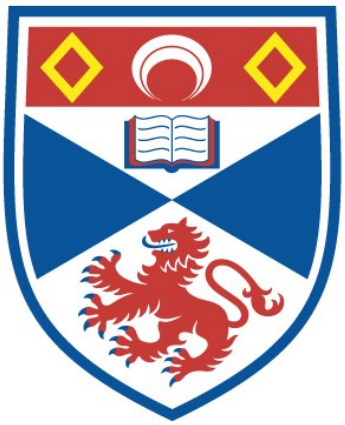
KANT'S CONCEPT OF THE GOOD

Bas Ben Martien Tönissen

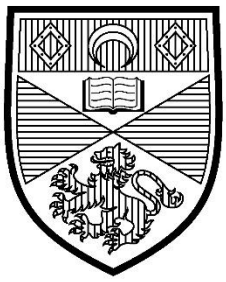
A Thesis Submitted for the Degree of MPhil
at the
University of St Andrews



2018

Kant's concept of the good

Bas Tönissen



University of
St Andrews

This thesis is submitted in partial fulfilment for the degree of MPhil

at the

University of St Andrews

September 24, 2017

Candidate's declaration

I, Bas Ben Martien Tonissen, do hereby certify that this thesis, submitted for the degree of MPhil, which is approximately 35,000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree.

I was admitted as a research student at the University of St Andrews in September 2015.

I confirm that no funding was received for this work.

Date 20 February 2018 Signature of candidate

Supervisor's declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of MPhil in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date 22 February 2018 Signature of supervisor

Permission for publication

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Bas Ben Martien Tonissen, confirm that my thesis does not contain any third-party material that requires copyright clearance.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

Printed copy

No embargo on print copy.

Electronic copy

No embargo on electronic copy.

Date 20 February 2018 Signature of candidate

Date 22 February 2018 Signature of supervisor

Underpinning Research Data or Digital Outputs

Candidate's declaration

I, Bas Ben Martien Tonissen, hereby certify that no requirements to deposit original research data or digital outputs apply to this thesis and that, where appropriate, secondary data used have been referenced in the full text of my thesis.

Date 20 February 2018 Signature of candidate

Abstract – Kant's concept of the good

This dissertation asks what Kant means when he talks about the good, and what role this concept plays in his ethical theory. It is divided in three chapters. The first examines the context in which this question was first asked, namely in a review of Kant's *Groundwork* by H.A. Pistorius. I analyse this review and Kant's direct response to it in his *Critique of Practical Reason*, where he clarifies that the good is the 'necessary object of the faculty of desire' and that it can only be determined 'after and by means of' the moral law. I argue that traditional law-first and good-first readings of these passages both fail, and that instead we should prefer Stephen Engstrom's reading which takes the good as an a priori concept of practical reason, whose content is determined by the moral law. The remaining chapters investigate this view, and specifically the strong guise of the good thesis to which it commits Kant. In the second I clarify the guise of the good and the specific version of it to which Engstrom's view is committed, which is one that holds all willing to aim at satisfying a condition of universal validity. I argue that self-conceit and despondency are two notions from Kant's psychology which provide a model for how non-moral willing can aim at universal validity. In the third chapter, I use this general framework to try and explain specific cases of non-moral willing. I find that the framework can adequately explain away diabolical willing as mere evil willing. It can also deal with frailty, though this requires departure from Engstrom's and Reath's views and the introduction of 'persistent illusion'. It has the same trouble dealing with listlessness that all Kantian views do. I conclude that Engstrom's view of the good is viable.

Table of contents

Acknowledgements

Insofar as the efforts of this dissertation are not entirely in vain, a great many people beside the author bear responsibility for that. First among them is undoubtedly my first supervisor Jens Timmermann, who has been a model of both philosophical mentorship and academic excellence and to whom I am forever grateful for all the doors he has opened for me. I also want to thank my second supervisor Justin Snedegar, whose advice – though I should have sought it out much more than I have – has tremendously benefitted this dissertation and me personally. I am further grateful to have enjoyed excellent feedback, conversation and advice from Andrews Reath and Stephen Engstrom; my annual reviewers, Ben Sachs and Alex Douglas; Michael Walschots and all the other active members of the St. Andrews Kant Colloquium, Lucas Sierra Velez, Stefano Lo Re, Kristina Kersa, Janis Schaab and Leslie Stevenson; Kate Moran; Martin Sticker; and from audiences at the 2017 Multilateral Kant Colloquium in Halle, the 2017 Annual Conference of the UK Kant Society in St. Andrews, and the St. Andrews Friday Graduate Seminar. Last but not least, I thank Nastia Grishkova for the boundless interest and patience she showed as I discussed my work with her, her sound philosophical advice and all the love and support she provided throughout the process of writing.

Introduction

> "This, then, is our first question: What is good? and What is bad? and to the discussion of this question (or these questions) I give the name of Ethics, since that science must, at all events, include it."[1]

In his opening chapter to *Principia Ethica*, 'The Subject-Matter of Ethics', G.E. Moore introduces ethics as a discipline concerned with the question of good conduct. It is evident to him, however, that this cannot be its most basic question. We cannot understand the complex notion 'good conduct' without a prior grasp of its most salient concept, 'good'. This motivates the rest of his project in the book, which is to give a conceptual analysis of the term 'good' and determine the kinds of objects that can fall under it. Moore's framing of the discipline is extremely intuitive and, I would conjecture, is even something of a default position. Surely we should engage in good conduct *because* it is good, and thus the task of ethics is to discover what goodness itself is. Pre-reflectively, it can be difficult even to make sense of an alternative position (besides an amoralist one); and surely the burden is on the proponent of such a position to demonstrate that their approach can make sense.

For a long time, however, Kant's ethics has been read as standing outside of this tradition. Instead of the good, the primary notions in his ethics are those of duty and moral law. When he does talk of the good, he often seems to want to define it reductively in terms of the moral law. In the *Religion,* for instance, "morally good" is explained elliptically as "agreement of the power of choice with the law".[2] In the *Groundwork* it is similarly made clear that a will is made good by its maxims having universalizability, i.e. by agreement with the moral law.[3] And in the *Critique of Practical Reason* he

[1] G.E. Moore, *Principia Ethica* (Cambridge: Cambridge University Press, 1971), 3
[2] Immanuel Kant, *Religion within the Bounds of Bare Reason*, trans. Werner S. Pluhar (Cambridge: Hackett Publishing Company), VI:161f. Henceforth *Religion*.
[3] Kant, *Groundwork to the Metaphysics of Morals*, trans. and ed. Mary Gregor and Jens Timmermann (Cambridge: Cambridge University Press, 2011), 4:437. Henceforth *GMS*.

seems very explicitly to argue for the priority of the moral law over the good.[4] The good has consequently received comparatively little attention in many of the seminal texts on Kant's moral philosophy. For a long time Kant counted as the primary representative of a 'deontological' tradition in ethics: one in which the notion of 'the right' takes priority over that of 'the good.'

This position is difficult to make sense of, primarily because it is hard to see how it can give a satisfactory rationale for moral obligation.[5] The deontological position can be easily caricatured as holding that we should follow the rules, simply because they are the rules – and rules are to be followed! Surely such a theory is unsatisfying without an explanation of why these rules in particular are the rules. However, we would expect such an explanation to bottom out in some relation these rules have to a good which they promote. Deontology rules out such an explanation by hypothesis, and may therefore leave us thinking of Kant's practical imperatives as, in Schopenhauer's phrase, "fallen from heaven"[6]. It is therefore unsurprising that the deontological understanding of Kant's ethics did not fare well in the twentieth-century metaethical landscape that Moore helped shape, and that many contemporary scholars sympathetic to Kant have since abandoned it. Barbara Herman begins her tellingly titled 'Leaving Deontology Behind' with the bold claim that "Kant's project in ethics is to provide a correct analysis of 'the Good', understood as the ultimate determining ground of all action."[7] Jens Timmermann has further argued that "Kant should not be tarred with the brush of deontology, because it threatens to obscure the most distinctive features of his theory".[8] Timmermann understands Kant as offering a theory of moral goodness, with *autonomy* rather than *duty* as its central buzzword.[9] A lively

[4] Kant, *Critique of Practical Reason*, trans. and ed. Mary Gregor (Cambridge: Cambridge University Press, 1997), 5:63. Henceforth *KpV*.
[5] See Barbara Herman, "Leaving deontology behind", in *The Practice of Moral Judgment* (Cambridge, MA: Harvard University Press, 1993), 210-211.
[6] Arthur Schopenhauer, *The World as Will and Representation vol. 1*, trans. E.F.J. Payne (New York: Dover, 1969), 1:84 (section 16)
[7] Herman, 210
[8] Jens Timmermann, "What's wrong with 'deontology'?", *Proceedings of the Aristotelian Society* 115(1) (2015), 85
[9] Ibid., 88

debate between proponents of a law-first approach and a good-first approach has emerged, thanks to which Kantian ethics has resurged as a serious contender to answer the vexing questions of modern metaethics. As Ebels-Duggan notes in an overview of this debate, neither side of this debate accepts the classic deontological picture of Kant, which now mainly lives on in the minds of his opponents.[10]

If someone of a broadly Moorean persuasion asked what Kant's theory says about the good, however, it would still be difficult to point them in the direction of a straightforward discussion in the literature. There is no paper or book simply called 'Kant on the good', surely a title that would seem to be low-hanging fruit in the large grove of publications on every aspect of Kant's philosophy. This is mere anecdotal evidence, of course, and does not by itself establish that there exist no meaningful and well-argued insights into Kant's view of the good, as in fact there do. It is striking, however, that the question is usually taken up indirectly as a means to making some other point, and that some of the most relevant texts have continued to be underexamined.

This dissertation makes a modest effort at providing a sustained account of the good in Kant's moral philosophy. In doing so I do not claim to say much that is very original. In fact I spend much of the dissertation simply elaborating the claims of others. I believe that there is at least some value, however, in putting together all the theorising that has been done on various aspects of this question and assessing what fits, and what does not; and in pursuing the concept of the good directly to the various places it takes us within Kant's wider philosophy. It is my hope that doing so will at least provide some original and more thorough support for (or grounds for rejection of) existing positions in the scholarly debate. In particular, I extensively examine a conception of good which sees it as exhaustively defined by being the formal object of the will or, which I will show to be an equivalent claim, as an *a priori* concept of practical reason. This conception is fairly new to the literature and has been put forward in

[10] Kyla Ebels-Duggan, "Kantian ethics" in Christian Miller (ed.), *Continuum Companion to Ethics* (New York: Continuum, 2011), 182

these respective terms primarily by Stephen Engstrom[11] and, following the former, Sebastian Rödl.[12] As I show in the first chapter, it has the potential to put Kant's ethics on firmer footing than other traditional interpretations – but this does require that serious exegetical and philosophical objections to it are overcome.

Kant is first challenged to elaborate his view of the good by H.A. Pistorius, author of the first full-length review of the *Groundwork*, and duly does so in chapter II of the Analytic of Practical Reason: 'On the concept of an object of pure practical reason.' This historical context is well-acknowledged, but there is little systematic study of the dialectic between Kant and Pistorius despite the fact that the latter's objections often predate well-known criticisms by later theorists. In the first chapter of this dissertation I provide a detailed study of the objections Pistorius raises. I trace out Kant's reply in the second Critique, first taking a traditional deflationary line that reduces the good to a concept that is merely derivative of the moral law. I follow this line as far as it will go, but find that it ultimately fails to explain the force of the moral law and cannot explain its refusal to address certain forms of moral scepticism. I also briefly consider, and reject, a moral realist interpretation of Kant which claims that we could base our concept of the good on something else than our motivational faculties. This interpretation, though held by some, saddles Kant with implausible metaphysical commitments, contradicts his epistemology on important points and is contradicted directly by textual evidence. Lastly, I show that introducing a more substantive view of the good as the necessary object of intentional action turns out to yield much more satisfying solutions to these puzzles and have some initial textual support from the second Critique.

This view commits Kant, however, to an ancient and newly controversial thesis in the philosophy of action: that all intentional action is performed *sub ratione boni* or, in a modern phrase popularised by

[11] Stephen Engstrom, *The Form of Practical Knowledge* (Cambridge, MA: Harvard University Press, 2009), 12
[12] Sebastian Rödl, "The form of the will" in Sergio Tenenbaum, ed., *Desire, Practical Reason, and the Good* (Oxford: Oxford University Press, 2010), 139

Velleman, 'under the guise of the good'. I consider the case for this thesis in the second and third chapters. In the second, I first provide a general overview of the recent debate, which serves to illustrate some of the pitfalls which an adherent to the guise of the good should seek to avoid and ways in which the claim is variously understood. I then move to discuss several versions of the thesis that can be attributed to Kant. These include the weak claim that only purely rational willing aims at the objective, i.e. moral good; the strong claim that *all* rational willing aims at the objective good; and the claim that all rational willing aims at *either* the objective or the subjective (prudential) good. I particularly focus on the case for the strong claim, to which Engstrom and Rödl, as well as Andrews Reath, are committed.

What I seek to contribute uniquely to the defence of this claim is a more solid grounding of it in Kant's moral psychology. This is particularly important because this account of willing owes us an explanation of what goes wrong in the case of 'bad' or defective willing. According to the strong thesis an agent whose willing does not accord with the moral law must either be confusedly or mistakenly aiming at the objective good, or not really be willing at all. Its defenders have consequently offered fairly general accounts of such bad willing as resting on a cognitive mistake. However, it is clear that bad willing is not a single phenomenon. There are many ways one's choice can fail to instantiate the moral law. One might be committed to an evil principle, take one's own interests to supersede moral demands, feel too weak to live up to the demands of morality, etc. What I argue is that Kant offers a rich moral psychology that yields compelling explanations of all these various phenomena, and that adherents of the strong thesis can and should be using it to defend their claim. I first explore the concepts of self-conceit and despondency, show that they underlie bad willing generally, and show that an agent suffering from these conditions can compellingly still be said to be making a claim to their actions being objectively good. I hope hereby to flesh out the nature of the 'cognitive mistake' gestured at by defenders of the strong claim.

In the third chapter, I apply these general insights to three specific categories of bad willing in order to see whether they can be explained with the resources available under the strong thesis. In doing so I considerably refine the moral psychology and the picture of will and choice (*Wille* and *Willkür*) that emerged from chapter 2. In the first section I consider diabolical willing (evil for evil's sake) and evil willing (placing self-love above the demands of morality.) I consider these together because I show that the former is impossible and that apparent instances of it reduce to the latter, of which Kant already offers a well-developed account that is consistent with the strong thesis. In the second section I turn to frailty, a phenomenon in which the agent's bad actions do not accord with their good will. I contend that Engstrom's and Reath's explanations of it are misguided, and rest on a misapplication of the *Wille*/*Willkür* distinction. However, I suggest that a different account offered by Sergio Tenenbaum, which avoids these problems, is available under the strong thesis. Lastly, I outline the most problematic case, which I call 'listlessness'. A listless agent is one that acts despite seeing little value in action altogether. This should not be possible according to the Kantian theory of action, but appears to describe a number of real cases including instances of clinical depression. Though I explore some promising strategies that might mitigate the damage, I contend that Kant's picture of agency lacks resources to deliver an ultimately compelling explanation. However, defenders of the strong thesis are no worse off in this regard than those of any other view, so long as they remain committed to Kant's picture of agency. Whether this should be considered a definitive reason to reject the theory depends on one's assessment of the reality and importance of listlessness, as well as the comparative strength of other theories of action in explaining this rather puzzling phenomenon.

In this dissertation I take an analytic approach to my historical subject, and combine exegesis and philosophical analysis. I am conscious of the methodological uncertainty of such a two-pronged approach, as one risks cherry-picking whatever method best helps push one's preferred narrative at any given moment. I would therefore like to clarify my methodology here to the best of my ability.

My interest is ultimately in the most philosophically compelling explanation of Kantian ethics, rather than the one that most faithfully reflects the views of the historical Kant. I am willing in principle to reject or revise aspects of the latter's theory if these can be shown to render it significantly less philosophically plausible. However, I also believe that Kant's texts combined offer such a rich and compelling picture of human agency that such revision will in practice seldom be necessary. Where a more appealing position is available, there is almost always textual evidence in its favour as well. Hence I usually appeal to philosophical merit as a tiebreaker between contradictory pieces of text or between two interpretations of a single piece of text, rather than between the text and a view wholly external to it. On the few occasions where I feel compelled to do the latter, I hope to also have shown that the view from outside is the more compatible with Kant's wider picture if not its details. In general, I take for granted some of the most basic Kantian commitments such as the autonomy of the will, the *Wille*/*Willkür* distinction, the universal character of morality and the identity of will and practical reason. While I try to explain them briefly in what follows, defending them would far exceed the purpose of this dissertation. Beyond this I assume, simply put, that Kant is a very good philosopher; and that one can attribute to him whatever position, consistent with these premises, he could have held in order to best respond to the philosophical challenges he faces. I further assume that his views are better articulated and thought through in his later work. Hence the most mature version of his ethical theory is the best philosophical theory to hold, and when in conflict we should give greater weight to views expressed in later works than those in, say, the *Groundwork*. I will not do much work here to defend this latter assumption. It will become clear in chapter 1, however, that it is particularly true for Kant's view of the good. The *Groundwork* offers very little clarity on the topic and Kant progressively appears to see the need to articulate clear views on it in his ethical works from the second *Critique* onward, rounding these views out with the moral psychology of the *Religion* and *Metaphysics of Morals*.

Chapter 1: Pistorius and the paradox of method

In 1786, Hermann Andreas Pistorius published the first substantial review of the *Groundwork*, writing anonymously in the journal *Allgemeine Deutsche Bibliothek*. Though not altogether negative, the review deemed Kant's project on the whole unsuccessful and presented a number of objections, many of which have continued to be pressed by later interpreters. That Kant took this review seriously is evident from the Preface to the *Critique of Practical Reason*: "In the second chapter of the Analytic I have, I hope, dealt adequately with the objection of a certain reviewer of the *Groundwork* (…), one who is devoted to truth and astute and therefore always worthy of respect".[13] Kant understood Pistorius as mainly complaining that "[in the *Groundwork*] the concept of the good was not established before the moral principle (as, in his opinion, was necessary.)"[14] He indeed devotes a substantial part of chapter II of the Analytic of Practical Reason to a defence of his 'paradox of method', which refers to the thesis that "the concept of good and evil must not be determined before the moral law (for which, as it would seem, this concept would have to be made the basis) but only (…) after it and by means of it".[15] The crux of Kant's reply appears to be that, if not through the moral law, the concept of good could only be empirically determined through the faculty of pleasure and displeasure and any principles derived from it would therefore lack the objective necessity which would characterise them as moral. This reply, or at least the gloss I give of it here, can hardly be said to have put the issue to bed even for those sympathetic to Kant's project. A number of interpreters have taken Kant at his word and taken the law as having absolute priority over the good, and have subsequently downplayed the importance of the latter in providing a complete account of Kantian ethics. As hinted at in the introduction, this last approach sometimes comes uncomfortably close to the deontological caricature – which is precisely what is already attacked in the Pistorius review. In response, others have rejected Kant's conclusion, and have

[13] *KpV*, 5:9
[14] Ibid., 5:8-9
[15] Ibid., 5:63

continued to press Pistorius' claim in trying to provide non-empirical conceptions of good on which the moral law could be based.

This is by no means an exhaustive characterisation of recent scholarship, which in fact exhibits an increasing tendency towards a more complex interpretation: paying more attention to the good while respecting Kant's stated views on the paradox of method. Nevertheless, in the light of this continued controversy, there is a surprising lack of in-depth study of the argument made in Chapter II of the Analytic and especially of Kant's engagement with Pistorius in that passage.[16] It seems often to be viewed as merely restating known Kantian views or unsurprising developments thereof. Beck for instance expects the reader to have been "lulled into a feeling of easy familiarity by the no doubt interesting, but certainly not surprising, development of the relation between principle and concept, between law and the good, in the beginning of chapter ii [until 5:65]".[17] Beiser similarly views Kant as "merely reiterating his conviction that the moral law can be determined apart from all utilitarian considerations" while this is "just the belief that Pistorius questions."[18] The passage is seldom considered in conjunction with the Pistorius review that gave occasion for it.[19] I consider it highly likely that paying closer attention to this context will helps us better understand what Kant was trying to say in this passage. To that end, in this chapter I first provide a reconstruction of Pistorius' criticisms in the

[16] Notable exceptions include John R. Silber, "The Copernican revolution in ethics: the good reexamined", in Robert Paul Wolff (ed.), *Kant: A Collection of Critical Essays* (London: MacMillan, 1968), 266-290, whose reading is similar to the one I offer in section II; and Volker Dieringer, "Was erkennt die praktische Vernunft? Zu Kants Begriff des Guten in der *Kritik der praktischen Vernunft*", *Kant-Studien* 93(2) (2002), 137-157. Neither makes reference to Pistorius' review, however, and in recent Anglophone scholarship the neglect is all the more striking. Beiser recognises the importance of Pistorius' review to the second *Critique*, noting that "many of the sections of the second *Kritik* are disguised polemics against Pistorius" in Frederick C. Beiser, *The Fate of Reason: German Philosophy from Kant to Fichte* (Cambridge, MA and London: Harvard University Press (1987), 188). He also provides a cursory overview of some of Pistorius' objections and Kant's replies in 190-192, though some of them are drawn from other Pistorius texts than the *Groundwork* review.

[17] Lewis White Beck, *A Commentary on Kant's Critique of Practical Reason* (Chicago: University of Chicago Press, 1960), 136

[18] Beiser, 190

[19] One extremely recent exception is Jens Timmermann, "The law and the good: Kant's paradox of method", forthcoming in Violetta Waibel (ed.), *Proceedings of the Vienna Kant Congress*, (Berlin: Walter de Gruyter).

relevant parts of his review.[20] I also briefly consider the positive view Pistorius puts forward, which he takes to be a sympathetic suggestion in the spirit of Kant's own writing, and Kant's reply to this (broadly Stoic) position. I then engage in a close reading of the paradox of method passage in order to reveal the exact nature of Kant's view of the good. I will proceed dialectically through several readings of it. First, I present the 'orthodox' law-first view and argue that it begs the question against important aspects of Pistorius' objections. I then proceed to a value-realist ('good-first') interpretation and argue that, while closer to Pistorius' own view, it is both philosophically and textually problematic and flatly contradicts important aspects of Kant's moral epistemology. Lastly, I argue that the seeming dichotomy between these approaches is due to a stronger reading of the paradox of method than warranted by the text. There is good reason to instead endorse Stephen Engstrom's reading, which holds that the good is determined by, but not conceptually derivative of, the moral law. This reading gives us access to Engstrom's own practical-cognitivist approach to the good, which I will present here in outline. I will show that this approach closely fits this particular passage and, *if successful*, shows Kant to have delivered a much more satisfactory reply to Pistorius' objections. I leave deliberately undecided for now whether it can in fact be successful, and rely on certain premises I do not here substantiate. Rather than arguing for practical cognitivism directly here, its introduction as a solution to the Pistorius problem is meant to provide a merely indirect argument to show that this is a view we would *like to be able to hold*. Its success in solving said problem motivates my closer investigation of the unsubstantiated premises, and the plausibility of the view as a whole, in the second and third chapters.

1.1 Pistorius' review

In the opening line of the *Groundwork* Kant extols the good will as the only thing that can be "taken to be good without limitation,"[21] to add on the next page that "a good will is good not because of what it

[20] I am deeply thankful to Michael Walschots for sharing with me a draft of his translation of Pistorius' review.
[21] *GMS*, 4:393

effects, or accomplishes, not because of its fitness to attain some intended end, but good just by its willing, i.e. in itself".[22] Pistorius begins by taking exception to this:

> "In this respect I would have wanted it to be preferable to the author above all to discuss the general concept of that which is *good* in general, and what is a good will in particular? Can a will that is good in and of itself, regarded as having no relation to an object of any kind, even be conceived?"[23]

Pistorius immediately identifies a notable gap in Kant's account of morality in the *Groundwork*: he never spells out a definition of goodness, despite clearly intending the term to play an important role from the very start.

In Kant's defence, perhaps he intends his usage here to be in some sense intuitive or pre-philosophical. This is after all still the part of the *Groundwork* concerned with 'common moral cognition.' If this is so, however, Kant seems to have seriously misjudged common moral cognition and its difficulties comprehending 'goodness in and of itself'. Absent a definition, it is difficult to see how goodness could be a property of a will irrespective of its objects. When we conceive of a will independent of its object, we are left with the mere activity of willing. This activity seems value-neutral at best. If we then try to conceive of how such a will could be good, the most intuitive thought seems to be that it is made good by willing a good object. This is the intuition which Pistorius continues to pursue:

> "Here I do not see how one can accept anything at all as completely and absolutely good without exception, or could call something good, *that* in reality would come to nothing good, and just as little [I do not see] how one could accept a will that is good absolutely and in and of

[22] *GMS*, 4:394

[23] Hermann Andreas Pistorius, '*Groundwork of the Metaphysics of Morals* by Immanuel Kant', trans. Michael Walschots, 27. Since this is unpublished, I use the page numbers from the German version reprinted in Bernward Gesang, *Kant's Vergessener Rezensent* (Hamburg: Felix Meiner Verlag, 2007).

> itself. But the will is meant to be absolutely good **not** in relation to some object, but **only** in relation to its principle or a law, for the sake of which it acts."[24]

Pistorius conceives of the good in what we might now call consequentialist terms. Inherent goodness is not a property of willing, but of states of affairs in the world. Willing cannot be good if it does not bring about good effects. A will can therefore only be good because of its relation to an object (i.e. some state of affairs) rather than from its relation to a principle, presumably because this principle would also derive its value from the states of affairs it helped bring about. He sees this as the only way to escape the vicious circularity he identifies in the Kantian conception of the good will as a will acting for the sake of principle:

> "But let this be so [*Es sey so*];[25] then I ask further: is it sufficient to establish a will as the Good that it acts merely according to any kind of principle or from respect to any law, be it as it may, good or evil? – impossible, thus it must be a good principle, a good law, and the question 'what is good?' turns back around, and if we have pushed it back from the will to the law, then we must now answer it here in a satisfactory way; i.e. we must eventually come to some kind of object or to an ultimate end of the law, and we must avail ourselves of what is material, because we cannot get by with what is formal in relation to either the will or the law."[26]

Granting that the will is made good by its law, surely not *any* principle will do. Kant of course agrees: since the goodness of the will cannot be dependent on a particular object and hence not on a material principle, the only remaining candidate is the mere *form* of willing under law, i.e. the categorical imperative.[27] But why is this a *good* principle? What makes willing under the representation of universal

---

[24] Pistorius, "Groundwork", 27. The last sentence is a correction to Walschots' draft by Jens Timmermann, to clarify that Pistorius is here representing Kant's view and not his own.
[25] I here emend Walschots' original "But this is so", an emendation which to the best of my knowledge he has taken on board for the final version at the suggestion of Jens Timmermann.
[26] Pistorius, "Groundwork", 27
[27] *GMS*, 4:402

law inherently good, if not some end that is furthered by it? (This objection, as previously noted, is what makes the deontological caricature look so unappealing.)

Pistorius raises a third objection, independent from issues of goodness. This objection is meant to show that the notion of a categorical imperative as Kant conceives of it – that is, as commanding absolutely unconditionally – is impossible. This motivates Pistorius in offering something of a positive alternative. The reason I include it for consideration here is that, as I will show later, Kant makes the argument that Pistorius' position on the good leads to absurdity and that his own position remains as the only alternative. Without a response to this third objection, however, it would turn out that *both* options are incoherent.

Pistorius says that despite Kant's rhetorical bluster about the categorical imperative being merely formal and valid independent from contingent material ends, it is only by reference to such ends that we can obtain the results Kant wants. His argument relies on Kant's example of the false promise. Here, I ask myself whether I could will a world in which it was universal law that promises were made to the promiser's own advantage without any intention of being kept. Kant swiftly concludes that I could not, because such a law would invalidate the very concept of promise. Pistorius agrees, but questions that this conclusion can be had by relying on pure reason alone. We could imagine a being who is rational but who does not have any interests, not even in truth or falsity. Such a being has no reason to will the enduring existence of the institution of promising. What is more:

> "Indeed, it seems to me to be completely unthinkable that a law could in fact be given to such a completely uninterested being, and that it could be necessitated morally, i.e. via representations, to observe it."[28]

Since this being lacks interests, the moral law has no utility for it. The only remaining ground for compliance with it would be merely cognitive: recognition of the law as true or reasonable. But, *ex*

[28] Pistorius, "Groundwork", 32

hypothesi, this being does not have interest even in what is true. It can then have no reason to respect the law. Therefore, Pistorius infers, morality must always presuppose some interest; and duty from a merely formal principle that abstracts from all interests is indeed the "empty delusion and a chimerical concept"[29] Kant had feared. There is then "no other moral law than a hypothetical law" based on the interest of rational beings, and:

> "Accordingly, the good will would be that will whose maxim is: do that which is in conformity and agreement with your and simultaneously the common interest of all rational beings. (…) For, higher and deeper than in the common nature of all rational beings, the rule of their will and their conduct cannot be sought. This principle becomes binding and turns into a law for me through the representation that my interest and that of all rational beings is one and the same".[30]

Pistorius does not spell out why he believes that all rational beings share the same interest. We might read him in roughly two ways. It is possible that he takes himself to be agreeing with Kant that all rational beings "in so far as imperatives suit them, namely as dependent beings" necessarily have happiness for an end. We can therefore have no higher duty than to impartially promote the general happiness of our fellow rational beings, which Pistorius takes to amount to treating them as ends-in-themselves. Of course, Kant himself did not infer from the fact that all rational beings have an interest in happiness that they have the same interest, since happiness is an indeterminate concept.[31] As agents vary between and even with themselves as to what constitutes their happiness, they may have an interest in widely different objects for the same end. What is more, agents take an interest only in their *own* happiness. Even if happiness were a determinate concept, our interest in it would not be common between us; rather, as we might both desire the same object for the sake of our own happiness, it

[29] *GMS*, 4:402
[30] Pistorius, "Groundwork", 33
[31] *GMS*, 4:418

would be very likely to conflict.[32] Pistorius does not appear to address, or take very seriously, this part of Kant's doctrine, and his positive conception would fail to get off the ground as a result.

However, it may not be fair to ascribe this sort of simple hedonism to Pistorius. Given his historical context and the remarks he makes elsewhere in the review, he appears to hold a view more akin to that of the Stoics. This is hinted at in his reply to Kant's 'teleological excursus',[33] when Kant makes the startling claim that happiness cannot be the end of rational beings since, if so, nature would have done better equipping us with instinct alone:

> "yes, I answer, if happiness through reason and happiness through instincts were the same thing, and between the two there is no difference other than that the former is weaker, inferior, and less pleasant than the latter, and if we are not at all permitted [first] to consider the fact that we are indebted to our own efforts for it, [second] to be aware of the pleasant addition, that it is in large part the work of our own self-activity – only then would it be more accurate and expedient to drive human beings to happiness through instinct, or what is the same, to give them an animal or instinctual happiness." [34]

Evidently Pistorius objects to Kant taking the difference between happiness through reason and happiness through instincts as being one of mere degree. It is slightly harder to see, however, what Pistorius thinks constitutes the *right* account of this difference. On the one hand it seems as if he wants them to be wholly distinct, with the latter qualifying as mere 'animal' happiness and the former, by implication, as 'human'. On the other hand, his only argument here is that in using reason, we take additional pleasure in the ownership we take of our work. This argument merely shows that Kant has the difference of degree running the wrong way, and that we can in fact obtain *more* happiness through

[32] Kant discusses the case of King Francis I and Emperor Charles V, who both desired Milan, in *KpV* 5:28.
[33] The term is from Jens Timmermann, *Kant's* Groundwork of the Metaphysics of Morals: *A Commentary* (Cambridge: Cambridge University Press, 2007), 23. It refers to Kant's argument in *GMS*, 4:395-396.
[34] Pistorius, "Groundwork", 28

reason than through instinct. Pistorius would then be equally vulnerable to his own objection that two kinds of happiness are here not distinguished; one is simply more or less of the other.

Pistorius is more helpful when discussing the aforementioned 'common interest of all rational beings' near the end of his review. Here he presents a starker contrast between what might be appropriately desired on the basis of instinct, and willed by reason. My particular interest, which we may take to be what instinct leads me to pursue, can clash with the "universal interest" to which reason guides me. In such cases I should always choose the latter.[35] The reason for that is that any particular interest is only a part of my nature, while my entire or true self is captured by my participation in the universal interest. My rational self for Pistorius is then my truest self. So in discussing Kant's formula of autonomy, he says "I […] cannot conceive of or wish for a more free legislation, other than that which gives expression to my own nature, and that which the Stoics express through this formula: *naturam, optimum ducem, tanquam Deum sequi, naturae conventienter vivere* [the nature of the best way to follow God, is to live conformably to nature] etc."[36] Since Pistorius does not admit that the will can be genuinely autonomous in the Kantian sense, that is, be determined wholly by principle without reference to an object, perhaps what he before called 'happiness of reason' is the object of a will that is in conformity with its true nature. This happiness of reason is a common interest of all rational beings and thus we must assume that its pursuit does not produce conflicts of will. To be maximally charitable to Pistorius, we may also assume that since the content of this happiness is determined by nature, it is in fact a determinate concept. Rational beings are bound to moral principles insofar as they all have an interest in procuring the happiness of reason. The latter is then the 'good' to which the good will is directed.

To recap, there are three salient challenges in Pistorius' review which I will focus on addressing.

[35] Pistorius, "Groundwork", 34
[36] Ibid., 36

1) Can Kant provide an independent definition of the good, i.e. in non-moral terms?

2) What makes the moral law a good law, or the right law for us to follow?

3) How could a being without interests, including cognitive interests, be bound by the law?

Believing that Kant cannot address these challenges, Pistorius proposes that a broadly stoic ethic saves what there is to be saved in Kant's theory. Grounding moral commands on a common interest in the good of the happiness of reason, he believes he is able to reconstruct the substance of the categorical imperatives as a system of hypothetical imperatives. [37]

1.2 Kant's reply in the second *Critique*

It is evident that, lacking the benefit of over two centuries of philosophical work on the text, Pistorius misreads or simplifies the doctrines of the *Groundwork* in a number of places; and *Groundwork* scholars may to various degrees be willing to say that his criticisms can be met by simply appealing to a more careful reading of the original text. However, the review calls attention to at least one important point:

[37] Interestingly, Pistorius also later reviewed the *Critique of Practical Reason*. If his *Groundwork* review has often received little more than lip service in the literature, this second review is almost completely neglected. (So far, it is also untranslated, and any translations that occur in this text are my own.) In fairness, compared to the first it is much less interesting. Pistorius spends much of it summarizing the book, and seems to find it much more to his liking than the *Groundwork*. (See H.A. Pistorius, "*Kritik der praktischen Vernunft* von Immanuel Kant", in Gesang*,* 85.) He continues to read perceptively and make some interesting objections, but these do not directly concern our topic here. I will here briefly set out the most notable ones. He first casts doubt on the fact of reason, which he takes to be that we must admit that there is a purely formal imperative that binds us, and which has now become the unproven foundation of Kantian ethics (86). Pistorius continues to doubt whether a completely formal moral imperative is possible and sufficient. The only completely formal law of reason, he says, is that of contradiction. There are cases, however, where two maxims are wholly opposed and both entirely self-consistent: reason cannot decide between these maxims solely on the basis of the law of contradiction. (87) For this, it would still have to admit the interests of rational beings into its decision, which again leads Pistorius to assimilate Kantian ethics to stoicism (88). He accuses Kant of focusing only on the intellectual side of human nature and neglecting its sensible side, which leads to a mistaken attribution of transcendental freedom to the human being (88-89). He also complains about the stringent nature of moral duty in Kant, which by abstracting from ends leaves too little room for wide duties concerning human welfare. This leads to such absurd conclusions as that a ruler should let his people perish, if the only means to save them was to break a promise – conclusions which Pistorius seriously doubts can in fact be endorsed by reason or common human understanding (90). This proves, says Pistorius, that the fact of reason does not in fact carry the weighty authority for all rational beings which Kant says it does, striking down all ends-directed reasoning that conflicts with the categorical imperative (91). Many of these complaints are familiar to Kant scholars, and are addressed either in later work (The *Metaphysics of Morals* especially comes to mind) or by a more careful reading/reconstruction of the text. I will not rehearse these points here. Pistorius does offer a somewhat interesting reply to the paradox of method passage, however, to which I will have occasion to refer below.

Kant's lack of an explicit definition of the good. As we have seen, it is due to this lack that Pistorius is able to find room for his counterproposal. I therefore want to bracket the question of Pistorius' exegetical qualities here, and focus instead on Kant's positive efforts in the *Critique of Practical Reason* to clarify and expand on his ideas so as to resist his reviewer's objections. I will show that a proper understanding of Kant's answer to the first objection allows us to derive satisfying answers to the second and third as well.

However, I will begin with some remarks on Pistorius' own proposed solution. It is worth pointing out that in the second chapter of the Analytic, Kant delivers a partial reply to Stoicism. In distinguishing the good from mere well-being, he approves of the Stoic sage who refuses to see his pain as an evil. This Stoic has correctly seen, Kant believes, the difference between an ill that negatively affects his well-being and an evil, in that the pain "did not in the least diminish the worth of his person but only the worth of his condition."[38] However, stoicism downplays too much the genuine importance of well-being to practical reason. "The human being is a being with needs, insofar as he belongs to the sensible world, and to this extent his reason certainly has a commission from the side of his sensibility which it cannot refuse, to attend to its interest and to form practical maxims with a view to happiness in this life and, where possible, in a future life as well."[39] Reason is not *merely* to be used as a tool for the attainment of our animal ends, but these nevertheless are a component of our self. Stoicism goes too far in denying their importance altogether and in insisting that our 'truest self' is found by complete abstraction from our sensible desires.[40]

That this reply occurs in the section devoted to dealing with Pistorius' review might suggest that Kant is here also addressing himself to Pistorius. There are some reasons, however, to resist this

[38] *KpV*, 5:60
[39] Ibid., 5:61
[40] A more extensive treatment of stoicism along these lines is given by Kant only later, in the Dialectic of Pure Practical Reason, in the context of determining the highest good. See *KpV*, 5:111-113, 5:126-127.

suggestion. Firstly, insofar as it refutes stoicism it does so indirectly. The part of the section that explicitly mentions the stoic sage does so only to endorse their view – the part from which Kant's criticism of the view can be drawn occurs slightly later in the text and does not directly refer to stoicism. Secondly, it is not obvious that Pistorius' stoicism took quite the form refuted above. He never explicitly advocates the abnegation of the sensible self, which he does consider a 'part of our nature'.[41] Though he does seem to hold that we should never choose this 'part' over our 'true and entire' nature, it is not clear that this is very far from Kant's own position. Kant himself enjoins us to make the interests of other rational beings our own[42] and to only pursue happiness insofar as it is qualified by agreement with the moral law.

Thirdly, Kant does not need a reply specifically to the Stoic principle if he can show that it belongs to a family of mistaken – because heteronomous – ethical theories *and* that there exists a genuine alternative in the form of a theory grounded in autonomy. It was only because of the perceived failure of the Kantian notion of autonomy in the *Groundwork* that Pistorius introduced the Stoic principle as the next best thing.[43] Stoicism is heteronomous insofar as its moral principles are founded on perfection as an object of the will, and perfection can only be relative to already given ends. The moral command is then based on these given ends, rather than given to reason on its own accord.[44] Part

[41] As noted above, he would later in fact complain that Kant neglected the sensible self too much. Pistorius, "*KpV*", 88.
[42] Kant, *Metaphysics of Morals,* trans. and ed. Mary Gregor (Cambridge: Cambridge University Press, 2015), 6:388
[43] Kant's remarks on stoicism suggest that he would agree with this appraisal, since the theory compares favourably to others. It agrees with his own in making virtue the supreme practical principle (*KpV*, 5:126) and "quite rightly" refused to make pleasure the motive for virtue (5:115).
[44] This refutation occurs explicitly in *KpV*, 5:41. An embryonic, but very brief version already featured in *Groundwork*, 4:443-444. This version only refers to the concept of perfection and does not mention stoicism explicitly as an ethic of perfection. Kant somewhat glibly asks to be "exempt from a lengthy refutation of all these doctrinal systems. It is so easy, and presumably so well understood (…) that it would only be superfluous labour." The passage is not very clear about the way in which perfectionism is heteronomous, and Pistorius might well be excused for not thinking that the criticism contained in it applied to his own ethical views. It is possible that the review contributed to Kant seeing that he needed to be clearer about his rejection of stoicism in the second *Critique*, where the view is much more extensively and respectfully discussed.

of a successful reply to Pistorius for Kant lies simply in vindicating his concept of autonomy, a large part of which needs to be done by dealing with Pistorius' other objections.

This first requires giving a definition of the good. Kant calls good and evil "the only objects of a practical reason."[45] He explains his meaning in the opening lines of chapter II of the Analytic:

> "By a concept of practical reason[46] I understand the representation of an object as an effect possible through freedom. To be an object of practical cognition so understood signifies, therefore, only the relation of the will to the action by which it or its opposite would be made real, and to appraise whether or not something is an object of *pure* practical reason is only to distinguish the possibility or impossibility of *willing* the action by which, if we had the ability to do so (and experience must judge about this) a certain object would be made real."[47]

An object of practical reason is such that it would be possible for a free will to act so as to effect it in the world. Kant clarifies his usage of 'possible' as follows. If the object itself is the determining ground of the faculty of desire, we inquire as to its physical possibility. However, if the moral law is the determining ground, we are concerned only with the moral possibility of the object.[48] Since a free will simply is a will determined by the moral law, the former can be laid aside and we can focus solely on the latter. Kant states that the proposition "we will nothing under the direction of reason except insofar as we hold it to be good or evil" is "indubitably certain".[49] Therefore, the good and the evil must be objects

---

[45] *KpV*, 5:58
[46] I here remove the interpolated 'of an object' (*'eines Gegenstandes*') which, through the *Akademie* edition, has found its way to most English translations including Gregor's. This interpolation has been thought necessary because of the incongruence of this opening line with the chapter heading, "On the concept of an object of pure practical reason". Its removal might be seen as stacking the deck in favour of Engstrom's interpretation as given below. However, I believe it warranted by the textual evidence. A concept *is* a representation: it represents objects as having a certain feature, in this case that of being "an effect possible through freedom." By falling under this concept, i.e. being "so understood", the object comes within the range of practical cognition, i.e. becomes "an object of practical cognition." If we can understand the text in this way I believe we should be hesitant to amend it. I take it that my argument below does not hang on this textual point either way.
[47] *KpV*, 5:57
[48] Ibid., 5:58
[49] Ibid., 5:60

of practical reason, and in fact the only objects of *pure* practical reason. (This does not justify Kant's claim, however, that they are the only objects of practical reason in general. Once impure practical reason is considered, the possibility arises of the inclinations suggesting other objects. I will bracket this worry for now, and assume that we can restrict this claim to pure practical reason. A more satisfying solution will be provided in section III.)

So construed, this passage plays into the deontological caricature, in which 'good' is reduced to a completely derivative concept merely indicating 'whatever the moral law enjoins.' This impression is strengthened by Kant's central statement of the paradox of method, "namely, that the concept of good and evil must not be determined before the moral law (for which, as it would seem, this concept would have to be made the basis) but only (as was done here) after it and by means of it."[50]

To explain this thesis, Kant proposes that we try things the other way around and begin by trying to determine the concept of the good, inferring from it the moral law.[51] The only other way to determine the good would be through the faculty of (dis)pleasure. In accordance with Theorem I of the *Critique*, principles based on the empirical conditions of pleasure and displeasure cannot carry objective necessity and therefore never yield *a priori* laws.[52] This is so because receptivity to pleasure or displeasure is a subjective condition, and it cannot be said *a priori* that they would occur in the same way for all rational beings. No genuinely moral law, therefore, could be based on a prior conception of the good.

So far the argument against Pistorius has been a mere *reductio*: understanding the good independently of the moral law yields absurdity, and therefore the law must be determined before the

[50] *KpV*, 5:63
[51] One might wonder why it is necessary to infer a moral law in the first place. This is because Kant conceives of the will as a causal faculty, and the concept of causality directly implies determination by some law. If we were to make do merely with some value concept like 'good', we would have a gap between that concept and its ability to determine the will.
[52] *KpV*, 5:21-22

good. In fact, I would suggest this is as far as the usual reading of the argument goes.[53] However, eliminating the alternatives does not automatically make Kant's view right. Philosophically there is no guarantee that *either* approach has to be correct: perhaps neither the good nor the law can be determined first. This option remains open so long as Kant leaves Pistorius' second and most vital question unaddressed: what makes the moral law itself good?[54]

One might suggest that this question rests on a category mistake: if goodness is a concept that follows from the moral law, it cannot be applied back to that law. No other credentials for the moral law are possible, or necessary, beyond autonomy: it is the law we give to ourselves. Good or not, it is our law and we have no other. This reading is most notably defended by Jens Timmermann, both

[53] See f.i. John R. Silber, *Kant's Ethics: The Good, Freedom, and the Will* (Berlin: De Gruyter, 2012), 276.

[54] Pistorius himself does appear to accept Kant's dichotomy in his review of the second *Critique*: "Here the question arises, whether the concepts of good and evil let themselves be given or determined prior to and without consideration for the moral law, or whether these concepts are first to be determined through the moral law after it is fixed [*festgesetzt*]. [In *KpV*] the latter is claimed, and it is shown that in the first case the good and evil could not be determined any other way than empirically, and that as a result the moral law, whose object would be this empirically found good and evil, could not be a universally necessary *a priori* law of reason [*allgemeines nothwendiges Vernunftgesetz* a priori]; **therefore only the second case remains**, namely that the good and evil are first determined only through and after the moral law, or that the concept must be derived from a preceding practical law, but cannot lie at the foundation of that law." Pistorius, "*KpV*", 81-82 (emphasis mine). Timmermann also notes that Kant is well aware that the indirect style of argument employed here cannot establish *why* his own position is correct, and relies on only two options being available (Timmermann refers to Kant, *Critique of Pure Reason*, trans. and ed. Paul Guyer & Allen Wood (Cambridge: Cambridge University Press, 1998), A790/B818, henceforth *KrV*.) He agrees that "[p]hilosophically, none of this can be taken for granted". However, the *reductio* succeeds in the dialectic with Pistorius since he and Kant agree on the assumptions "that there are valid moral imperatives and that ethics must have something to say about goodness as well as laws, that there are only two ways of relating them to each other etc." (Timmermann, "The law and the good", 5-6f). I think it is correct that Pistorius would be unwilling to entertain the third option, that neither notion can be determined independently, since he is no moral sceptic. I have some reservations about attributing to him the assumption 'that there are valid moral imperatives', depending on whether 'moral' is here understood in the strictly Kantian sense. As we saw, Pistorius thinks all imperatives are ultimately hypothetical, though he holds that the ends of rational beings are such that there are some that are universally shared (and he appropriates the term 'categorical' to refer to these.) Pistorius himself would undoubtedly call these imperatives moral. Kant would not, however, since he reserves the name 'imperative of morality' for the categorical imperative proper. (Kant, *GMS*, 4:416.)

independently[55] and in the context of Kant's reply to Pistorius.[56] I agree with it as an interpretation of the *Groundwork*, but this argument was not enough to satisfy Pistorius:

> "if the author understands by [autonomy] that the will gives itself a law without considering whether this law is good for something, (…) then this high-handed legislation seems to me to be a blind process, and not much different from that which one calls stubbornness, which means: *stat pro ratione voluntas* [let my will stand in place of reason]."[57]

Pistorius demands a reason for an agent to subject themselves to the moral law. If we accept that the law the will gives itself is the moral one, this question exhibits a kind of radical moral scepticism which Kant was likely not interested in addressing. The *Groundwork* speaks to an audience which already accepts the moral law, but worries about its metaphysical foundation.[58] In the second *Critique*, Kant further cements this anti-sceptical stance by introducing our consciousness of the moral law as a "fact of reason", that "forces itself upon us of itself as a synthetic a priori proposition (…) it is not an empirical fact but the sole fact of pure reason which, by it, announces itself as originally lawgiving (*sic volo, sic jubeo* [thus I will, so I command].)"[59] Thus the normative force of the moral law is ingrained in reason itself, and no further reason can or need be given for it. It is worth noting the Latin quote at the end of this passage, which is taken (albeit misquoted) from the very same line in Juvenal from which Pistorius quoted: "*hic volo, sic iubeo, stat pro ratione voluntas.*"[60] Given how clearly the review was on Kant's mind in writing this book, this is unlikely to be a coincidence. It seems that Kant is here mocking Pistorius' demand and takes the fact of reason to show just how misguided it is. Where Pistorius invoked

---

[55] Interpreting *GMS*, 4:431, he states: "we decide to side with morality *because its law is essentially our very own law*." Timmermann, *A Commentary*, 103.
[56] "If – as a consequence of the will's autonomy – we know that a certain law applies to the human will, our job is done." Timmermann, "The law and the good", 12.
[57] Pistorius, "Groundwork", 36
[58] Timmermann, *A Commentary*, 129-130.
[59] *KpV*, 5:31
[60] Juvenal, *Satire VI,* 223. In *The Satires*, ed. John Ferguson (Bristol: Bristol Classical Press, 1979).

Juvenal to underline that the will cannot count as a reason, Kant uses him to restate that normative authority rests in the will itself.[61] While I share Timmermann's reading completely, it is unsatisfying to read Kant as merely refusing to engage with the question. There must be some sort of argument for *why* the moral sceptic is not worth addressing. I will return to this question later, but first I will consider a completely different approach to the Pistorius problem.

1.3 Value-realist solutions

Above, I took for granted Kant's claim that the good could only be determined through either the law, or the faculty of (dis)pleasure. This premise is often rejected by proponents of a more value-realist interpretation of Kantian ethics. Hills states that "it is not value that [Kant] rejects in favour of principle, but conceptions of the good that are based on happiness or desire."[62] Guyer calls Kant's disjunctive inference "invalid, because it fails to admit the possibility that there might be an object of the will that is not suggested by contingent inclination but that is in some sense necessary – that is not, as it were, suggested by the lower faculty of desire, but by a higher faculty of desire."[63] Most value-realists, including these two, take Kant's insistence that "rational nature exists as an end in itself"[64] to mean that this object is rational nature, humanity, and/or freedom.[65] The moral law is then taken to have its basis in promoting this object as an independent good.

Rejection of the premise does not come cheap, however, within Kant's wider system. A higher faculty of desire in Guyer's sense could only be higher in the sense that the representations it connects with the feeling of pleasure have their source in the understanding rather than in the senses. However,

[61] I thank Jens Timmermann for pointing out to me that these were in fact two halves of the same quote, and the potential significance thereof.
[62] Alison Hills, "Kantian Value Realism", *Ratio* XXI(2) (2008), 191
[63] Paul Guyer, "The form and matter of the categorical imperative", in *Kant's System of Nature and Freedom* (Oxford: Oxford University Press, 2005), 152
[64] *GMS*, 4:429
[65] These terms are all related but not identical, and which one a realist picks as her ultimate value will affect her theory significantly. I do not think anything hangs on it here, however, and will not differentiate various types of value realism.

the understanding in Kant is by itself a mere cognising faculty. It does not have any ability to determine choice.[66] It is still only through the mediation of pleasure that we are motivated to will the good. The only real higher faculty of desire is that which is determined by merely formal laws of the will, i.e. pure practical reason.[67]

This leaves open the possibility that the good is a motivationally inert concept, cognition of which only translates into action due to the mediation of reason or pleasure. But by what faculty is it cognised, if not reason or the senses? Only one remains: the understanding, which is "the faculty for bringing forth representations itself, or the spontaneity of cognition."[68] However, the understanding cannot by itself cognise objects. For this it requires the input of intuition. And "intuition can never be other than sensible, i.e. (…) it contains only the way in which we are affected by objects. (…) The understanding is not capable of intuiting anything, and the senses are not capable of thinking anything."[69] Kant therefore rejects the possibility of any kind of moral intuition.[70] It is through how the good affects us motivationally that it is cognised:[71] it is "a necessary object of the faculty of desire (…) in accordance with a principle of reason."[72]

There is good reason to hold that the good can be cognised only motivationally. Kant clearly states that the good is an object of *practical* cognition. Practical cognition is knowledge of what *ought* to be, and thus intrinsically carries motivational force. If we were to cognise something as motivationally inert, that is, theoretically, by definition that thing could not in itself be the good. It would merely be some object, cognition of which somehow activated our motivational faculties. This is problematic in

---

[66] *KpV*, 5:23
[67] Ibid., 5:22
[68] Kant, *KrV*, A51/B75
[69] Ibid., A51/B75
[70] Also seen in his rejection of moral sense theory in *KpV*, 5:38. See also Beck, 128.
[71] A similar argument is made in Oliver Sensen, *Kant on Human Dignity* (Berlin: Walter de Gruyter, 2011), 26; and in Silber, *Kant's Ethics*, 274.
[72] *KpV*, 5:58

two ways. Firstly, the connection between cognition and motivation is unclear on this account, and seems to involve a naturalistic fallacy. Secondly, even if the fallacy can be avoided this good cannot be an adequate basis for the moral law. This law would not be the will's own, but would be given to it by this mysterious object external to it; it would be heteronomous, rather than autonomous lawgiving.

1.4 Good as a priori concept of practical reason

We have so far treated the good as conceptually secondary to the moral law. Stephen Engstrom points out that this reading is not warranted by the text. "Kant does not claim that the moral law is prior to the *concept* of the good. He only says that this concept must not be *determined* [*bestimmt*] prior to the moral law but only after it and through it."[73] This leaves open the possibility, which Engstrom defends, that the concept 'good' exists prior to and independently of the moral law.[74] It is what he calls an "a priori concept of practical reason".[75] To say that the good is an a priori concept of practical reason combines two claims about this concept. It is firstly to claim that it gains its content from the form of our cognitive capacities itself, rather than through experience. It is secondly to identify practical reason as the particular capacity whose form gives it said content.[76] This claim is much more controversial than Engstrom appears to give it credit for. The exact phrase "a priori concept of practical reason" is never

[73] Engstrom, 179

[74] Talk of 'after' and 'existence prior to' might bring to mind temporal priority. It is worth clarifying that this cannot be what Kant has in mind, since both the good and the law are present from the start in the reasoning process of the agent. Nor is this a claim about conceptual priority; it is not the case that analysis of the one concept (law) will yield the other (good) as an analysans. (One should note that proponents of the traditional reading of the paradox of method would in fact claim that the law has conceptual priority over the good.) Rather, this claim concerns deliberative priority: one cannot talk about or bring to mind the moral law without already having the concept of the good and/or having brought it to mind. In terms of conceptual priority, I understand Engstrom as claiming that both concepts are necessary for a full understanding of the other and that hence neither has conceptual priority over the other. I thank Ben Sachs for pointing out the potential confusion here and helping me clarify this point.

[75] Engstrom, 12

[76] See Reinhard Hiltscher, "Begriff a priori" in Marcus Willaschek et al. (eds.), *Kant-Lexikon* (Berlin: Walter de Gruyter, 2015), 1: 239-240.

used by Kant, and the most recent edition of the *Kant-Lexikon* does not include the good in what is meant to be an exhaustive list of a priori concepts.[77]

What Engstrom means by his claim is intuitively plausible, however. To him, practical reason is first and foremost a faculty for knowledge. As reason, it is in the business of making judgments, which are knowledge claims. As practical, its judgments cannot be descriptive ("what is"), but must be prescriptive ("what ought to be"). Adherents of this interpretation think of 'good' as having the same role in practical judgment that 'true' has in a Fregean account of theoretical judgment. On such an account, 'true' is a formal concept that is constitutive of judgment: to make a judgment *p* is to assert that *p* is true. Truth is not itself a predicate that can feature in a judgment, but is the "form of predication" shared by all judgments.[78] Analogously, the good is the form of predication of practical judgments: to judge that something ought to be is to judge that it is good. The good is an a priori concept because it is contained in the very form of the cognitive faculty of practical reason.

To see how this might fit with Kant's view, recall that the good is a necessary object of the faculty of desire in a rational being. That is to say, to rationally desire something is to subsume it under the concept of good – which is just a complicated way of saying that we desire things because we judge them in some way good, and are averse of things we judge bad. This 'good' need not yet be understood as moral, i.e. unconditional good. Practical reason is also in the business of making instrumental or prudential judgments, judging something good *for some end* or good *for one's happiness* respectively. This makes significantly more intuitive the claim that good and bad are the *only* objects of practical

[77] The four elements on this list are 1) the categories, 2) mathematical concepts, 3) ideas and 4) a priori concepts with a posteriori validity (Ibid., 239-240.) Out of these four, Engstrom's good would be best placed under the header of 'ideas', since Kant equates these with concepts of reason. ("A concept made up of notions, which goes beyond the possibility of experience, is called an idea or a concept of reason." *KrV*, A320/B377.) It appears to me that Kant's usage of idea, as a regulative concept whose objective practical reality is guaranteed only by the fact of reason, does not sit well with the role Engstrom has in mind for the good, as constitutive of practical judgment in general.

[78] Rödl, 139; see Gottlob Frege, "Der Gedanke" in Ignacio Angelelli (ed.), *Kleine Schriften* (Hildesheim: Georg Olms Verlagsbuchhandlung, 1967), 347.

reason, since now practical judgments based on inclination can also be seen to involve the concept of good. As such, good and bad are *necessary* concepts of practical judgment by virtue of the *form* of the will, i.e. of practical reason. They are therefore indeed *a priori* concepts of practical reason.

What Kant claims according to Engstrom is that without the moral law, good is an *undetermined* concept. It is indeterminate in the sense that objects are subsumed under it haphazardly, with no rule delineating necessary and sufficient conditions for its application. As we have seen, while happiness is necessarily judged good, it cannot provide such a rule: no object has a necessary connection to happiness, which is itself an indeterminate concept. These connections are merely empirical. This will not do: good is to be "appraised by reason and hence through concepts, which can be universally communicated, not through mere feeling, which is restricted to individual subjects and their receptivity".[79]

That the good must be appraised through universally communicable concepts makes sense if we believe Engstrom's claim that practical reason is a faculty for knowledge. For our practical judgments to count as knowledge, they must satisfy the formal conditions of cognition.[80] Engstrom spells out these conditions as the necessary agreement between judgments, which consists in the subjective and objective universal validity of these judgments: they must be valid both for all subjects, and for all objects falling under the same concept.[81] Such judgments are therefore made under what he calls the *presupposition of universality*: "that *it is possible for every subject with the capacity for practical knowledge to share (not only* in abstracto *but also in use) the practical judgment that every such subject is to act as determined in the particular judgment when in the conditions on which it is based*."[82] This presupposition is equivalent to the Formula of Universal Law: "act only in accordance with that maxim

[79] *KpV*, 5:58
[80] See Engstrom, 70.
[81] Ibid., 115-116.
[82] Ibid., 125-126.

through which you can at the same time will that it become a universal law."[83] If Engstrom's interpretation holds, we can understand moral requirements as grounded in cognitive requirements.

When reason gives itself the moral law, it is really imposing on its practical judgments a rule that allows for the consistent and coherent application, i.e. the determination, of the concept good. In asking whether we can will that our maxim become universal law, the categorical imperative prevents us from willing a contradiction. It also imposes on our practical judgments a kind of condition of publicity: all rational agents should be able to make the same judgment of goodness. What gives these impositions their normative import is that they are the conditions for practical judgments to constitute *knowledge*.

We are now in a position to make sense of Kant's attribution of unconditional goodness to the good will alone. A good will is a will whose judgments necessarily agree with the conditions of practical knowledge, as it is guided by respect for the moral law which expresses these conditions. If it were not, any agreement with those conditions remains merely accidental. This means nothing guarantees the correctness of our practical judgments, and we cannot rest secure in our pursuit of all other goods. It remains possible in theory that I live my life accidentally making only correct practical judgments and pursuing genuine goods in conformity with the moral law.[84] However, even so my judgments are in an important sense blind. I am unable to connect them into a body of practical knowledge, because I lack an understanding of their common concept 'good'. Lacking this understanding I also cannot fully guard myself against the possibility of practical error. I can sustain my agreement with the moral law (the legality of my maxims) only by hoping that my inclinations continue to push me in the right direction, and that I do not encounter a situation in which they tempt me to judge in a way that is inconsistent with the presupposition of universality. I must therefore necessarily judge it good to have a good will, as both the condition of all further pursuit of goodness *and* an instantiation of goodness in and of itself.

---

[83] *GMS*, 4:421.

[84] This might be the case for the "soul so attuned to compassion" of *GMS*, 4:398, who adopts a maxim of beneficence from the simple joy they naturally find in helping others.

While Kant never makes explicit a response to Pistorius' case of the interest-free being, which remained totally indifferent to the universalisation of a maxim to make false promises, we can now infer it from what was said above. If such a being is truly not taking an interest in anything, it is not engaged *at all* in the enterprise of practical reason. The conditions on the use of the concept 'good', i.e. the moral law, still apply to such a being due to the public nature of the concept. Since it does not make practical judgments, however, it simply never uses the concept of goodness. It could then also never act intentionally. Pistorius is right to say that the law cannot be applied to such a being, but surely, it is no great objection to a moral theory that it applies to agents only. This gives us a much more satisfying understanding of Kant's refusal to address the sceptic. Once a being with reason displays agency, it is necessarily judging those things it acts for to be good; once it has used that concept, it can be held to the conditions that determine its correct application, i.e. the moral law. A self-consistent sceptic would thus be a being which refused to be an agent, which if even conceivable is too marginal to address.

1.5 Conclusion

Pistorius challenged Kant to do three things: to give a definition of the good independent of the moral law, to explain why the moral law was a good principle, and to justify how the moral law could apply to a being independent of its interests, under the assumption that this being lacked even a cognitive interest in the truth. These challenges are all, implicitly or explicitly, addressed in the second chapter of the Analytic of Practical Reason. On a traditional reading of this passage, all three are rejected as essentially misguided. It is not possible to define the good independently of the moral law, because then it would have to be through the faculty of (dis)pleasure which cannot yield objectively necessary practical principles. The moral law is not a good or bad principle; it is simply that principle which we give ourselves, springing from our own autonomy, from which it derives its normative force. And Kant's moral philosophy is not meant to speak to a being which refuses to take an interest in morality.

I have shown above that none of these responses are entirely wrong, but they represent Kant as merely dismissing Pistorius' objections. This does not remove all of the force of the objections and does not do justice to Kant's evident concern with addressing them.

By contrast, Engstrom's interpretation provides Kant with the ammunition he needs to give a much more decisive reply. There is much to say about the good before the moral law is ever introduced: it is an a priori concept of practical reason, which as a necessary object of desire is involved in all practical judgment. However, this concept remains indeterminate because it cannot satisfy the demands of practical knowledge. It can only be determined by the moral law, which imposes on it the condition of being applied in a consistent and universally valid manner. Aspiring to a good will, determined by the moral law, is the only way in which we can guarantee that our judgments obtain the status of practical knowledge; that is, it is the only way for us to guarantee that we are pursuing the genuine good. And the interest-free being of the moral sceptic can be seen to be outside of the moral domain for a very clear reason; it refuses to engage in *any* practical judgment, and can therefore not even count as an agent. The moment it *does* decide to act, interested or not, the conditions of practical knowledge still apply to it willy-nillyf, and it must acknowledge their binding force on pain of relinquishing its own agency.

What the above has *not* shown, however, is the truth of Engstrom's central premise: that the good is the a priori concept of practical reason. This was taken for granted merely to demonstrate the force of the resulting conception of Kant's ethics. It is not at all obvious that it is correct, however, and it entails a number of further substantive and controversial commitments in the philosophy of action. In the following chapter I will evaluate its truth by tracing out some of these commitments, most notably the 'guise of the good', and attempting to find in Kant's theory of the will the resources to defend them.

Chapter 2: The good as necessary object of the will

The view that the good is the a priori concept of reason promises to deliver an elegant, and minimal, foundation for Kantian ethics which gives both the good and the moral law an indispensable role. Practical reason operates with the concept 'good' by virtue of making any practical judgments at all. The moral law is needed to provide this concept with the kind of content that could constitute knowledge, and operating with the moral law in view is a necessary condition of coherent practice.

At the basis of the view, however, lies the idea that all practical judgment – and hence, all action informed by reason – constitutively asserts the goodness of the action the agent chooses. That is to say, insofar as practical reason is involved in action an agent cannot act without judging the action good in some way. This thesis is known in recent literature as the 'guise of the good', and though it has historically been endorsed by a large majority of Western philosophers it has become the focus of significant controversy in the philosophy of action. Joseph Raz sets out the thesis as follows:

> *GG* "Intentional actions are actions taken in, and because of, a belief that there is some good in them."[85]

There are a number of salient features of this claim worth unpacking. Firstly, it remains ambiguous on what is meant here by good. This allows for, at the very least, a weak and a strong reading. On the weak reading, 'good' is taken in a general sense including subjective good, i.e. instrumental and prudential good. On the strong reading, which Raz himself does not intend here, we take good to refer to the objective good (where that would likely, but does not necessarily, mean the moral good.)

Secondly, whether read weakly or strongly, GG should be understood as a conceptual thesis[86]. It does not merely state the contingent fact that intentional actions are mostly done with some good in mind; it claims that there is a conceptual connection between the intentionality of an action and the

[85] Joseph Raz, "On the guise of the good", in Sergio Tenenbaum (ed.), *Desire, Practical Reason, and the Good* (Oxford: Oxford University Press, 2010), 111
[86] Ibid., 112

belief that there is some good in it. An action not performed under this belief could not count as intentional. Thirdly, this is a thesis about the belief of the agent, not about the actual goodness of their action. It is in principle compatible with the truth of GG that no good action has ever been performed, so long as agents were all acting under the mistaken belief that their actions had some good in them.

Fourthly, though the claim here is about intentional action, in the literature the guise of the good is encountered just as often as a claim about desire or volition. As will be seen below Kant also discusses it in these terms. I take it that Kant, as well as most other participants in this debate, would understand the claims to be equivalent. This assumes that there is a strong conceptual connection between desire and intentional action, such that an action is intentional only insofar as the agent desires to act in that way. Such a picture of action appears to be already present in Aristotle when he describes the guise of the good.[87] There is no reason to believe that Kant strongly diverges from this framework. He defines the faculty of desire in *KpV* as "a being's faculty to be by means of its representations the cause of the reality of the objects of these representations".[88] Desire is, in Engstrom's phrase, "efficacious representation"[89] and if we further assume that all intentional action involves acting on the basis of representations (which seems uncontroversial to me and would surely be Kant's assumption) it follows that these representations by definition must be desires. Insofar as abandoning this picture is a philosophically viable option, I take it that one thereby abandons much of Kant's theory of agency already. I will therefore bracket such worries in what follows and use volition, desire and intentional action somewhat interchangeably unless otherwise indicated.

---

[87] "Thus there is really only one thing that produces movement, the faculty of desire. If there were two such things, intellect and desire, they would do so in accordance with some form in common, and in fact it is not clear that the intellect produces movement without desire, wishing being a type of desire, and the movement produced by reasoning being invariably accompanied by that produced by wishing, while desire even in the face of reasoning produces movement, a type of desire being appetite." Aristotle, *De Anima* (*On the Soul*), trans. Hugh Lawson-Tancred (Middlesex: Penguin Books, 1986), 433a.

[88] *KpV*, 5:9n. An almost identical formulation occurs at *MS*, 6:211.

[89] Engstrom, 27

In the eponymous paper that popularised the term 'guise of the good', J. David Velleman has criticised it for portraying the agent as "let's face it, a square" and "ignoring those agents who are disaffected, refractory, silly, satanic, or punk."[90] Through various counterexamples, he – and others after as well as before him – seeks to prove that the thesis does not do justice to the complexity of our moral psychology, and cannot make sense of decision-making that falls afoul of the norm. Opponents of GG like Velleman do not, of course, deny that it describes the vast majority of human action. They rather deny that it can describe *all* such action. They also deny that it is a conceptual truth, and therefore hold that instances of actions that are not performed under the guise of the good can nonetheless count as genuine instances of intentional action. Adherents tend to respond by looking for a way in which the agent in each counterexample can still be said to act on some, obscure or confused, value commitment. If this proves impossible or is ruled out by stipulation, their other strategy is to argue for some reason why the counterexample is not a genuine exercise of practical reason – that it is either inconceivable outside of thought experiment, or does not qualify as intentional action.

In this chapter I examine whether, and what version of, the guise of the good can in fact be plausibly attributed to Kant. I begin by examining the textual evidence, focusing again on the second *Critique*. I demonstrate that there are at least two possible interpretations of this evidence, each of which yields two very different views on Kant's position. One implies only the orthodox, very weak thesis that the pure rational will is directed to the good. The other supports Engstrom's and Reath's stronger claim that in fact all rational volition, pure or impure, subsumes its object under the concept of (objective) good. This directly implies that we hold all our intentional actions to a rather high standard of universal validity.

Since the strong thesis is the more controversial one, I then outline its philosophical and textual presuppositions in greater detail. The aim here is to clearly get in view how this position generally seeks

[90] J. David Velleman, "The guise of the good", *Noûs* 26(1) (1992), 3

to explain cases of non-moral willing, i.e. those cases that fall outside of the scope of the weak thesis. I first deal with Kant's Incorporation Thesis, which at first sight may appear to support the strong thesis, and argue that it would be a mistake to read it as doing so. More plausibly, the explanation for non-moral action is to be found where Reath seeks to locate it: in Kant's theory of self-love and of self-conceit in particular, where self-conceit can be interpreted as a tendency to universalise self-love. This explanation as it stands is insufficient to cover the many varieties of non-moral willing. It does offer a model, however, for a moral-psychological explanation of these varieties. I introduce despondency, a concept mostly present in Kant's *Lectures on Ethics*, as a counterpart to self-conceit with a similar function of rationalising non-moral action and making one's maxims appear universally valid. Equipped with this understanding of non-moral willing, in the next chapter I will be able apply it and further explore it by examining various types of such willing.

2.1 Kant and the Guise of the Good

That Kant endorses some version of GG is obvious. He discusses it in the second *Critique* as the "old formula of the schools, *nihil appetimus, nisi sub ratione boni; nihil aversimur, nisi sub ratione mali* [we desire nothing except under the form of the good; nothing is avoided except under the form of the bad.]"[91] He states that this formula is "at least very doubtful if it is translated: we desire nothing except with a view to our *well-being* or *woe*, whereas if it is rendered: we will nothing under the direction of reason except insofar as we hold it to be good or evil, it is indubitably certain and at the same time quite clearly expressed." He accordingly defines the good as "the necessary object of the faculty of desire" and the bad of "the faculty of aversion, both, however, in accordance with a principle of reason."[92]

What is clear from these remarks is that Kant rejects (or holds 'very doubtful') a subjectivist version of the thesis, according to which we desire things only because they are good for ourselves. Such

[91] *KpV*, 5:59
[92] Ibid., 5:58

a theory of desire would do no justice to the experience of duty, as an unconditional command that overrides self-interest.[93] Instead, Kant seems to say that our willing is always directed to the (moral) good or evil. Read without qualification, this claim seems even more wildly implausible: we very often desire things that are not morally good or that we do not view as morally good. Comparing the two interpretations Kant gives of the scholastic formula, however, shows two important differences in their scope. Firstly, the first claim concerns desiring [*begehren*] while the second merely speaks of willing [*wollen*]. Secondly, the first claim is made unrestrictedly while the second applies only to willing "under the direction of reason [*nach Anweisung der Vernunft*]." To understand exactly what these restrictions amount to is to understand Kant's position regarding GG.

Kant uses 'desire' in a broader sense than 'will' throughout his works. The faculty of desire [*Begehrungsvermögen*] as we have seen before is defined as "the faculty to be by means of one's representations the cause of the objects of these representations."[94] Desire does not by itself depend on reason and can wholly be determined by sensible representations, particularly pleasure. The will is "[t]he faculty of desire whose inner determining ground, hence even what pleases it, lies within the subject's reason" and is identified with practical reason.[95] Kant's denial of the subjectivist claim is not intended to deny that we very often desire things merely for our own well-being. That this would be true of all desiring, however, is denied by the second claim: the possession of practical reason means that determining grounds outside of our own well-being are available to us, with practical reason directing the faculty of desire to the good.

It is strange for Kant, however, to speak of 'willing under the direction of reason'. This would seem to be tautological, since willing is *defined* as desire directed by reason and the will itself is *identified with* practical reason. It is not obvious what willing 'not under the direction of reason' would

[93] See *GMS*, 4:400, and *KpV*, 5:86.
[94] *MS*, 6:211
[95] Ibid., 6:213. Though these definitions occur in later work, they correspond to Kant's usage in *GMS* and *KpV*.

amount to. Perhaps Kant merely intended to emphasise this point so that the distinction between *wollen* and *begehren* would not be lost on his audience. Nevertheless it is odd for him to praise as 'quite clearly expressed' a proposition containing what he should have recognised to be superfluous, and possibly misleading, elements.

Another possibility is that Kant is here still conflating will (*Wille*) with power of choice (*Willkür*). In later work these two will be disambiguated, and it becomes clear that *Wille* is what commentators have called the *legislative* faculty while *Willkür* is the *executive* faculty.[96] "Hence the will [*Wille*] directs with absolute necessity and is itself *subject to* no necessitation. Only *choice* [*Willkür*] can therefore be called *free*."[97] *The will* cannot give any law but the moral, since no other law has its determining ground solely in reason (i.e., in the will itself.) To will the law is not yet, however, to choose it. *Willkür* is free in that it has the ability to choose maxims in conformity with this law, but does not do so necessarily; it can also let itself be determined by the inclinations, and form maxims that take no account of the law instead. This allows Kant to reconcile the claim that the moral law is the will's own, and that every rational being wills it for themselves, with the fact that they very often choose to disregard the law.[98]

At the time of the second *Critique* however, Kant has not yet made the distinction verbally explicit and is still using *Wille* for both faculties, claiming both that it is a capacity to free choice and that it is directed wholly by the moral law.[99] This suggests that by "willing under the direction of reason" he is

---

[96] See f.i. Wuerth, 241, and Andrews Reath, "Did Kant hold that rational volition is *sub ratione boni*?", in Mark Timmons and Robert E. Johnson (eds.), *Reason, Value and Respect: Kantian Themes from the Philosophy of Thomas E. Hill, Jr.* (Oxford: Oxford University Press, 2015), 9.

[97] *MS*, 6:226

[98] This tension in what he called Kant's dual notions of "'Good' or 'Rational Freedom' and (…) 'Neutral' or 'Moral Freedom'" was famously pointed out by Sidgwick. Henry Sidgwick, "The Kantian conception of free will", *Mind* 13(51) (1888): 407. See for a discussion of the scholarship that arose in reply to this objection Wuerth, 239-240.

[99] A similar point is made in Silber, *Kant's Ethics*, 66-69. Silber sees Kant as working towards a solution to the problem posed by his *Groundwork* view of the will, namely that only autonomous willing can be viewed as free, in *KpV*. It is only a "partial solution", however, and raises new problems only solved once *Wille* and *Willkür* are adequately distinguished. Jens Timmermann has told me that he instead believes the development to be verbal rather than philosophical, as Kant already appears to have the distinction in mind in the *Groundwork* but to express it with increasing clarity over time. Either view seems to be consistent with what I argue here.

prefiguring his later distinction. He wants to make very clear that he was here referring to what would later be *Wille* proper, that is the will as lawgiving faculty wholly directed by reason. The claim Kant views as 'indubitably certain' would then be quite minimal: that the will as pure practical reason necessarily aims at the good.

By contrast, he had claimed earlier in the book that a "pathologically affected will of a rational being" is capable of forming maxims that conflict with "the practical laws cognised by himself."[100] Hence this will is capable of determination both by the law it gives itself, and by external forces (inclinations). I take it that the will as described here is what would later become *Willkür*, and that Kant is using phrases like 'pathologically affected' and 'under the direction of reason' to distinguish very heterogenous phenomena which he is increasingly realising cannot fit under a single concept 'will'. The conflict he describes would not be possible unless the pathologically affected will/*Willkür* were able to formulate maxims that were not directed to the good, in opposition to the laws given to it by the will. So Kant's later endorsement of the scholastic formula is not intended to extend to *Willkür*, and does not imply an agreement with the stronger GG given above.

I think this picture, though not explicitly argued for, is something of an orthodoxy in Kant scholarship.[101] The pure will is the good will, and so trivially aims at the good; the imperfect, human *Willkür* need not. However, the quotation leaves room for an opposing interpretation. Kant says "we will nothing under the direction of reason except *insofar as we hold it to be* good or evil" [emphasis mine]. Holding the object of desire to be good or evil is a significantly weaker condition than the object *being* good or evil, and allows for the possibility that in a large amount of cases we hold this belief mistakenly.

---

[100] *KpV*, 5:19

[101] It receives explicit defence in f.i. Stefano Bacin, "'Under the guise of the good': Kant and a tenet of moral rationalism", forthcoming in Violetta Waibel and Margaret Ruffing, eds., *Natur und Freiheit. Akten des 12. Internationalen Kant-Kongresses* (Berlin/Boston: De Gruyter), 6-9; and Thomas Hill, Jr., "Personal values and setting oneself ends", in *Human Welfare and Moral Worth: Kantian Perspectives* (Oxford: Oxford University Press, 2002), 263. I thank prof. Bacin for sharing his article with me.

Since the pure will is unfree to make such mistakes, this would suggest that Kant is here speaking of impure willing as well. If this is so my earlier suggestion that 'willing under the direction of reason' referred to *Wille* must have been mistaken, and Kant must instead be speaking of *Willkür* which is under the direction of reason insofar as it receives part of its incentives from *Wille* and is guided by prinicples of instrumental reason, particularly hypothetical imperatives (if you will the end, will the means).[102]

Thomas Hill, Jr. considers this possibility and rejects it. "Kant's phrase 'under the direction of reason' probably refers to the definite prescriptions of pure practical reason, not the qualified 'rules' and 'counsels' of instrumental reason. (…) hypothetical imperatives do not give decisive directions but always leave us the option of abandoning our ends and suspending a particular way of pursuing happiness, and so they do not give us unequivocal 'direction'."[103] The argument is true but does not seem to me to prove the point. Why does the direction [*Anweisung*] involved have to be definite and unequivocal? Just because I can abandon the end to which my hypothetical imperative is directed, does not mean it does not give me direction while I pursue that end.

Jens Timmermann has raised an objection to me along similar lines. A hypothetical imperative may urge me to perform a bad action, based on an end I hold. Timmermann argues that I still would not have reason to perform that action, because reason itself (in the form of the categorical imperative) speaks against it. To perform it regardless would hence be *against* the direction of reason. Surely, however, I would at least have *a* reason to perform the action, although this reason is subject to defeat by the moral reason counting against it. The bad action is evidently intentional, and intentional action is performed for reasons. Timmermann's assumption then appears to be that I act under the direction of reason only if I do what I have reason (i.e., all-things-considered) to do and not merely what I have *a* reason to do. This assumption strikes me as unwarranted. Pure reason and instrumental reason may

[102] See *GMS*, 4:417.
[103] Hill, Jr., 263.

conflict, and though what pure reason recommends is the rational choice this need not mean that reason is out of play once one fails to make that choice. (The rational choice is not to steal because this is evil. But to put it in extreme terms, once that bridge is crossed there is still a clear difference in rationality between carefully coordinating a break-in, and running head-first into the closed door armed with only a rubber chicken.) Again, that the direction reason gives is not unequivocal is not a reason not to see it as direction.

Similarly, it is unclear that "the faculty of desire (…) in accordance with a principle of reason" (of which the good is the necessary object) is supposed to refer to the pure will alone. Hypothetical imperatives are also principles of reason, but apply only to an impure will. Even if 'principle of reason' is here meant to refer to the categorical imperative, the human *Willkür* is always bound by the categorical imperative although it does not always heed it. It is not clear that that would not qualify it as a 'faculty of desire under a principle of reason'.

This dispute can be boiled down to one about two readings of the sentence "[t]he only objects of a practical reason are therefore those of the *good* and the *evil*." The first interpretation reads this sentence *de re*, as referring to what is *actually* good and evil. Since only a pure practical reason wills the good and avoids the evil without fail, the weak thesis follows:

WGG All pure rational volition necessarily aims at the good.

The second interpretation reads the sentence *de dicto*, where good and evil are concepts that collect all objects of practical reason. Hence, the claim is not that practical reason only takes the actual good and evil for its object; but that it takes things for its object only insofar as it takes them to be good or evil. This yields the stronger thesis:

SGG All rational volition necessarily aims at what it holds to be the good.

The theses are not supposed to be mutually exclusive, and in fact *WGG* is uncontroversial and should be accepted by adherents of *SGG* as well (but not vice versa.) What *SGG* actually amounts to as a Kant interpretation when fully spelled out is:

*SGG** All rational volition necessarily aims at what it holds to be the good,

where pure rational volition necessarily aims at the good,

and impure rational volition necessarily does not aim at the good.[104]

This interpretation would say that we only rationally desire things, i.e. intentionally perform actions, when we subsume them under our concept of (objective or moral) good. That is, we take ourselves to be morally justified in our actions. However, we are only correct in doing so if we are actually guided fully by the will, i.e. in cases of pure rational volition. Insofar as our desire, though involving reason, gives priority to sensible incentives we are mistaken. (SGG* does not yet make any claims as to how aware we may be of our mistake.)

It is important to distinguish this interpretation from the 'regular' GG. That thesis states only that intentional action/rational volition aims at 'some good', rather than 'the good'. That is to say that the value the agent sees in their action need not be objective value, and may be purely subjective value ('good for me'). This claim is equally compatible with (but not necessarily entailed by) WGG. It is likely

[104] By *Religion*, 6:22-23f, we can say "aims at the evil." Since the moral law is always an incentive in us, any act not motivated by it requires some form of resistance to the law, which is therefore evil; there is no morally neutral space in between. I choose to leave "does not aim at the good" in the main body of the text here because it is an equivalent formulation that makes clearer Kant's position in the debate on the guise of the good. His usage of 'evil', which deviates quite far from its current use, might obscure that position and make it sound like a much less plausible one. It is worth pointing out here, however, that Kant does not hold the somewhat implausible position that there exist no morally indifferent actions (*adiaphora*). *Adiaphora* are distinguished from forbidden and required actions as "merely permitted" actions in *MS*, 6:223. He explicitly answers the question "Are there, then, *adiaphora* as such?" affirmatively in the *Lectures*, naming actions performed for mere physical sustenance as an example. Kant, "Mrongovius", *Lectures on Ethics*, trans. Peter Heath, ed. Peter Heath and J.B. Schneewind (Cambridge: Cambridge University Press", 1997), 29:615. What Kant claims in the *Religion* is rather that volition itself, or the determination of the power of choice, cannot be morally indifferent and has to either be good or evil. An agent who performs a morally indifferent action from a good disposition (*Gesinnung*), i.e. an appropriate prioritisation of the moral law over inclinations and an understanding that the inclinations are to be restricted according to the demands of the law, is pursuing the good (because they act from the right motive); one who does so from an evil disposition pursues the evil, even though their action is the same.

that interpreters who deny that Kant holds SGG* will attribute the weaker GG to him instead. They would hold that desire and intentional action are directed *either* at the moral, or the natural good (pleasure/pain.) This goes back to Kant's view, discussed in part 1, that apart from the will itself only pain and pleasure can be an incentive to action. Note however that if GG is false, then *a fortiori* SGG* is false. In what follows, my primary concern is with evaluating SGG*. GG will be a useful baseline to invoke, however, since many counterexamples appear to equally put it in doubt. Often the most fruitful strategy will be first to see whether a certain counterexample can be understood under GG at all, and only then to see whether it can be further reduced to an instance of SGG*.

2.2. Willing under the presupposition of universality

The implications and commitments associated with SGG* have been most clearly set out by Andrews Reath. He equates the view that rational volition aims at the good with the view that "rational volition constitutively understands itself to satisfy a condition of universal validity. That is, rational volition is based on practical reasoning aimed at judgments of goodness that make a tacit claim to universality. (…) This amounts to the admittedly controversial claim that all rational volition is tacitly guided by something like the Universal Law version of the categorical imperative as its formal or internal constitutive norm."[105] To see how these formulations are equivalent, we should recall that Kant stipulated that goodness be a universally communicable concept appraised through reason. To judge that *x* is good is therefore to judge that '*x* should be done' is universally valid.

This view is partly motivated by the strongly intellectualist picture of choice Reath has outlined in earlier work, in which "one chooses to act on an incentive of any kind by regarding it as providing a sufficient reason for action, where that is a reason with normative force from the standpoint of others, not just that of the agent."[106] "This conception of choice presupposes that all rational action carries an

[105] Reath, "Did Kant hold that rational volition is *sub ratione boni*?", 6
[106] Andrews Reath, "Kant's theory of moral sensibility", in *Agency and Autonomy in Kant's Moral Theory* (Oxford: Clarendon Press, 2006), 18

implicit claim to justification."[107] He shares this intellectualism with Stephen Engstrom,[108] and for all intents and purposes in what follows I will take them to defend the same view unless otherwise indicated. Their position entails that we do not intentionally deviate from the moral law without in some way rationalising our deviation, and justifying it to ourselves and (hypothetical) others; finding reasons, however spurious, that we can make an exception to the law in this case and claiming that other agents would acknowledge this exception to be valid from their standpoint as well.[109]

It might seem as if this view receives textual support in the *Religion*, in what Henry Allison has called the Incorporation Thesis.[110] This thesis holds that "the freedom of the power of choice has the quite peculiar characteristic that it cannot be determined to an action by any incentive *except insofar as the human being has admitted the incentive into his maxim* (has made this a universal rule for himself, according to which he wills to conduct himself)."[111] Translated this way, the thesis seems straightforwardly to suggest that choice always involves setting up a universal rule, hence acting under the presupposition of universality. However, the German is much more ambiguous than this. What Pluhar renders as 'universal rule' reads "*allgemeinen Regel*" in German. *Allgemein* can be translated both as 'general' and 'universal', depending on the context.[112] An instructive comparison is its occurrence in the first formulation of the categorical imperative. Here "*allgemeinen Gesetz*"[113] is widely rendered "universal law" in English. After all, a law carries necessity: a law cannot merely apply generally to most cases, but must be equally binding on all subjects in the relevant domain (i.e. all rational beings) to be a genuine law at all. "General law" would hence be oxymoronic. Rules however have a different modal status. Rules may admit of exceptions and can apply to only one agent. I can

[107] Reath, "Kant's theory of moral sensibility", 19
[108] See Engstrom, 49-50.
[109] See Reath, "Kant's theory of moral sensibility", 20
[110] Henry Allison, *Kant's Theory of Freedom* (Cambridge: Cambridge University Press, 1990), 40
[111] *Religion*, 6:23-24
[112] I thank Jens Timmermann for alerting me to this linguistic point and for instructive discussion of it.
[113] *GMS*, 4:402

make it a rule for myself to play basketball three times a week, without thinking that the rule applies to you and expecting you to show up for practice. Hence a rule can be general without being universal, and the former translation seems more appropriate here than the latter.

This should come as no surprise, since the Incorporation Thesis equates *allgemeinen Regeln* with maxims.[114] Maxims are explicitly differentiated from practical laws at the start of the second *Critique*: maxims are subjective practical principles and apply only to the will of the subject, whereas laws are objective and apply to the will of every rational being.[115] Bacin has therefore objected to Reath that since maxims do not aim to satisfy a condition of universal validity, they cannot be understood as judgments of goodness. Since we act according to maxims, this means our actions do not necessarily aim at universal validity.[116]

The Incorporation Thesis then does not advance the case for Reath's view. Reath himself quotes as the most important textual evidence for his view this passage from the second *Critique*:

> "we find our pathologically determinable self, even though it is quite unfit to give universal law through its maxims, nevertheless striving antecedently to make its claims primary and originally valid, just as if it constituted our entire self. This propensity to make oneself as having subjective determining grounds of choice into the objective determining ground of the will in general can be called *self-love*; and if self-love makes itself lawgiving and the unconditional practical principle, it can be called *self-conceit.*"[117]

Of course, "primary and originally valid [*als die ersten und ursprünglichen geltend*]" does not obviously mean the same as 'universally valid.' The first sentence seems to concern the way the pathologically

---

[114] The equation already appears in *GMS*, 4:420-21n: "A maxim is the objective principle for acting, and must be distinguished from the *objective principle*, namely the practical law. The former contains the practical rule (…) and is thus the principle according to which the subject *acts*; but the law is the objective principle, valid for every rational being, and the principle according to which it *ought to act*".
[115] *KpV*, 5:35
[116] Bacin, 6-7
[117] *KpV*, 5:74

determinable self seeks to obtain primacy over the rational self, not whether its claims are taken as valid from the perspective of other agents as well. More promising is the claim in the second sentence that it makes its subjective determining grounds into objective determining grounds. Objectivity after all entails intersubjective validity, hence universality.

This passage is also important because it indicates the way we should understand the actions of agents who do not act in conformity with the moral law, in a manner that could preserve the guise of the good. All such agents, Kant seems to indicate, are acting under a maxim of *self-love*. This dichotomous view is further supported by Theorem II of the second Critique:

> "All material practical principles as such are, without exception, of one and the same kind and come under the general principle of self-love and one's own happiness."[118]

Chapter I saw some of the motivation for this rather stark view of human motivation. Kant need not deny that we sometimes act against our own interest for the sake of others, even in cases where the moral law does not demand it. However, what he denies is that interests outside of our own can have any direct bearing on our faculty of desire. For them to become motivating, they must first be related to our own faculty of desire by means of pleasure. This pleasure then allows us to view our happiness such that part of it is the pursuit of the ends of others, and hence we pursue those ends 'under the general principle of our own happiness.'

All immoral maxims are maxims of self-love, but not all maxims of self-love are immoral. Kant claims that:

> "Pure practical reason merely infringes upon self-love, inasmuch as it only restricts it, as natural and active in us even prior to the moral law, to the condition of agreement with this law, and then it is called rational self-love. But it strikes down self-conceit altogether, since all claims to

[118] *KpV*, 5:22

self-esteem[119] that precede accord with the moral law are null and quite unwarranted because certainty of disposition in accord with this law is the first condition of any worth of a person".[120] There is nothing wrong in principle with taking an interest in the satisfaction of one's inclinations, provided one gives due priority to the moral law. Hence maxims of self-love can be rational, and have genuine universal validity. When they do not, however, they are liable to lapse into self-conceit.

Reath explains self-love as "a tendency to treat one's inclinations as objectively good reasons for one's actions, which are sufficient to justify them to others." Self-conceit then is "a tendency to treat oneself or one's inclinations as providing reasons for the actions of *others*, or take one's desires as sources of value to which they should defer."[121] A self-conceited agent believes themselves worthy of a special kind of moral esteem, one that takes deliberative priority over the claims of the moral law. When their inclinations conflict with the moral law, therefore, they take themselves to be fully justified in the pursuit of their inclinations – to have a moral justification for their behaviour that could command the assent, and in fact support, of other agents.[122] Self-conceit then seems a perfect expression of the model of volition Reath proposes. Rather than be satisfied saying that they give priority to self-love, though they know it is wrong, the self-conceited agent engages in an elaborate exercise in self-delusion in order to be able to view their willing as universally valid.

Stefano Bacin objects to Reath's reading of self-conceit as making a claim to universality. According to Bacin the sense in which self-conceit is law-giving is not the same sense in which the

---

[119] I here correct Gregor's rendition of the German *Selbstsschätzung*, which she translates "esteem for oneself", at the suggestion of Jens Timmermann. The difference is subtle but important; 'esteem for oneself' suggests that one demands the esteem of others, while 'self-esteem' involves only the agent's view of themselves.

[120] *KpV*, 5:73

[121] Reath, "Kant's theory of moral sensibility", 15

[122] As always, these moral-psychological descriptions should be read with the opacity of maxims in mind. It is unlikely that a self-conceited agent would consciously assent to claims such as "my interests outweigh the demands of morality" or "you should place my wishes above your own." It might surprise them just as well to find that on close examination of their actions and practical reasoning, one can infer maxims that do accord with these beliefs rather than with a pure commitment to the moral law. Conscious and unapologetic self-conceit is possible, but extreme, and I suspect it will quickly look maniacal. For my purposes here it is not necessary to differentiate various levels of self-conceit.

categorical imperative is. Rather, self-conceit "gives a law by imposing the will of the subject on others."[123] This law does not aim at consistency, since the will of the subject "can change, as it is not governed by law in the first place."[124] I find this argument rather puzzling. From the fact that the self-conceited agent demands from others an undue respect, it does not follow that this demand is in fact law-giving. It merely indicates that *if* the agent had the power to impose their will as law on others, they would do so. However, lacking this power, the only agent on which they can impose that will is themselves. The universality of the claim consists in the *mere wish* that others would impose a law on themselves which afforded similar respect to this agent, and the belief that they would be correct in doing so. In this, it is not substantially different from the moral law which the agent can similarly take to be universal, while lacking the power to impose it on others. Taking self-love as an unconditional practical principle, surely the agent does take it as the law from which to derive their maxims, and to that extent self-conceit is genuinely law-giving. It is true that the will of this agent can change and is not governed by a true law. That a will can change, however, does not mean that it does not aim at consistency and takes itself to be making a genuinely universal claim at the time.

These considerations are supported by Kate Moran's thorough examination of self-conceit. Moran also holds that self-conceit need not involve giving law to others, and that "the fundamental failure of self-conceit is a failure in the way an agent addresses herself morally."[125] Self-conceit "does not make self-love the unconditional practical principle in the sense that it *imposes* an agent's self-love upon others. Rather, self-conceit makes self-love the unconditional practical principle in the sense that it reconstructs an agent's moral self-conception, including her capacity for moral self-evaluation, around self-love and heteronomy."[126] Moran also adds considerable specificity to the picture of self-conceit

[123] Bacin, 7-8
[124] Ibid., 8
[125] Kate Moran, "Delusions of virtue: Kant on self-conceit", *Kantian Review* 19(3) (2014), 438
[126] Ibid.f, 439

given by Reath and Bacin. She first describes it as "the tendency to construct a fiction that one is abiding by [the constraint of impartiality] when one has actually failed to heed it."[127] She makes a strong case that self-conceit appears throughout Kant's work as a tendency to overstep the boundaries of reason, and to use distorted reason to construct an illusion that the agent *cannot* make mistakes in their reasoning.[128] In the practical sphere, it is "the illusion that the agent is perfectly virtuous", which means the agent turns a deaf ear to the call of the moral law from the belief that *any* action they choose to perform must already have moral worth.[129] To believe one cannot make mistakes in their reasoning is to believe one's reasoning to be universally valid, even if the agent mistakenly believes the validity to derive from their own person rather than the nature of their actions.

The passage in which Kant speaks of the moral law 'infringing on self-love, but striking down self-conceit' juxtaposes rational self-love and self-conceit in such a way that it might appear that self-conceit is identical to, or co-extensive with, irrational self-love. That would imply that it bears the burden for explaining all cases of bad willing. This view would be highly implausible, since self-conceit is a very specific phenomenon that could not cover such a wide generality of cases. Many agents do not believe themselves perfectly virtuous even as they engage in bad willing; they just consider themselves justified in that particular instance. While undoubtedly many of us hold ourselves in excessive moral esteem, the opposite also seems possible. It must then be possible to engage in other forms of irrational self-love.

Moran points out that self-conceit has a counterpart in Kant's moral psychology: despondency [*Kleinmüthigkeit*]. Rather than elevating themselves above others, the despondent agent harbours "doubt as to man's capacity for ever attaining to the moral law, whereby we give up all effort to

[127] Moran, 422
[128] Ibid., 427
[129] Ibid., 438

approach it, and declare ourselves incapable of improving or elevating our worth."[130] While self-conceited agents ascribe a natural perfection of the will to themselves, despondent agents ascribe it only to others, with whom they compare themselves unfavourably. Since they share with their counterparts the tendency to think of virtue as a matter of such natural perfection, they give up on the very possibility of being virtuous themselves.[131]

We might imagine despondency to manifest itself in different ways. It can lead to a quiet resignation and inaction, from the sense that one 'can't do anything right'. But it may equally lead to an almost aggressive indulgence of one's inclinations, and a deliberate if angry (because one remains plagued by the pangs of conscience) rejection of moral norms. As a particularly poignant example, consider an agent who has grown up surrounded by gang-related crime and abhors it morally, but has internalised the belief that 'people like me' are not born for anything better and can't get by without participating in that system. We could well imagine such agents being so frustrated with their situation that, rather than confining themselves to whatever crime seems unavoidable, they adopt an attitude that 'nothing matters' and become vicious criminals, made more cruel by their loathing for themselves and their circumstances. (Of course, both attitudes may well occur over time in the same agent.) In the latter case they would still count as despondent, since they reject the moral law out of a sense of their own inferiority and give up the prospect of moral improvement. Since the moral law is rejected as an incentive, self-love would have to remain as the primary incentive in both these cases. Despondent agents follow their own inclinations, even if this is accompanied by frustration or self-hatred at their perceived lack of moral fortitude. Despondency therefore exemplify unreasonable self-love without self-conceit.[132]

---

[130] Kant, "Vigilantius", in *Lectures on Ethics*, 27:611

[131] Moran, 426

[132] Moran, interestingly, raises the question whether unreasonable self-love is possible without self-conceit and answers it affirmatively, but does not use despondency as an example. (In conversation on July 25th 2017, she agreed however that it would be a good example.) Instead she argues that "an honest assessment of one's

There is a plausible claim to be made that despondency, like self-conceit, involves a claim to universal validity on the part of the agent. At the very least we can see here an attempted rationalisation and justification: the agent seeks to justify the priority they give to self-love by lamenting the alleged defectiveness of their moral disposition and the difficulty of following the moral law. It seems unlikely that this justification is directed only at themselves, and that they would not attempt to excuse themselves from their moral duties from the perspectives of others in the same terms. (An agent who complained of the difficulty of being moral but did not believe this held up as a justification for not being so would more likely fall under the header of frailty, which will be further discussed below.) In claiming such a justification, the agent also appears to commit themselves to accepting that same excuse from other agents who are in a similar position to themselves. That is *not* to say that a despondent agent can no longer have moral expectations or hope for others, and would not be upset when faced with immoral behaviour. It is rather to make the claim that they would be disposed to excuse or understand the behaviour of the other if they understood that the other's thought process was similarly despondent.[133] Thus they, despite their mistake, continue to reason in accordance with a condition of universal validity.

The foregoing accounts of self-conceit and despondency have been in rather extreme terms, making it sound as if they are relatively stable features of a person's character. However, I do not think

---

wrongdoing is possible, even if it is difficult or painful" through conscience. (Moran, 439). This appears to me to prove that self-conceit can be overcome, but not that it is not involved in the wrongdoing to begin with.

[133] It is interesting to ask whether there is a genuine asymmetry here with self-conceit. On the one hand the self-conceited agent might seem disposed to grant the claims of others to excessive moral esteem, if he recognised or believed them to be in a similar position. We can easily imagine someone justifying their perceived right to additional esteem using phrases like 'for someone in my position' or 'people like me' (implying, of similar moral quality or talent). They here clearly imply that their status derives from properties that can be shared by others. However, self-conceit also implies a kind of arrogance and a strong disposition to rationalise. It appears to me that a self-conceited agent, when confronted with someone who is objectively in the same circumstances and of similar status as themselves, would likely try to find a way to rationalise that this person did not deserve the same level of moral esteem after all – their justification is, after all, spurious and they are driven by self-love which does not extend to other persons. The answer to this question appears to me to have no bearing on the overall point I am making, and I will leave it unanswered here.

they need to be seen that way. They are features of character insofar as they are predispositions to rely on certain explanations or excuses for our moral failure in certain situations, but it may be more helpful in what follows to view self-conceit and despondency as failures that can take place on the level of individual choices and actions. Otherwise humble and virtuous people may exhibit self-conceit in a moment of weakness, when they decide to indulge themselves and ignore their duties for a while. Insofar as they believe themselves justified in doing so they are behaving in self-conceit, and thus prove that this disposition exists in their character – even if we would strongly hesitate to refer to them as 'self-conceited people' on the whole. We can similarly imagine someone momentarily despairing of the possibility of doing one's duty and giving up, without completely renouncing their faith in the reality of virtue and the moral law. Their action is despondent even if their character may not wholly be.

2.3 Conclusion

The textual evidence appears to be indecisive at best as to what version of the guise of the good Kant personally held. The view that the pure will necessarily aims at the good is certainly his; but there is additional textual space to argue that he also held that the impure will takes itself, mistakenly, to do the same. This then leads to a view of willing as involving a strong presupposition of universality, so that agents act intentionally under the idea that their action is justified by reason and hence from the perspective of all other agents.

Both self-conceit and despondency are excellent models for an explanation of immoral behaviour on Reath's account, in that both are certain kinds of cognitive failures that ground a mistaken claim to universal validity. Self-conceit has agents unwittingly making themselves and their own self-love, rather than the moral law, the source of that universal validity. They not only make exceptions to the law for themselves, thereby treating it as a competing demand on par with that of self-love rather than as giving unconditional commands, but they do so in the belief that such exceptions are justified from the point of view of other agents. This happens because self-conceited agents becomes convinced

that virtuous as they believe they are, they cannot or do not make mistakes in their moral reasoning. The opposite holds for despondent agents: they hold themselves to be defective to the point of being incapable of moral behaviour, and hence stop trying. The end result is often similar: the moral incentive is deprioritised, clearing the way for self-love to determine the agent's behaviour. Agents believe themselves justified in this inversion of incentives by their alleged moral incapacity, and expect others to recognise this justification. Once a self-conceited or despondent agent has accepted these spurious justifications as universally valid, behaviour that from the moral point of view is evidently wrong may become entirely intelligible.

Though there is no reason in principle for this list to be exhaustive, we should expect that (with the deflated understanding of these phenomena outlined in the last paragraph of section 2.2) we could explain a large amount of moral failure in terms of these two. Therefore, in the section that follows, I will address various kinds of moral failure and try to map out the underlying cognitive failure. Again, if Reath and Engstrom are correct we should expect that all of them can be explained either on the model of self-conceit and despondency, or somehow shown not to be genuine instances of rational volition.

Chapter 3: Non-moral willing and the strong guise of the good

In the previous chapter I explored the general outlines of an argumentative strategy for Kantian defenders of the strong guise of the good thesis SGG*. They should argue that all rational volition involves a claim to universal validity, of the kind aimed at by the Formula of Universal Law. This involves agents seeking justification for their behaviour from the perspective of others, not merely their own. I suggested that self-conceit and despondency offered these theorists plausible moral-psychological models for the kind of cognitive error that might lead agents to make such claims falsely or mistakenly, and that hence these phenomena should be expected to rear their heads in a wide array of instances of bad willing.

In this chapter I apply that insight to actual cases of bad willing, in order to find whether and how adherents of the strong thesis are able to explain such cases. In coming up with these cases I have made use of Kant's own taxonomy of moral failures, as well as the various examples from the literature on the guise of the good. Though I believe to have explored the theoretical space that Kant's theory of agency offers quite widely, there is no way I could prove that this is in fact an exhaustive categorisation of the kinds of moral failure that are realistically or conceptually possible. I have no doubt that cases could be raised which do not neatly fit this list, though one might then argue about whether these cases are actually possible or correctly described. However, in considering the examples below I hope also to further develop and render more robust the picture of agency here offered, in such a way that extending that picture to new instances would require judicious extrapolation and application rather than revision.

I first look at diabolical willing, i.e. willing evil for evil's sake, which Kant holds to be impossible. I show that this position, which is not very well explained in the published work, both naturally follows from his thinking and fits ordinary moral consciousness. I then examine the case of Hitler, which Silber holds to be a counterexample. I argue that though the reality of Hitler (and similar historical figures)

does put pressure on Kant's optimistic theory of conscience, we have no reason to attribute to such dictators a truly diabolical will. What the example does show is that bad moral theory or ideology has the potential to override moral feeling, a position which I think is consistent with Kant's theory despite not being explicitly articulated in it. I conclude that all willing that appears diabolical can be explained instead as 'ordinary' evil willing, a phenomenon which Kant's theory is well-equipped to deal with and which I briefly include in the same section.

Secondly, I look at frailty, which is Kant's term for a phenomenon closely related to weakness of will or *akrasia* (incontinence). Frail agents will what is good, but fail to perform in accordance with their will. Hence their intentional action does not accord with their own conception of the good, and would seem not to fall under the guise of the good. Reath and Engstrom's response to this phenomenon is to emphasise the *Wille*/*Willkür* distinction and argue that since the will is not acted on by the frail agent, their action does not count as rational volition. I show that this response risks generalising over all non-moral willing and collapsing the strong thesis into the weak one. Engstrom avoids this risk by giving a different account of willing, emphasising that willing is about the adoption of maxims and the formation of character rather than about individual choices. His strategy confuses the notion of will, however, in a way unwarranted by the text. I argue that a different, and simpler strategy would suffice to deal with frailty: Tenenbaum's appeal to persistent illusion. Knowing that an action is not good does not necessarily dispel its appearance as good, and when appearances and moral knowledge conflict it is intelligible – though irrational – under GG for the frail agent to pursue the apparent good.

Thirdly, I look at the most difficult case and one Kant does not directly cover: the listlessness case. This case, inspired by but extending on a suggestion by Velleman, has the agent losing track of their sense of value altogether and acting intentionally, but without thinking of their action as good (in either the moral or the prudential sense.) Despondency explains some of this phenomenon, but cannot help recover it under the guise of the good. It is tempting to argue instead that an agent who reasons

without value conception is not in fact engaged in practical reason at all. This conceptual claim does not appear to capture the phenomenon well, however. I contend that orthodox Kantianism could treat listlessness as a degenerate case of practical reason, which it therefore does not have to account for in the same way it does for the 'natural' moral failures. While this response is only philosophically satisfying insofar as the orthodox picture is accepted, the strong thesis is at no disadvantage in explaining the phenomenon compared to other Kantian views.

3.1 Evil willing and diabolical willing

One of the main criticisms of the guise of the good is that it appears that there is also a 'guise of the evil.' That is to say, some philosophers contend that there are cases of deliberately evil willing, in which the agent knowingly performs an action *because* they hold it to be evil. Such an agent has what Kant would call a "malicious reason" [*boshafte Vernunft*] and becomes a "diabolical being."[134]

In this context, the literature often invokes Milton's Satan and his famous exclamation in *Paradise Lost*: "Evil, be thou my good!"[135] If we are committed to GG, we must say that Satan is simply rejecting the good because he thinks the evil is better – hence he is merely espousing a Nietzschean sort of view that what is normally considered good is not of genuine value, while seeing value in what is conventionally evil.[136] Satan therefore remains committed to acting "under the aspect of some good."[137] Velleman criticises this "rather sappy Satan"[138] because he has lost precisely what made him so striking: a knowing and willing commitment to *evil* for its own sake. "The ruler of Hell doesn't desire what he wrongly thinks is worthy of approval; he desires what he rightly thinks isn't."[139]

---

[134] *Religion*, 6:35
[135] John Milton, "Paradise lost", in Gordon Campbell (ed.), *The Complete English Poems,* (London: David Campbell Publishers, 1992), 225 (IV:110)
[136] This, as Velleman notes, is the answer offered in Elizabeth Anscombe's original discussion of the passage. G.E.M. Anscombe, *Intention* (Oxford: Basil Blackwell), 75. I use 'Nietzschean' here in the loose sense of a kind of thinking historically associated with Nietzsche; I make no claim, and would in fact seriously doubt, that this is the historical Nietzsche's real view.
[137] Anscombe, 75
[138] Velleman, 19
[139] Ibid., 18.

Kant himself appears to rule out the possibility of a diabolical will being instantiated by a human being:

> "To think oneself as a freely acting being, and yet as released from the law commensurate with such a being (the moral law) would be tantamount to thinking a cause operating without any laws (…): and this is contradictory. (…) a reason that absolves one from the moral law, a *malicious reason*, as it were (an absolutely evil will), contains too much, because the opposition to the law would thereby itself be elevated to an incentive (for without any incentive the power of choice cannot be determined), and thus the subject would be turned into a *diabolical* being. - Neither of the two, however, is applicable to the human being."[140]

Kant's argument is that the will, as a causal power, cannot act without a law governing its causation. It is not clear from this passage, however, why it could not renounce the moral law in favour of an opposite 'evil law'; why it could not make it its general maxim systematically to act in whatever way that the moral law forbids. In part, it seems as if this comes down to Kant's notorious moral optimism. Following the above passage he says:

> "The human being (even the basest), no matter in what maxims, does not, as it were, in a rebellious manner renounce the moral law (by revoking his obedience to it). Rather, the law thrusts itself upon him irresistibly, and if no other incentive acted against it, then he would also admit it into his supreme maxim as sufficient determining basis of his power of choice [*Willkür*], i.e., he would be morally good."[141]

For Kant, there is no renouncing or escaping the moral law. It always provides an incentive that is by itself sufficient to determine the human being to action.

[140] *Religion*, 6:35. The other possibility to which 'neither of the two' refers is that of the human being as merely animal being.
[141] Ibid., 6:36.

Even if this is accepted it still leaves unclear why this incentive could not be consciously resisted. It might be possible in principle for an agent to simply adopt a satanic maxim and ignore the moral law. Though this would not count as a complete rejection, it would amount to the same in practice. However, the question is whether we could make sense of an agent actually making this choice. It is not hard to see why someone might see some good in what is *de facto* evil, for instance with an eye to personal gain. This would however immediately disqualify them from being truly diabolical. For a diabolical will, the evil must *in itself* give the incentive to action.

This is absurd within the Kantian framework for a number of reasons. The first is a simple conceptual claim: the evil is defined as the necessary object of aversion, the good as that of desire. To desire the evil, therefore, is *ipso facto* to make it your good. This of course scans perfectly as an interpretation of 'Evil be thou my good': Satan cannot commit himself to evil without subsuming it under the good.

Of course, an opponent is unlikely to be persuaded just by an argument from definition; they may reply that the definition was simply stacked against them from the start. In response, a Kantian might show that there is no theoretical space for the possibility of the evil acting as an incentive. Recall that in Kant's theory practical reason only has two incentives: self-love (that is, the sum of all our inclinations),[142] and the moral law. For a diabolical will to be possible, it cannot be motivated by *either* of these incentives. We must therefore effectively postulate a third incentive, call it the evil law, which acts as a polar opposite to the moral law. We must also reproduce the great pains Kant took to demonstrate the possibility of practical reason being motivated purely by the moral law, this time for the evil law. Hence, we need a notion of 'anti-respect' for the evil law. We need a notion of 'anti-autonomy' to follow this evil law, different from heteronomy *and* different from Kant's notion of autonomy, which can only consist in following the law reason gives itself (i.e. the moral law.) To make anti-autonomy work we

[142] *KpV*, 5:73

might even need an 'evil reason' to give this law, as well as a 'fact of evil reason', etc. Though it would be marginally interesting to see how far one can get in working out this anti-moral system in Kant, I will not pursue these thoughts here. I think it is safe to say these notions would be extremely difficult to make sense of and would lack a clear basis in human moral experience. True evil, or the truest evil we can know in real experience, is to be conceived of in terms of heteronomy and hence in connection with sensible incentives to which the will yields; it cannot be pursued for its own sake.

These arguments rely quite heavily on the Kantian framework, but it is plausible to think that here this framework tracks the way action is conceived of in ordinary moral consciousness. This is to say that Velleman's Satan cannot exist outside of literature: to represent someone as truly committed to evil in this way is cartoonish, and for an actual agent with real practical reason to be committed in this way is inconceivable.[143] In fact, one might in this example say that Milton's Satan is such a compelling character *because* he is presented as having real, understandable motivations, and his motivations are understandable *because* he has a positive value conception he pursues.[144] 'True villains' in fiction are usually represented somewhat shallowly and without drawing much attention to their motivations. (The classic trope being 'mad scientist' characters whose goal is something as incoherent as 'to destroy the world': an aim which we would mostly associate with cartoon villainy.) To 'deepen' their character almost always means providing justifications for their motives, such that an audience can find it believable that in their own mind the villain is in fact pursuing some positive value conception.[145] It is

[143] Silber attributes this position to Kant. He appears to me misleadingly to suggest that Kant holds it explicitly. I find no evidence of this in the text and Silber does not produce any. Nevertheless, I agree with him that this is a position Kant should, and likely did, hold given his theoretical commitments. Silber, *Kant's Ethics*, 110-112.

[144] Satan says that "to do aught good never will be our task/ But ever to do ill our sole delight/ As being the contrary to his high will/ Whom we resist." (Milton, 153-154 (I:159-162)). It is certainly true that Satan is knowingly and willingly committed to doing moral evil, but only under the guise of resistance to God's will. This suggests that he does attribute positive value to such resistance.

[145] There are exceptions. Cathy Ames from Steinbeck's *East of Eden* comes to mind as a well-developed character who has proven compelling to many, yet who can truly be said to have a diabolical will. (Steinbeck opens chapter 8, where Cathy is first introduced, with "I believe there are monsters born in the world to human parents." John Steinbeck, *East of Eden*, in Robert DeMott (ed.), *Novels 1942-1952* (New York: Library of America, 2001), 385.) It is noteworthy, however, that Steinbeck had to defend the very *possibility* of her character from those who felt it was

difficult to see how a character who was truly motivated solely by evil could be fleshed out in a believable way.

Silber suggests in his essay 'Kant at Auschwitz', however, that history does in fact feature such characters; and that it is a fundamental deficiency of Kant's moral psychology that he was too optimistic about human nature to account for them. He states that "Kant's theory can comprehend the motivations of an Eichmann, a functionary whose efficiency and zeal were motivated almost entirely by careerist concerns; but it cannot illuminate the conduct of a Hitler."[146] According to Silber, "in Hitler we confront not an absence of self-directed will but, together with Stalin, one of history's ultimate examples of focused malevolent volition."[147] Kant can only dismiss Hitler as a deeply irrational madman, but this does no justice to the fact that his irrationality is an expression of his freedom and that Hitler "consciously chose evil."[148] This would suggest, if true, that we must yet try to find room within Kant's moral psychology to do justice to the empirical facts.

Silber is undoubtedly correct that Kant did not imagine the monstrous evil of the Third Reich as he was writing his works (as few thinkers at the time could have.) I seriously doubt, however, that Hitler is so incomprehensible from the perspective of Kant's moral psychology. Of course, whatever can be said about Hitler's motivations is mere speculation and inference from his behaviour. (This is not only so

---

unrealistic. (See for this some of the reviews collected in Joseph R. McElrath, Jr., Jesse S. Crisler, and Susan Shillinglaw (eds.), *John Steinbeck: the Contemporary Reviews* (Cambridge: Cambridge University Press, 1996), particularly Mark Schorer's review "A dark and violent Steinbeck novel", 391-393. Schorer refers to "Cathy Ames, most vicious female in literature, whose story if we accept it at all, we accept at the level of folklore, the abstract fiction of the Social Theatre, of a Witch beyond women" and complains that "There are defects in Mr. Steinbeck's imagination, certainly. He has always been fascinated by depravities that he seems helpless to account for", 392.) That Steinbeck and some of his readers *did* believe in the possibility might be taken to count against the view I defend here. However, the fact remains that Steinbeck cannot make Cathy intelligible beyond simply asserting her satanic nature. Furthermore, he calls his character a 'monster' and suggests that her humanity is in some deep way deficient. Taken together with Schorer's remarks and the theoretical framework laid out above, I take this still to support the Kantian position that a diabolical will is a "transcendent superstition" that does not apply to the human will (Silber's phrase: *Kant's Ethics*, 111).

[146] Silber, "Kant at Auschwitz" in *Kant's Ethics*, 333

[147] Ibid., 332

[148] Ibid., 335

because he is dead: the opacity of maxims means that even the living Hitler could not have given completely reliable testimony as to his own motivations.)[149] What is clear is that Hitler displayed a near-maniacal determination and steadfastness in pursuing a goal that by all accounts, his reason should have told him was an evil one. Part of what makes him uniquely terrifying is that his evil was methodically thought out and calmly carried through. It was not the mere result of violent temper or a misguided sense of self-preservation, and clearly involved the use of reason. Yet Hitler's behaviour does not suggest that he suffered the pangs of conscience, or that his reason urged the moral law upon him.[150]

The memory of Hitler and other twentieth-century dictators may be one of the reasons contemporary readers, including myself, have trouble taking the 'most hardened scoundrel' passage in the *Groundwork* seriously:[151] that "there is no one, not even the most hardened scoundrel, if only he is otherwise in the habit of using reason, who – when one presents him with examples of [moral conduct] – does not wish that he too might be so disposed. But he cannot easily bring this about in himself, just because of his inclinations and impulses; while at the same time he wishes to be free from such inclinations, which he himself finds burdensome."[152] Hitler does not seem to conform to this model; in fact, there seems to be something perverted in imagining Hitler as a slave of burdensome inclinations, wishing but failing to be rid of them. So far, then, I am willing to follow Silber in thinking that Kant was too optimistic. As a matter of empirical fact, it appears possible to be a scoundrel so hardened (an agent so debased) that one does not wish to be otherwise; to have a somewhat stable and reasoned commitment to evil.

---

[149] See *Religion*, 6:20: "Now, through experience one can indeed notice unlawful actions, and also (at least in oneself) that they are consciously unlawful; but one cannot observe the maxims, not even always in oneself, and hence the judgment that the agent is an evil human being cannot with assurance be based on experience."

[150] I am no historian, of course, and am not familiar enough with Hitler's biography to know whether there is any evidence that he in fact ever did suffer pangs of conscience. I here assume the contrary for the sake of Silber's argument.

[151] Even Timmermann notes, for instance, that here "Kant's optimism is astonishing." Timmermann, *A Commentary*, 143.

[152] *GMS*, 4:454

At first glance, the psychological claim I am suggesting we abandon appears to occupy an important place in the structure of Kant's system. This is reaffirmed by Kant's descriptions of moral feeling and moral death in the *Doctrine of Virtue:*

> "[Moral feeling] is the susceptibility to feel pleasure or displeasure merely from being aware that our actions are consistent with or contrary to the law of duty. (…) No human being is entirely without moral feeling, for were he completely lacking in receptivity to it he would be morally dead; and if (…) the moral vital force could no longer excite this feeling, then humanity would dissolve (…) into mere animality".[153]

Denying moral feeling to characters like Hitler, then, means robbing them of their humanity entirely – and consequently absolving them of the duties that are associated with our humanity. This is evidently unacceptable. Crucially, however, moral feeling is supposed to come from awareness of our observance of duty. This makes it possible that an agent who is suitably confused about morality could have false awareness of their (non-)observance of duty and hence lack the appropriate moral feeling in a given case. The appearance of moral death then belies the underlying cognitive failure: one is simply seriously misguided about the *content* of the moral law.

It seems clear to me that at least a part of Hitler's moral failure is genuinely cognitive in this way. He quite probably sincerely held the beliefs upon which the Third Reich was based, including that certain races and groups had no place in the Reich and were undeserving of moral respect. Though these beliefs do not count as 'moral' in the thick, Kantian sense (because they lack universal validity), undoubtedly Hitler took them to have moral import and count as moral beliefs.[154] In acting on them he must therefore have taken himself to act for the sake of the good: bizarrely, in Hitler's eyes an inability

[153] *MS*, 6:399-400

[154] Or descriptive beliefs with moral import: for instance, one can be committed to the dignity of humanity but hold the descriptive belief that certain races lack humanity, and therefore fail to afford them moral consideration.

to carry out his plans because of moral compunctions would look like weakness of will.[155] We have no reason to assume, furthermore, that Hitler would not have displayed moral feeling correctly in cases that lay outside of the scope of his Nazi principles. As Silber notes, he was charming and attentive to those around him, especially to little children.[156] He reportedly abhorred animal suffering and there is little reason to think that he would not have been happy to have done a good deed or dejected at having failed to help a friend in need. Thus even Hitler is not morally dead: rather, he has engaged in a misuse of reason that, under the influence of heteronomy, has produced in him a false awareness of what he believes is the moral law and thereby numbed him to contradictory evidence from confrontations with the actual moral law.

This explanation attributes a strong role to an agent's theoretical beliefs about morality, including the power potentially to override moral feeling and conscience and deafen the agent to the call of the moral law. Of course, we should not conclude that this process will occur in any agent who holds non-Kantian moral beliefs – this would invalidate Kant's entire theory of moral feeling, by narrowing its application only to the small group of agents who already hold the correct moral theory. I take it to be restricted to a small class of people I will refer to as 'ideologues', where I intend the use of that term to be value-neutral. What distinguishes ideologues from other human beings is that they pay an extraordinary attention to their beliefs, and make a conscious and directed effort to bring their behaviour in line with these. An ideologue will actively seek to correct what they perceive to be an 'errant' moral feeling. This description will apply to the Nazi and the Stalinist, but may equally cover

[155] Silber credits Hannah Arendt for noticing this fascinating, if horrific inversion: "Evil in the Third Reich had lost the quality by which most people recognize it – the quality of temptation. Many Germans and many Nazis, probably an overwhelming majority of them, must have been tempted *not* to murder, *not* to rob, *not* to let their neighbors go off to their doom (..), and not to become accomplices in all these crimes by benefitting from them. But, God knows, they had learned how to resist temptation." Hannah Arendt, *Eichmann in Jerusalem: A Report on the Banality of Evil* (New York: Penguin Books, 1963/2006), 150.
[156] Silber, "Kant's Ethics", 331

certain (harmless) moral philosophers, whose moral theory produces counterintuitive results.[157] These agents may genuinely believe that their moral beliefs are well-reasoned, and that they cannot trust feelings that would inform them otherwise. Alternatively, in many of these cases there may be deeper personal motives that make it unthinkable for the agent to abandon their theory. The Nazi might have a lot to gain personally from operating as a cog within his society's system. The communist civil servant may be afraid of the consequences of even seeming to question party doctrine. The moral philosopher might be afraid to doubt the theory on which they have built their career. And quite likely, agents whose moral beliefs support particularly abhorrent deeds (like the Nazi) might be subconsciously terrified to confront the gravity of their own deeds and beliefs from the genuine moral standpoint, and therefore seek to numb their moral feeling so that they will no longer be prompted to occupy that standpoint. In the case of the moral philosopher, this may look a lot more benign – they may more or less consciously train themselves to align their response to moral situations with their theory, rather than their moral feeling. Note that all this is a form of self-conceit: ideologues place an inordinate trust in their own twisted logic over moral feeling and the dictates of reason, and render themselves insusceptible to the criticism of other members of the moral community which might have steered them back in the right direction.

This position is speculative on my part and not directly based on Kant's text. I do however think it plays a background role in some of Kant's writing. In the *Groundwork* he explains that "[common human reason in practical use] stands just as good a chance of hitting the mark as a philosopher can ever expect; indeed it is almost more sure in this than even the latter, because he can have no other principle, but can easily confuse his judgment with a host of alien and irrelevant considerations and deflect it from the straight course. Accordingly, would it not be more advisable (…) not to let

[157] I say 'certain' because it is not at all my intention to imply that all non-Kantian moral philosophers are ideologues – merely to point out that it is possible to be an ideologue in a field in which that will ordinarily be much less dramatic and harmful than it can be in the political or religious spheres.

[philosophy] lead common human understanding away from its fortunate simplicity for practical purposes"?[158] Of course, he answers 'no' to this question, since philosophy (in the form of a critique of practical reason) can provide the clarity we require to escape the 'natural dialectic' of practical reason, "a propensity to rationalise against strict laws of duty" from seductive inclination.[159] This assumes however that this philosophy is done right, according to the critical method; in fact Kant thinks that moral philosophy which does not rest on a metaphysics of morals, like his own, cannot even be real philosophy. Since it will inevitably intermingle pure and empirical principles it "infringes on the purity of morals themselves by this intermingling and proceeds contrary to its own end."[160] All this establishes two points. Firstly, that insofar as Kant thinks moral philosophy worth doing he is only talking about moral philosophy that begins from critical philosophy. Secondly, that Kant attributes to 'bad' philosophy a real power to corrupt ordinary moral thinking and confuse an otherwise sound capacity for judgment. Whether we should classify Nazi ideology or communist dogma as bad philosophy may be up for debate, but certainly they play a similar cognitive role and yield moral judgments.

The numbing effect of ideology can be powerful but there is no reason we should believe that it would be complete, and so we need not contradict Kant's insistence that no human being can be completely deaf to the call of conscience. This is why I spoke before of only a 'somewhat stable' commitment to evil. The most hardened scoundrel may not wish to be good, but neither can he completely insulate himself from conscience. It is always present in him dimly and, in the right circumstances, able to break through his mental defences and raise his awareness of his own wrongdoing to the level of consciousness. This is important because it entails two conclusions about the agent. Firstly, that all agents are *redeemable*: there is no evil so deep that a change of heart is impossible, though circumstances may make it very difficult. Secondly, as a consequence, that all agents

[158] *GMS*, 4:404
[159] Ibid., 4:405
[160] Ibid., 4:390

can be held *responsible*: they can never pretend that it was not within their power to realise that they were doing wrong. Kant explains in the *Religion* that conscience is "a consciousness that is by itself a duty" and that "one ought to venture nothing at the risk of its being wrong (*quod dubitas, ne feceris*!)".[161] He introduces his own example of an ideologue in the form of an inquisitor, who believes that faith requires of him that he condemn a heretic to death. Since faith does not admit of any evidence that could absolutely guarantee the truth of his convictions, the inquisitor cannot rule out the possibility of error in his judgment; but conscience does permit him to know for certain that putting the heretic to death would be morally wrong.[162] If he decides to condemn the heretic after all, he will be acting wrongly on a principle of *probabilism*, i.e. that an action is justified if one merely has the opinion that it may well be right, and therefore in direct violation of his duty of conscience.[163] Thus so long as an agent has access to conscience as a power of judgment, they have a duty to be conscious of its commands. The same applies to Hitler: though he may have been quite convinced that his actions were justified, Kant holds that conscience prevented him from ever being *certain*.

This conclusion is particularly important to adherents of SGG* because it gives them the tools to defeat a potentially dangerous objection to SGG*. Samuel Kahn argues that SGG does not allow us to distinguish "culpable and inculpable ignorance", where the former is a kind of ignorance for which the agent is or should be morally responsible. In Kahn's example, a doctor tells a medically untrained woman to administer penicillin to her sick friend. The woman does so diligently out of the good intention to help her friend, not knowing that the latter is in fact allergic to penicillin. The woman's ignorance is inculpable: insofar as she had an obligation to inform herself before treating her friend, her obligation as a layperson was to seek sound medical advice, which she did in good faith. The doctor's

[161] *Religion*, 6:186
[162] Ibid., 6:187
[163] See Jens Timmermann, "*Quod dubitas, ne feceris*: Kant on using conscience as a guide", *Studi Kantiani*, XXIX (2016), 162-166.

ignorance is culpable: as a medical professional, she had an obligation to inform herself about the patient's medical history. Wanting to just be done with this patient, she ignored that duty and is responsible for the resulting allergic reaction.[164] Kahn argues that on GG, we can explain the moral difference between the two cases: the woman did what she considered to be objectively good, while the doctor knowingly did what was only subjectively good in defiance of the known objective good.[165]

On SGG, however, we are forced to say that wrongdoing consists in misrepresenting an action to oneself as objectively good. The doctor did wrong insofar as she mistook her own interests for objectively good. But the woman similarly did wrong in mistaking the administering of penicillin for an objective good. Thus both were pursuing what they genuinely perceived to be the objective good. Thus, argues Kahn, SGG lacks the resources to distinguish their errors clearly, and appears to absolve the doctor from blame (or conversely implicate the woman just as much as the doctor.)[166] Kahn fails to appreciate the difference conscience makes for culpability, however. The woman has fulfilled her duty of conscience and can be reasonably certain her action is in fact right; the ignorance that leads to its wrongness is of a theoretical nature, and conscience could not have remedied it. (Unless one argues that *quod dubitas, ne feceris* also implies that she should have been conscious of a duty to completely verify the patient's medical history. In that case, however, the example collapses entirely since her ignorance would count as culpable, and SGG is also validated.) The doctor, by contrast, acted on a principle that is morally uncertain and failed to live up to her duty of conscience. Therefore though both represented their action to themselves as objectively good, the doctor (morally) should have known that this representation was at the very least uncertain and is therefore culpable for her ignorance in a way the untrained woman is not. So long as we are able to maintain Kant's fairly robust moral psychology,

[164] Samuel Kahn, "The guise of the objectively good", *Journal of Value Inquiry*, 47(1-2) (2013), 8
[165] Ibid., 10-11
[166] Ibid., 12

then, we need not worry so much about SGG* reducing all our errors to errors of mere ignorance and losing track of culpability.

All of the above clears the way for a general explanation of evil willing in Kantian terms. Evil agents cannot be motivated by evil alone; nor can they be considered merely and inculpably ignorant, even if they are strongly under the spell of ideology. Despite appearances, all are guilty of the same fundamental mistake: they have given motivational priority to self-love, rather than the moral law, and have built an illusion of justification on the basis of treating the resulting principles as universally valid. This means we may plausibly understand all evil agents as self-conceited or despondent in some way. Self-conceit and despondency both function to suppress conscience and moral feeling, but as we have seen they cannot bring about complete moral death. Hence the agent's error can also never be seen as innocent or as the product of delusion so deep that it is beyond the scope of moral responsibility and appraisal. Though in this section I have focused mostly on quite dramatic cases of evil, which have historically been viewed as giving Kant and the adherents of SGG* the most trouble, it should be noted that this explanation generalises to all actions in which the agent fails to accord due priority to the moral law, whether they be Hitler or someone who does not return a library book. What they have in common is that they treat the moral law as an incentive to be traded away against their inclinations, rather than as giving unconditional commands. Evil willing reduces, then, to an error in (practical) judgment that stems from self-delusion.

However, some agents do wrong despite having completely sound judgment. They understand their duties and the unconditional nature of the moral law, yet in moments of weakness fail to act in accordance with this understanding. Philosophers tend to refer to such cases of dissonance between willing and action as 'weakness of will', while Kant has preferred the term 'frailty'. I turn to this phenomenon in the next section.

3.2 Frailty

At the outset of this section it appears worth examining whether and to what extent the Kantian phenomenon of frailty is the same as that of weakness of will, and/or *akrasia* (incontinence) as these are discussed in the general philosophical literature. In his influential paper on the subject, Donald Davidson characterises weakness of the will (which he identifies with incontinence) as follows:

> "In doing *x* an agent acts incontinently if and only if: (*a*) the agent does *x* intentionally; (*b*) the agent believes there is an alternative action *y* open to him; and (*c*) the agent judges that, all things considered, it would be better to do *y* than to do *x*."[167]

This echoes a description Aristotle gives of *akrasia* (incontinence). The incontinent person in Aristotle "acts in one way though already persuaded that he should act in another."[168]

Compare this to the way Kant introduces the phenomenon of frailty in the *Religion*:

> "Willing I have indeed, but performance is lacking; i.e. I admit the good (the law) into the maxim of my power of choice; but this good, which objectively, in the idea (…) is an insurmountable incentive, is subjectively (…) the weaker (by comparison with inclination) when the maxim ought to be complied with."[169]

Kant's rendition of the phenomenon is in principle consistent with Davidson's and Aristotle's.[170] It is also, however, much more specific than theirs in its characterisation. In Davidson and Aristotle, the possibility is open that the agent is weak merely in failing to comply with their own subjective beliefs about the right course of action. By contrast, the Kantian frail agent wills the objective good. This is so because

---

[167] Donald Davidson, "How is weakness of the will possible?", in *Essays on Actions and Events* (Oxford: Oxford University Press, 2001), 22

[168] Aristotle, *Nicomachean Ethics*, trans. and ed. Roger Crisp (Cambridge: Cambridge University Press, 2000), 1146b

[169] *Religion*, 6:29

[170] By which I do not mean to imply that there is only one, or an uncontroversial, understanding of *akrasia* in Aristotle. References to 'Aristotle's position' here refer loosely to that position which emerges from the quoted passages of the *Nicomachean Ethics*, serving merely to indicate that such a position exists and has historic weight in Western philosophy. I take no stance on how it fits with Aristotle's other writings and whether the view should be attributed to the historical Aristotle himself.

Kant describes frailty directly in terms of the *Wille/Willkür* distinction.[171] As seen before, an agent's will is necessarily directed to the good. It competes with the incentive of inclination, however, in determining the power of choice. Frailty covers a select number of cases in which the agent is well aware that the moral law is objectively the strongest incentive, but still fails to make their choice accordingly due to the subjective force of inclination. (Where the agent lacks this awareness and treats the moral law as an incentive that can be weighed off against inclination, they have already lapsed into *impurity*.[172] In this situation the agent is vulnerable to self-conceit and despondency as I have analysed these phenomena above.) In this it is in line with the Aristotelian conception, in which the incontinent is explained as "pursu[ing] the present pleasure" in spite of their rational choice.[173] This underlines Engstrom's insistence that Kant's moral philosophy, as a species of practical cognitivism, is continuous with the tradition that stems from Plato and Aristotle rather than breaking away from it.[174]

Unsurprisingly but significantly, Kant's conception of frailty trades on his specific theory of the will as a faculty for moral knowledge. It does not cover instances of weakness of will in which an agent fails to make a choice that would be better according to a mistaken value scheme. Such cases would presumably have to fall under impurity, since the agent has already admitted non-moral incentives into their value scheme. Nor does it hold for cases that lack a clear moral element, such as Davidson's 'toothbrush case'.[175] And of course Kant's theory of choice as always being determined by one of two incentives directly rules out the possibility of 'incoherent incontinence'; that is, an agent intentionally

---

[171] This is another reason for me to stick to Kant's usage rather than adopting the more common 'weakness of will', a term which sounds misleading from the Kantian perspective as it locates the weakness in the wrong faculty.
[172] *Religion*, 6:30
[173] Aristotle, *Nicomachean Ethics*, 1146b.
[174] Engstrom, 25
[175] Davidson, 30. In this case Davidson is lying in bed when he realizes he has not brushed his teeth. Despite judging that the sensuous pleasure of staying in bed actually outweighs the negligible negative effect of skipping just one brush, he cannot help himself getting up to brush his teeth, citing his nagging sense of obligation. Since the obligation involved here is not of the moral kind, this is simply a case of competing inclinations. It would be an interesting exercise to see how Kant would explain this kind of prudential irrationality. However, as this would have no bearing on the guise of the good here under consideration, I will not further pursue this here.

acting against their better judgment without any discernible motive. (Listlessness, which I will discuss in the next section, also involves the agent not acting on either of these incentives; however, their action is not incoherent nor obviously lacking in motives, and so this claim should not be taken to extend to listlessness.)

The reality of frailty would seem to be one of the best arguments against GG, and SGG* in particular, since it involves acting intentionally against one's idea of the good. An adherent of GG could swiftly reply that, by Kant's own admission, the agent is still pursuing *some* value: namely subjective over objective value. There would be nothing incoherent about this in principle. This reply is quite unsatisfactory as it stands, however. Insofar as a choice appears as good, it does so by comparison to the alternatives of which the agent is aware: and in this situation the agent is quite explicitly aware of the superiority of the alternative. GG would be close to trivial if 'good' in it referred only to the *pro tanto* good, since there is probably some good to be found in almost *any* possible choice. Surely, what GG means to say is that an agent makes a choice in the conviction that their choice is 'good' in a more interesting sense: good in terms of being the preferable option, all things considered. Even if this is doubted, it seems to me that on a psychological level the extent to which a choice seems good is significantly affected by the value of the other choices which are ruled out by it, and that the frail agent does not view their choice as good in this way as they are making it.

Nevertheless this response can be developed into one with more merit, by appealing to Kant's distinction between the objective and subjective level. Frail agents are aware that the moral law should objectively be an insurmountable incentive for them. But in their subjective experience, it does not appear to them that way; there, inclination does in fact present itself as the bigger incentive, and hence its recommendations appear as a greater good. Sergio Tenenbaum argues that it is crucial to understand such 'appearances of the good' in a manner analogous to theoretical appearances. The latter appearances can not only be illusory, he points out, but the illusion can be persistent. In such cases,

even knowing that we are being presented with an illusion will not correct the appearance.[176] The lines in the Müller-Lyer illusion continue to appear of unequal length to us even when we are aware that they are equally long. Similarly, the inclinations continue to present themselves to us as making a claim to the good even after we have become aware that this claim is restricted by the moral law. This is also why the inclinations have their quality of temptation, and why reading Kant's philosophy does not have the effect of an immediate moral conversion.

Certain behaviours of an agent presented with a perceptual illusion seem to be intelligible even if irrational. For instance, we would not complain that a child's behaviour was unintelligible if it was transfixed in fear by a dark humanoid shadow in the room at night, even though it knows from daylight experience that it is only seeing a coat and hat on a coat rack. The child itself will likely recognise its fear as a kind of cognitive weakness, but feel powerless to stop it. Similarly, a defender of GG can claim, an agent can be intelligibly pursuing a mere appearance of the good in full knowledge that this is irrational. The illusion of goodness does not disappear, nor does it lose its motivating force, in the face of practical knowledge. So while the agent *knows* they are not doing what is all-things-considered good, it may still *appear to them* as if they are. Since both courses of action appear under the guise of the good, either choice is intelligible from the perspective of GG.

This strategy, though not completely worked out here, appears a very promising one to salvage GG. It is not evident that it can also rescue SGG*, however, due to its reliance on distinguishing the objective and subjective levels of value. SGG*'s most prominent adherents have sought to deal with the phenomenon of frailty in other ways.

Reath does not pretend to have a fully developed answer to what he calls weakness, and he somewhat confuses the matter by using this term and not clearly distinguishing impurity and frailty. He

[176] Sergio Tenenbaum, *Appearances of the Good: An Essay on the Nature of Practical Reason* (Cambridge: Cambridge University Press, 2007), 40

initially appears to suggest that, "rather than in a simple failure of choice (*Willkür*) to follow practical reason (*Wille*)"[177] it should be located in the cognitive dimension of volition. Weakness occurs when agents have unclear principles or reason defectively from their principle to its application. While this undoubtedly explains certain kinds of behaviour, it cannot apply to frailty. The frail agent is well aware of the correct principles and how to apply them. Reath appears to be aware of this and suggests the following:

> "A back-up possibility for handling certain kinds of weakness is to treat them as intentional action that is not the result of volition. In this respect, the thesis is less strong than it might initially appear since it need not imply that all intentional behaviour understands itself to satisfy a condition of universal validity. Such choices would be intentional in the sense that they issue from a faculty of desire – a capacity by means of one's representations to realize the objects of these representations. But they would not be rationally willed since they do not engage the distinctive capacity for rational volition, which initiates action from a principle understood to satisfy a condition of universal validity. (…) Instances of weakness understood in this way would not, strictly speaking, be "weak willing", but volitional failure – failures to exercise the will."[178]

Unfortunately Reath considers this only a 'back-up possibility', and his exploration of it ends with this passage. Stephen Engstrom appears to suggest a similar reply, though he also never quite appears to view frailty (which he also does not distinguish from general weakness of will) as a significant challenge to his views. For Engstrom, the possibility of weakness of will demonstrates the difference between willing and mere intending. Only willing, says Engstrom, understands itself as making a judgment that an action is good. Hence weakness of will involves a conflict between intention and 'one's better judgment' or will, in which "the will is thwarted or hindered by an opposing sensible desire; and

[177] Reath, "Did Kant hold that rational volition is *sub ratione boni*?", 49
[178] Ibid., 41-42

to the extent that the latter prevails in the conflict, action contrary to what is willed may ensue."[179] Engstrom still regards it as a cognitive failure, i.e. a deficiency in one's knowledge of the good.

Both of these responses capitalise on Kant's own application of the *Wille/Willkür* distinction, and argue that the frail agent satisfies SGG* because their willing (but not their choice) does in fact conform to a condition of universal validity.[180] As we observed above, the will of a frail agent is directed to the good, and recognised as having the appropriate authority. The failure occurs only at the level of *Willkür*. John Silber describes this failure as a bifurcation of the action into two acts of *Willkür*: one to choose the immoral course of action, the other to immediately condemn itself for doing so.[181] The latter act is occasioned by conscience and hence reflects what the agent actually wills: to adhere to the moral law.

This response is only successful, however, if – as Reath observes – the thesis is much less strong that it initially appeared. There is a strong feeling that goalposts are being shifted if now SGG* is suddenly taken to be restricted to the will. As noted before, the claim that the will (as practical reason) is necessarily directed to the good (WGG) is quite uncontroversial among Kantians, and it is evident that Reath and Engstrom have wanted to say something significantly stronger. Their claim rather seemed to be that any act of choice (intentional action) in which the will is at all involved takes the good for its object, and hence takes place under the presupposition of universality. Since, by Silber's analysis of it, the will is involved in frail agency it too should fall under this presupposition.

Reath of course does not strictly deny that the will is involved in frail agency. By his analysis, it rather 'fails to be exercised' and takes a back-seat to a "sub-agential system". The will is present in the

[179] Engstrom, 45

[180] This response is not yet clearly stated in their published work. However, I have learned in personal conversations in March and April of 2017 that both Engstrom and Reath have since become more aware of the problem posed by frailty to their theory. They both indicated that their present response would be along these lines.

[181] Silber, *Kant's Ethics*, 106

sense that the agent had the capacity to exercise it, but did not. Hence the volition of the agent does not qualify as rational volition, and SGG* applied only to rational volition.

I agree with Reath that frailty is a type of irrationality, which means that it cannot count as fully rational volition. I am not comfortable, however, with the claim that it does not count as rational volition at all. My worry is that this argument appears to generalise to all kinds of improper volition under discussion in this chapter: in all of them, the agent has a capacity to exercise the will but fails to do so for various reasons. While undoubtedly it is true that all these behaviours count as irrational in that sense, this cannot be sufficient to deny them status as rational volition. Otherwise, SGG* would again amount to WGG: the rather boring claim that purely rational volition aims at the good. It is already evident that Reath intends to defend a stronger claim than that.

Engstrom would reply that there is in fact a difference between frailty and impurity, such that Reath's point would not generalise over both. The latter he calls a "conflict in the will", the former "a conflict between practical thought and willing (practical judgment) proper".[182] What distinguishes them is that the will goes beyond making an individual decision, as one does when suffering from frailty. Rather, to will is a matter of adopting maxims and thereby "constituting one's character". The tension in frailty is precisely that the isolated action does not accord with the maxims the agent has in fact adopted, and hence with the structure of their overall character. That the will is to be understood like this is already suggested by Kant's discussion of the good will at the very start of the *Groundwork*.[183]

[182] I here base myself on notes from a conversation with Engstrom on March 17th, 2017, in which I have tried to capture his literal wording to the best of my ability.

[183] As Timmermann interprets it: "Kant is thinking not of isolated good intentions or individual actions but rather of good moral volition overall: a morally good character (which will then express itself in good acts of will, and consequently good action.)" Timmermann, *A Commentary*, 17. He is not very explicit about his reasoning, but presumably the textual argument would be that Kant speaks of the good will as something that can be "present" in a continuous sense and as something with which a being can be "adorned". Kant also compares a good will to other qualities of character such as wit and resolve. *GMS*, 3:393.

This view is consistent with Engstrom's conception of the will as a capacity for practical knowledge. Agents can know that *p* without consistently instantiating that knowledge. Gamblers may be familiar with all of the basic principles of probability (*p*) but still make irrational bets on the basis of a 'lucky feeling' or an irrelevant superstition, perhaps even while professing that they know the odds are against them. It is important to note here the presence of a conflicting thought *q* ('the roulette ball fell on black the last four times, surely it will be red *this* time!') which explains their behaviour, but which the gamblers likely would not count as knowledge (as it is neither justified nor true). One might imagine them saying things like 'I *know* it doesn't work like this, but…' and even being ready to dissuade a friend in the exact same situation from taking a bet based on *q*. Nor need we suppose that they have momentarily forgotten *p*, or are unaware of its conflict with *q*. We still would not take this behaviour as evidence that the gambler in fact did not know that *p*, or has a kind of 'cognitive character' that does not *really* know that *p*. We would, however, probably ascribe to them some kind of cognitive weakness – a type of theoretical frailty, that is susceptible to the attraction of irrelevant factors overriding the intellect's better judgment. Thus, we have a conflict between the agent's knowledge and their thinking, rather than a conflict between different knowledge claims. Engstrom's claim is that moral frailty should similarly be understood as a conflict between a good will, as a body of correct practical knowledge (i.e. a character) and an errant practical thought. This conflict is more or less transparent to the agent and while their thought is action-motivating, they do not take it to make a knowledge claim (have universal validity.)

By itself this explanation of frailty is plausible enough, and compatible with a view of will as a faculty that knows, and commands, the moral law alone. The same cannot be said for the contrast with impurity as a 'conflict in the will.' This requires the will being able to make conflicting judgments, and such a notion of will differs markedly from the one which contrasts with *Willkür*.[184] So far I have

[184] Engstrom does acknowledge and accept the *Wille*/*Willkür* distinction, however, in Engstrom, 49.

understood *Wille* as intrinsically guided by its own form, i.e. the form of the moral law, and as one of the two incentives of *Willkür*. Engstrom however appears to reverse their relation: for him "all exercise of the free power of choice [*Willkür*] is also exercise of the will", where he understands the will as the capacity for practical judgment.[185] As textual evidence he points to a puzzling passage from *MS* which occurs in the very same section in which Kant makes the *Wille/Willkür* distinction:

> "not only choice but also mere wish can be included under the will (*Unter dem Willen kann die Willkür, aber auch der bloße Wunsch enthalten sein*)".[186]

If Kant here really means to say that choice is a mere aspect or subspecies of will, this directly contradicts the preceding paragraph of his text. However, it seems likely that Kant means to make a quite different point than Engstrom needs. He introduces choice not by means of a direct contrast to will, but to *wish*: the faculty of desire *without* a "consciousness of the ability to bring about its object by one's action". The will is subsequently defined as the "faculty of desire whose inner determining ground (…) lies within the subject's reason". The point Kant wants to make is not that all choice is will. It is rather that will can "determine the faculty of desire as such" and hence can be an incentive to wish as well as to choice. Choice is 'included' under the will insofar as it is one of the two sub-faculties of desire under the will's influence.[187] So while will is involved as an incentive in every exercise of choice, it is not itself exercised with every choice. It is exercised only in those choices that are rational, i.e. moral. Hence will proper, contra Engstrom, cannot be in conflict with itself; and all forms of moral failure, not just frailty, are conflicts between the agent's choice and their will.

Of course Engstrom is still correct that the difference between frailty and impurity lies in character: the frail agent has a good character but fails to live up to it at all times, while the impure agent treats the moral law as an incentive on par with the inclinations, which is a severe character flaw.

---

[185] Engstrom, 65
[186] *MS*, 6:213
[187] Ibid., 6:213

The difference lies not in the will, however, but rather in the "dispositional act" by which these agents prioritise the two incentives of *Willkür*. The concept of disposition (*Gesinnung*)[188] emerges in the *Religion* in the discussion of good and evil character, and signifies an agent's most general or 'supreme' maxim. One's disposition is "the first subjective basis for the adoption of maxims",[189] and can only be good or evil. A good agent gives priority to the moral law, an evil agent to self-love. The dispositional act is that act by which *Willkür* adopts one or the other supreme maxim.[190] The dispositional act of choosing the good maxim does not guarantee the goodness of all of the agent's particular acts, but neither can these acts completely fail to relate to the disposition.[191] (Hence the guilt and self-condemnation associated with frailty, which is absent for an agent who acts from impurity.) Since the dispositional act is an act of *Willkür*, the distinction between frailty and impurity should be drawn in almost the opposite way to Engstrom's. Both are conflicts between an act of *Willkür* and the will. However, in frailty *Willkür* is also in conflict with itself: an individual act of choice conflicts with the dispositional act. If one wanted then to preserve Engstrom's reply and say that the kind of rational volition to which SGG* applies is that which constitutes character, this amounts to claiming that SGG* only extends to the dispositional act. This view would be significantly weaker than the one Engstrom actually proposes, and could hardly be said to support the claim that the good is the *a priori* concept of all practical reason.

So it appears that Reath's and Engstrom's explanations of frailty fall short. Frailty may differ from impurity and other kinds of moral failure, but it is still rational volition and hence should be covered by SGG*. I propose that they instead look to Tenenbaum's argument from persistent illusion as an answer. To briefly recap that argument, Tenenbaum suggests that actions which are commended by the inclinations appear to us as good, and they continue to do so even after we discover that this

---

[188] Pluhar translates it 'attitude', but I here prefer to remain consistent with Silber's more expressive rendition 'disposition.' I believe the point does not depend on the specific translation preferred.
[189] *Religion*, 6:25
[190] Silber, *Kant's Ethics*, 101; see *Religion*, 6:31.
[191] Silber, *Kant's Ethics*, 101-102

appearance is illusory (i.e. that the moral law in fact forbids these actions.) A tension then develops between what we know to be good and what appears to us to be good, and this tension need not be resolved in favour of knowledge.

One reason an adherent of SGG* might be sceptical of this argument is that they are required to argue that appearances of the good make a claim to universal validity, and continue to do so in the face of a contradictory and rationally more compelling claim. I fail to see, however, why the argument cannot be transformed to reflect this in a plausible way. Of course that which appears as good to the agent is only subjectively good, and the agent is aware of this fact. The agent however also has a good disposition, i.e. prioritises objective over subjective good. If the subjective good appears to agents only as subjective, it is not clear why it could tempt them to act contrary to their disposition and against the moral law. If, however, the appearance of good gives the illusion of objective goodness (even against the agent's better judgment) this temptation is less mysterious.[192] The agent is weighing off two appearances of objective goodness, one veridical but the other with inclination (hence subjective goodness) on its side. The latter presents itself to him or her in errant thoughts which seek to rationalise their choice, such as "you know you deserve this", "everyone would understand" or "they would do the same in your shoes." While these thoughts are contradicted by the agent's judgments they still have the force of the illusion which they help sustain. Thus frail agents are still engaged in rationalising their

[192] Kant talks of "the inner practical illusion of mistaking a subjective element in the grounds of action for something objective" in the *Anthropology*, but this does not seem to refer to illusions of *goodness*. Rather, the illusion lies in agents mistaking the subjective objects of their imagination for things in themselves. Kant thinks this somewhat confusing description applies to children playing soldier, but also to gamblers. However, that these illusions concern the "grounds of action" still suggests that the illusion concerns a value judgment and not merely a perceptual one. Kant, *Anthropology from a Pragmatic Point of View*, trans. and ed. Mary Gregor (The Hague: Martinus Nijhoff, 1974), 7:274-5. There is similarly a passage in *KpV* where Kant speaks of "an optical illusion in the self-consciousness of what one *does* as distinguished from what one *feels*"; this appears to refer to the fact that there is a contentment with oneself associated with acting for the sake of the law, and that this feeling of contentment can appear to be the incentive for moral behaviour where in fact this incentive is the law itself. *KpV*, 5:116. Importantly, therefore, my usage of 'illusion' here does not correspond to Kant's and there is no direct textual evidence for the idea of persistent illusions of goodness. I do take it however to be consistent with Kant's theory and to plausibly follow from it.

choices and presenting these to themselves as objectively justified; but they do so in full, if momentarily suppressed knowledge that this justification is actually spurious and will not quell their conscience for very long.

One important difference should be pointed out between ordinary, i.e. perceptual, persistent illusions and persistent illusions of goodness. The former occur in relatively rare and special circumstances, as our perceptual capacities are well-equipped under normal conditions to a) clearly perceive appearances and b) adjust themselves to an illusory appearance in such a way that the illusion is broken, as soon as the illusion is seen through once. By contrast, on this theory we must hold that 'practical' persistent illusions occur under normal conditions in great number, essentially whenever an agent is tempted by a subjective good. This disanalogy is consistent, however, with Kant's view of human nature and practical reason. Practical reason differs from the theoretical, after all, in being subject to two incentives pulling in different directions, and even the highest degree of virtue attainable to a human being does not release them from the pull and temptation of the sensible incentive.[193] Hence the appeal to persistent illusion appears fully consistent with Kant's general theory of the frailty of our nature.

3.3 Listlessness

Unlike evil and frailty, listlessness is not a familiar concept from Kant's writings on moral psychology. To my knowledge, in the general literature it also does not go by any widely agreed upon name the way "weakness of will" does. The term 'listlessness' occurs in work by Alfred Mele[194] and I have found it to be the most descriptive of the phenomenon in question, though I will deviate somewhat from his treatment of it.

[193] *MS*, 6:383

[194] See Alfred R. Mele, "Motivation: essentially motivation-constituting attitudes", *The Philosophical Review*, 104(3) (1995), 403, and "Internalist moral cognitivism and listlessness", *Ethics*, 106(4) (1996), 732-734. I thank Nastia Grishkova for being the first to suggest this term to me, independent of Mele's work.

Essentially, the phenomenon I have in mind occurs when an agent appears to act in ways that do not clearly reflect *any* sense of value in their action. One such case is one of Velleman's counterexamples to GG, which I will dub the 'crockery case'. In this case Velleman has fallen into a deep despair and convinced himself that it is futile to act for any value whatsoever. As a result he stops leaving his house and refrains from any action to pursue his conception of the good. However, he does not stop acting altogether. He may be taken by a fancy to smash some crockery in his house. Not, Velleman insists, because he think this will somehow help alleviate his despair; he is so dejected that he may not even see this as possible, and if he does acknowledge the possibility he is too depressed to act on it. It rather strikes him that this act of crockery-smashing would be such a worthless thing to do, and he sees this as fitting his present self-conception as worthless.[195] One might dismiss his action as nonsensical, but this is the very point; with all the bearings of practical reason and intentionality he commits himself to an action which he knows to be nonsensical, because it is nonsensical.

Less frivolously and more realistically, we might look to cases of clinical depression.[196] I want to be very careful here not to make any general claims about this condition, which is obviously complex and manifests itself in very different ways for different people. First-hand accounts suggest, however, that some patients experience their actions during a depressive episode in a way that may put pressure on GG. They describe losing track of the very idea that their actions might have value, and simply 'doing things'. (References to clinical depression in what follows are intended to be exclusively to agents who fit this description). GG would then seem to imply that, since intentional action is defined by the pursuit of good, depression renders them incapable of intentional action. This is a very strange (and possibly offensive) description of what is going on. Patients are still choosing to do the things they do, with an intention of doing them. While undoubtedly many of their actions are performed out of habit and to

[195] Velleman, 20-21
[196] Mele does so in "Internalist moral cognitivism and listlessness", 733.

that degree 'mindlessly', they are not automatic and are, or can be, the result of some deliberative process. There is simply a certain kind of answer to the 'why'-question that they cannot give, namely one in terms of value.

In the Kantian scheme, listless agents appear not to be motivated by *either* of the incentives of practical reason. They are despondent and do not take an interest in the moral law, but neither can their behaviour be said to be motivated by self-love. (This is not at all to imply that the clinically depressed are not morally motivated, as they very often are. Kant's own example of the misanthrope, who feels no sympathy for others but continues to act beneficently to them out of sheer duty,[197] could with minor alterations be seen as such a case. Here I am rather thinking of Mele's example of Eve, a woman who sees herself as morally obligated to care for her sick uncle. When her husband and children die in a plane crash Eve becomes clinically depressed and as a result, though she continues to believe she is morally required to help him, she entirely stops being motivated to do so.[198] Mele does not say that she stops helping her uncle as a result; assume here that she would.) Like self-conceit and despondency, I take it that listlessness need not be a complete condition so that these incentives are not exhibited in *any* of the agents' behaviour. An agent is listless insofar as they perform *some* actions intentionally without viewing them as good. Listlessness is the disposition to perform actions in this way. In the sorts of cases I have here used as examples, we could expect this disposition to manifest itself in quite a high number of the agents' actions – though it appears difficult to imagine it doing so with complete consistency. I will leave undecided here whether isolated listless actions in an otherwise motivationally healthy agent are possible.

By way of a first reply, the defender of SGG* might wonder whether genuine listlessness is possible. Maybe Velleman does not view smashing his crockery as useful or productive, but it is not

[197] *GMS*, 4:398
[198] Mele, "Internalist moral cognitivism and listlessness", 734

entirely random that *this* is the action he chooses to perform. Such an action comes from an urge; and though this urge is itself random and irrational, choosing to satisfy it is everything but. Raz suggests that an urge inherently provides a reason to satisfy that urge, and thus that – absent stronger reasons against such action – an agent pursues genuine value in satisfying even an irrational urge.[199] This claim has a good deal of intuitive evidence for it. The satisfaction of an urge tends to produce immediate satisfaction, and conversely the failure to do so can lead to a nagging feeling that is both unpleasant, and interferes with other pursuits. Raz further argues that the fact that agents do not feel that their action has such value is due to their being "conceptually confused about reasons":[200] they believe, mistakenly, that there is no reason to satisfy urges, presumably because they confuse the irrationality of the urge itself with the irrationality of choosing to satisfy it. That they still act as they do reveals that their surface-level beliefs do not match their motivations and their actual value commitments. Nothing about this story seems to me inconsistent with the basic Kantian picture of agency. That picture then further requires us to believe that, when an agent intentionally acts to satisfy an urge, the incentive at play is self-love – a supposition that seems perfectly reasonable, insofar as the satisfaction involved could be counted under the header of pleasure.

With a sufficiently developed notion of what an urge is, such a story would seem to extend relatively well to the clinical depression case. Agents in this scenario report feeling like they are on 'auto-pilot', being driven to do things quite normally but having lost a clear sense of why. It is plausible that much of what they do is done out of habit. We could then further say that part of the phenomenon of habit formation is that it creates urges in the agent to conform with their habits. An agent once had a justification which they took to be sufficient to begin performing a certain action regularly; the resulting urge provides this agent with a reason to perform this action even in situations where the initial rational

[199] Raz, 125
[200] Ibid., 125

support is absent, either because the situation is itself different or because the agent no longer takes that support to be sufficient.[201] The latter description is particularly likely to hold for clinically depressed agents. Habit then gives us a rational origin story for their urges that prevents us from painting them as completely erratic agents who are the puppets of whatever random urge seizes control of them. At the same time, the crockery case shows that random urges *can* still occur – there is no reason we should completely exclude them from our account.

There are two main problems with this story. The first is that it is uncomfortably speculative. We essentially posit an, unfalsifiable and unverifiable, urge for any behaviour that we cannot otherwise explain, alongside the unsubstantiated claim that *all* behaviour without a higher-level explanation can be explained with urges in this way. The second, and worse is that it seems that it can bear the weight of GG but not SGG*. Agents who decide to give in to an urge do something that is good for themselves, i.e. provides them with pleasure in the form of satisfaction. There is no indication however that they represent their choice to themselves as universally valid, and no obvious reason that they might. One might say that they would believe that 'I should satisfy my (harmless/morally unproblematic) urges' is a universally valid judgment, and this may be correct. But we have got to this point by arguing that the agents are conceptually confused about urges; they do *not* see themselves as intentionally giving in to an urge and rather believe there is *no* reason (in terms of value) they do what they do. The judgment is never on their mind to be universalised.

If listless action does not involve a claim to universal validity, the other option for adherents of SGG* is to deny that it is proper rational volition. There is, after all, a kind of pathology here that prevents practical reason from functioning properly. This is of course not to say that listless agents lack practical reason or do not make use of it. It is rather to say that certain generic claims about practical reason cannot hold for them.

[201] One might note that this story fits Davidson's toothbrush case, referenced in fn. 174.

An obvious reply to this would be to say that frail and evil agents exhibit similar defects in their practical reason. However, a major difference is that Kant views both propensities as wholly natural and as intrinsic to the human being.[202] The human will, as that of a finite being, is necessarily influenced by inclination which gives rise to these volitional failures. By contrast, listlessness is an extraordinary state that does not belong to the essence of the human will and in fact goes directly against it. As such, an account of practical reason and of moral psychology owes more of an explanation of evil and of frailty than it does of listlessness. Compare it to a theory (rulebook) of basketball: this requires answers to personal fouls that naturally occur on the field, as players exert their bodies in space in imperfect ways, but not to teams showing up to a game with only a single member. (Of course, there is one kind of blanket answer in the rulebooks: such a team forfeits the game, much like a listless agent may be said to forfeit the game of practical reason, and as a result the rest of the rulebook fails to find application.)

This argument has force, and may well be sufficient to relieve Kantian defenders of SGG* of their duty to respond to the listlessness challenge. One remaining source of unease is that it merely asserts Kant's picture of the human being, without further motivating it. One might well argue that a propensity to listlessness is present in every human being, or is at least a natural feature of the will to which various humans can be subject to various degrees. The fact that clinical depression is a recognised pathology counts against that claim, but not very strongly. Listlessness may occur in less severe or pathological forms, and many if not all humans may have a propensity to certain forms of depression depending on their circumstances. The Kantian view of human agency may just be an inadequate portrayal of the complexity of our inner lives.

Be that as it may, this reply still puts Engstrom and Reath in a better position than initially thought. The remaining objections now count against Kantian views of agency generally, not SGG* specifically. As it is not the purpose of this dissertation to evaluate those views as a whole and as these

[202] See *Religion*, 6:29-33

objections require significant amounts of speculation, I will not pursue them further here. My conjecture is that listlessness is a phenomenon that generally puts pressure on traditional accounts of agency including Kant's, especially insofar as they are committed to the guise of the good.

3.4 Conclusion

My contention, having considered all these cases of volitional failure, is that SGG* works. That is to say, there are ways to explain volitional failure that are consistent with the claim that rational volition constitutively aims at the good, and always represents itself as satisfying a condition of universal validity. Granted, explaining all volitional failure in this way requires considerable speculative effort and, though I have tried to show the contrary where possible, does not always yield the most intuitive or obvious picture of what is going on. Its descriptions in isolation would not be sufficient to recommend SGG* over a more relaxed view. The question is first and foremost whether SGG* is an acceptable consequence of a theory which otherwise appears to have considerable interpretative and philosophical merit. In the case of listlessness, we have seen that SGG* stands on equal footing with other (Kantian) views, as it rejects the burden of having to explain the phenomenon by denying it the status of normal or natural rational volition. For evil willing, it paints a quite compelling picture of agents being led astray by self-conceit and despondency into feeling genuinely justified in their immoral pursuit. And frailty could be explained in terms of the agent failing to rid themselves of a persistent illusion – not, as Engstrom and Reath would have it, as a degenerate case of rational volition or a conflict of will with itself.

Of course, what will be most objectionable to many philosophers is not so much the substantive explanations offered by SGG* for the other cases, but its stringency in insisting that these explanations must hold at all times. Surely *some* villains will conceive of themselves as pursuing some twisted version of universal good, but why would we believe that *all* of them do? Is it not more parsimonious in isolated

cases to view them as simply failing to care about doing good and consequently doing wrong, perhaps not as an end in itself but therefore no less deliberately?

The answer to this question probably remains yes. What I hope to have shown, however, is that such answers are not helpful in making this behaviour intelligible in general. Lacking complete transparency of motive, there is unfortunately no way to know for sure whether at some level the presupposition of universality plays a role in all practical judgment. If we suppose that it does, however, the resulting theory is successful at delivering a unified explanation in which all morally problematic actions are rendered intelligible. This then means that we need no longer be concerned that counterexamples prevent us from adopting the conceptually attractive claim that the good is the a priori concept of practical reason, with all its positive implications for the force and justification behind Kant’s ethical theory.

Concluding remarks

This dissertation asked the Pistorius question 'what does Kant mean when he talks about the good?' It has by no means given a definitive, or even a complete, answer to that question. I have not examined all of the possible answers in equally great detail and have completely avoided talking about the highest good, an important notion in the second *Critique* which is the locus of endless controversy but undoubtedly has to be a part of the complete story about the good. Instead, having unfolded Kant's dialectic with Pistorius, I was rather quickly side-tracked by what I hope to have shown was the most intriguing possibility, namely the view that the good is the a priori concept of practical reason. It is necessarily involved in all practical judgments, and to be understood as making a claim about the action the agent intends to undertake that purports to be universally valid. This view has the potential of justifying the normative power of the categorical imperative in terms of generally accepted normative principles about knowledge, thereby also providing a more solid justification of Kant's moral anti-scepticism. Most importantly, it addresses Pistorius' challenge in a way that avoids the problems arising from making either the good or the law primitive, by presenting them as mutually independent but supporting concepts that both play an indispensable role in our agency.

The basic idea of the good as a priori concept is the foundation of a very complex picture, which I have only partly been able to elaborate here. I have focused on the kinds of explanations it can give of moral failure, for the simple reason that I take this to be the most obvious objection to the theory and one that directly relates to our topic of the relation between law and good. It was also a topic where I felt much productive work was, and of course still is, to be done given the somewhat sketchy explanations Engstrom and Reath have so far provided.

I took it as my aim here to vindicate SGG*, not because I necessarily agree with it, but in order to see whether it could be done at all. While my conclusion is that it can, the reader may understandably not be willing to follow me in all the steps and philosophical sacrifices I had to make in order to arrive at

that conclusion. Insofar as a case is made here for Engstrom's view it is an indirect one: the view can consistently be held in the face of certain objections, and is an attractive one in understanding Kant against the backdrop of the historical context. I also hope to have shown that the view requires emendation and elaboration at least on certain of its details, as the canonical version from Engstrom and Reath does not get the phenomenon of frailty quite right and is at best implicitly adequate in its treatment of evil and listlessness.

Above all, however, I hope to have shown that Kant's view of the good was significantly more sophisticated and frankly, interesting than the deontological caricature has made it seem. As a result it is also significantly better able to stand up to modern philosophical scrutiny. The Pistorius question to Kant was not a trifling one, and one whose answer affected the very foundations of his ethical system. Kant did not mean to displace the good or to dislodge it from its central role in ethics, even if perhaps it is also not the case that his theory should be seen primarily as an analysis of it like Moore's is. It is a theory of agency, above all, locating the good in an examination of the desires and actions of rational beings. It also offers a quite developed moral psychology to explain and understand these desires and actions, and the relation they bear both to the agent's perceived good and to the objective good.

Aside from all the ways I fear I may have misconstrued and misrepresented arguments for and against from the secondary literature, a reconstructive project like this one always raises the worry of having added material which the historical author himself could not have provided or would not have, if presented with the same problems. While I have partially tried to account for that when discussing my methodology, I have little doubt that I may have gone too far in places and attributed to Kant views that are distinctly foreign to him, if they are even philosophically plausible. I only hope to have paid enough attention and respect to the texts that any such additions may be close to their spirit and therefore acceptable as parts of a still-Kantian view. Despite my best efforts I am also conscious of my inadequate knowledge of the literature in both pre-Kantian and modern theory of agency, of which I have only been

able to sample so much in support of this dissertation. As a result, some of the arguments here may cover ground that feels old to those more familiar with that literature, or have failed to take account of important objections and advances made therein. In the worst case the arguments may be rendered irrelevant as a result, or be of interest exclusively from an exegetical point of view.

Despite all these inadequacies, I flatter myself that, while I leave it to the reader to judge my treatment of these issues, this investigation has at least led me to examine some aspects of Kant's moral theory and psychology that are very worthy of further research. I am thinking here of such topics as the role and relation of self-conceit and despondency, the power of ideology, the implications of frailty and the phenomenon of listlessness. I also believe the dialectic with Pistorius remains a fertile ground for scholarship, as other interpretations may still be found that shed a different light on a rather dense, mysterious, and curiously neglected section of the *Critique of Practical Reason*.

Bibliography

Allison, Henry. *Kant's Theory of Freedom*. Cambridge: Cambridge University Press, 1990.

Anscombe, G.E.M. *Intention*. Oxford: Basil Blackwell, 1963.

Arendt, Hannah. *Eichmann in Jerusalem: A Report on the Banality of Evil*. New York: Penguin Books, 1963/2006.

Aristotle. *De Anima* (*On the Soul*). Translated by Hugh Lawson-Tancred. Middlesex: Penguin Books, 1986.

Nicomachean Ethics. Edited and translated by Roger Crisp. Cambridge: Cambridge University Press, 2000.

Bacin, Stefano. "'Under the guise of the good': Kant and a tenet of moral rationalism". Forthcoming in Violetta Waibel & Margit Ruffing, eds., *Natur und Freiheit. Akten des 12. Internationalen Kant-Kongresses*. Berlin/Boston: De Gruyter, 2017.

Beck, Lewis White. *A Commentary on Kant's Critique of Practical Reason*. Chicago: The University of Chicago Press, 1963.

Beiser, Frederick C. *The Fate of Reason: German Philosophy from Kant to Fichte*. Cambridge, MA and London: Harvard University Press, 1987.

Davidson, Donald. "How is weakness of the will possible?" In *Essays on Actions and Events*. Oxford: Oxford University Press (2001): 21-42.

Dieringer, Volker. "Was erkennt die praktische Vernunft? Zu Kants Begriff des Guten in der *Kritik der praktischen Vernunft*". *Kant-Studien*, 93(2) (2002): 137-157.

Ebels-Duggan, Kyla. "Kantian ethics". In Christian Miller, ed., *Continuum Companion to Ethics*. New York: Continuum (2011): 168-189.

Engstrom, Stephen. *The Form of Practical Knowledge: a Study of the Categorical Imperative.* Cambridge, MA: Harvard University Press, 2009.

Frege, Gottlob. “Der Gedanke”. In Ignacio Angelelli, ed., *Gottlob Frege: Kleine Schriften*. Hildesheim: Georg Olms Verlagsbuchhandlung, 1967: 342-362.

Guyer, Paul. “The form and matter of the categorical imperative”. In *Kant’s System of Nature and Freedom*. Oxford: Oxford University Press, 2005: 146-168

Herman, Barbara. “Leaving deontology behind”. In *The Practice of Moral Judgment*. Cambridge, MA: Harvard University Press, 1993: 208-240.

Hill, Jr., Thomas E. “Personal values and setting oneself ends”. In *Human Welfare and Moral Worth: Kantian Perspectives*. Oxford: Oxford University Press, 2002: 244-274.

Hills, Alison. “Kantian value realism”. *Ratio* XXI(2) (2008), 182-200.

Juvenal, *The Satires*. Edited by John Ferguson. Bristol: Bristol Classical Press, 1979.

Kahn, Samuel. “The guise of the objectively good”. *Journal of Value Inquiry*, 47(1-2) (2013): 87-99.

Kant, Immanuel. *Anthropology from a Pragmatic Point of View*. Translated and edited by Mary Gregor. The Hague: Martinus Nijhoff, 1974.

Critique of Practical Reason. Translated and edited by Mary Gregor. Cambridge: Cambridge University Press, 1997.

Critique of Pure Reason. Translated and edited by Paul Guyer and Allen Wood. Cambridge: Cambridge University Press, 1998.

Groundwork of the Metaphysics of Morals: A German-English Edition. Translated and edited by Mary Gregor and Jens Timmermann. Cambridge: Cambridge University Press, 2011.

Lectures on Ethics. Translated by Peter Heath. Edited by Peter Heath and J.B. Schneewind. Cambridge: Cambridge University Press, 1997.

The Metaphysics of Morals. Translated and edited by Mary Gregor. Cambridge: Cambridge University Press, 2015.

Religion Within the Bounds of Bare Reason. Translated by Werner S. Pluhar. Indianapolis/Cambridge: Hackett, 2009.

McElrath, Joseph R., Crisler, Jesse S., and Shillinglaw, Susan, eds. *John Steinbeck: the Contemporary Reviews.* Cambridge: Cambridge University Press, 1996.

Mele, Alfred R. "Motivation: essentially motivation-constituting attitudes". *The Philosophical Review*, 104(3) (1995): 387-432.

"Internalist moral cognitivism and listlessness". *Ethics*, 106(4) (1996): 727-753.

Milton, John. "Paradise lost". In *The Complete English Poems,* edited by Gordon Campbell. London: David Campbell Publishers, 1992: 148-441.

Moore, G.E. *Principia Ethica*. Cambridge: Cambridge University Press, 1903/1971.

Moran, Kate. "Delusions of virtue: Kant on self-conceit". *Kantian Review*, 19(3), 2014: 419-447.

Pistorius, Hermann Andreas. "*Grundlegung zur Metaphysik der Sitten* von Immanuel Kant". (26-38).

"*Kritik der praktischen Vernunft* von Immanuel Kant". (78-98). Both in Bernward Gesang, *Kant's Vergessener Rezensent*. Hamburg: Felix Meiner Verlag, 2007:.

"*Groundwork of the Metaphysics of Morals* by Immanuel Kant". Translated by Michael Walschots. Draft, final version to appear in Michael Walschots, *Background Sources to Kant's Practical Philosophy*.

Raz, Joseph. "On the guise of the good". In Sergio Tenenbaum, ed. *Desire, Practical Reason and the Good*. Oxford: Oxford University Press, 2010: 111-137.

Reath, Andrews. "Kant's theory of moral sensibility: respect for the moral law and the influence of inclination". In *Agency and Autonomy in Kant's Moral Theory*. Oxford: Clarendon Press, 2006: 8-32.

"Did Kant hold that rational volition is *sub ratione boni*?" In Mark Timmons and Robert

E. Johnson, eds. *Reason, Value, and Respect: Kantian Themes from the Philosophy of Thomas E. Hill, Jr.* Oxford: Oxford University Press, 2015.

Rödl, Sebastian. "The form of the will". In Sergio Tenenbaum, ed., *Desire, Practical Reason and the Good*. Oxford: Oxford University Press, 2010: 138-160.

Schopenhauer, Arthur. *The World as Will and Representation, vol. 1*. Translated by E. F. J. Payne. New York: Dover, 1969.

Sensen, Oliver. *Kant on Human Dignity*. Berlin: Walter de Gruyter, 2011.

Sidgwick, Henry. "The Kantian conception of free will". *Mind*, 13(51), 1888: 405-412.

Silber, John R. "The Copernican revolution in ethics: the good reexamined". In Robert Paul Wolff (ed.), *Kant: A Collection of Critical Essays*. London: MacMillan, 1968: 266-290.

Kant's Ethics: The Good, Freedom, and the Will. Boston: De Gruyter, 2012.

Steinbeck, John. *East of Eden.* In *Novels 1942-1952*, edited by Robert DeMott. New York: Library of America, 2001.

Tenenbaum, Sergio. *Appearances of the Good: An Essay on the Nature of Practical Reason*. Cambridge: Cambridge University Press, 2007.

Timmermann, Jens. *Kant's* Groundwork of the Metaphysics of Morals*: A Commentary*. Cambridge: Cambridge University Press, 2007.

"What's wrong with 'Deontology'?" *Proceedings of the Aristotelian Society*, 115(1) (2015): 75-92.

"Quod dubitas, ne feceris: Kant on using conscience as a guide". *Studi Kantiani*, XXIX (2016): 163-168.

"The law and the good: Kant's paradox of method". Unpublished draft. Final version to appear in Violetta Waibel, ed., *Proceedings of the Vienna Kant Congress*, Berlin: Walter de Gruyter, forthcoming.

Velleman, J. David. "The guise of the good". *Noûs* 26(1) (1992): 3-26.

Willaschek, Marcus, Jürgen Stolzenberg, Georg Mohr and Stefano Bacin, eds. *Kant-Lexikon, Band I*. Berlin: Walter de Gruyter, 2015.

Wuerth, Julian. *Kant on Mind, Action, and Ethics*. Oxford: Oxford University Press, 2014.