
—— (1996d), *The Sources of Normativity*. Cambridge: Cambridge University Press.
Nagel, T. (1979), 'Moral Luck', in *Mortal Questions*. Cambridge: Cambridge University Press.
—— (1986), *The View from Nowhere*. Oxford: Oxford University Press.
Nozick, Robert (1995), 'Choice and Indeterminism', in T. O'Connor (ed.), *Agents, Causes, and Events*. Oxford: Oxford University Press.
Vaida, I. C. (2009), 'A New Kantian Solution to the Third Antinomy of Pure Reason and to the Free Will Problem', *Southern Journal of Philosophy*, 47: 403–31.
Williams, B. (1981), 'Moral Luck', in *Moral Luck. Philosophical Papers 1973–1980*. Cambridge: Cambridge University Press.
—— (1985), *Ethics and the Limits of Philosophy*. Cambridge, MA: Harvard University Press.