Franz Knappik, Erasmus Mayr*

"An Erring Conscience is an Absurdity": The Later Kant on Certainty, Moral Judgment and the Infallibility of Conscience

https://doi.org/10.1515/agph-2019-1004

Abstract: This article explores Kant's view, found in several passages in his late writings on moral philosophy, that the verdicts of conscience are infallible. We argue that Kant's infallibility claim must be seen in the context of a major shift in Kant's views on conscience that took place around 1790 and that has not yet been sufficiently appreciated in the literature. This shift led Kant to treat conscience as an exclusively second-order capacity which does not directly evaluate actions, but one's first-order moral judgments and deliberation. On the basis of this novel interpretation, we develop a new defence of Kant's infallibility claim that draws on Kant's account of the characteristic features of specifically *moral* judgments.

1 Introduction

In his later writings on moral philosophy, Kant repeatedly expresses the view that the verdicts of conscience are infallible. Thinking of conscience as liable to error, Kant claims, is not just false, but "an absurdity".[1] This insistence on the infallibility of conscience is prima face highly surprising. First, it is hardly a generally held, or intuitively plausible, view that conscience is entirely immune to error. Even if it were granted that the verdicts of conscience are generally reliable, why could people not be badly misguided in (some of) their moral beliefs and to this extent have a 'warped' conscience? To rule out the possibility of such cases as an

1 MpVT 8:268; MS 6:401. – The abbrevations used are explained at the end of the article. In quoting Kant, we use (with minor modifications, and unless stated otherwise) the translations indicated in the abbreviations section, and give volume and page numbers of the Akademieausgabe.

***Corresponding authors: Franz Knappik,** Department of Philosophy, University of Bergen, Postboks 7805, 5020 Bergen, Norway; franz.knappik@uib.no; **Erasmus Mayr,** Department of Philosophy, Universität Erlangen-Nürnberg, Bismarckstrasse 1, 91054 Erlangen, Germany; erasmus.mayr@fau.de

"absurdity" is, at the very least, a remarkably strong claim.[2] Second, there is an additional reason why the infallibility claim is particularly problematic in Kant's own case: for in several of his lectures on moral philosophy, Kant himself explicitly affirms that conscience *can* be erring.

Kant's account of conscience has attracted considerable attention in the recent literature, and the exegetical and systematic puzzles raised by his infallibility claim have not gone unnoticed.[3] But, as we are going to argue, none of the proposed readings of Kant's account of conscience in general, and of his infallibility claim in particular, has been fully satisfactory. We will therefore offer a novel interpretation that centres around two claims. The first claim is that Kant's infallibility thesis must be seen in the context of a major shift in Kant's general views on conscience that took place around 1790, and that has not yet been fully appreciated in the literature. As we will argue, this shift led Kant to adopt an entirely novel and unconventional account of conscience. On his later view, Kant sees conscience not as a faculty that issues moral evaluations of (types of) actions – as did almost all traditional views of conscience, including Kant's own earlier account – but rather as a *second-order* capacity to evaluate one's own first-order moral *judgments*. Our second claim is that once Kant's later account of conscience is properly understood, a plausible defence of the infallibility claim can be mounted if, in addition, some key features of Kantian *moral* judgments are taken into account as well.

Our discussion will set off with a brief presentation of some crucial passages from Kant's works in the 1790s where the infallibility claim is made (Section 2). We will then go on to examine and criticize the readings of Kant's views on conscience and its infallibility that have been offered in the current literature (Section 3). Section 4 will compare Kant's later claims on conscience with discussions of conscience in his earlier lectures, where he advocates a modified view of Baumgarten's orthodox account of conscience as a first-order, fallible capacity. This comparison will motivate our contention that Kant's views on the nature of conscience underwent a radical change around 1790 and that the infallibility claim belongs to his later, exclusively second-order account of conscience. Dis-

2 Even authors like Bonaventure and Aquinas who assume that conscience is infallible in *some* of its functions (namely, as source of general moral knowledge) grant that in other functions (the application of general knowledge to particular actions), conscience can easily go wrong: cf. footnote 32 below.

3 The following recent studies are either entirely dedicated to Kant's account of conscience, or include substantial discussions of it: Hill 2002a, 2002b; Hoffmann 2002; Timmermann 2006; Moyar 2006; Wood 2008; Ware 2009; and Esser 2013. Earlier treatments include Paton 1979 and Lehmann 1980.

tinguishing between these two phases of Kant's views will allow us to develop a new account of how Kant's infallibility claim can be made sense of, which draws heavily on Kant's account of specifically *moral* judgments (Sections 5–7). We will conclude by pointing out that, while Kant's infallibility claim can be defended in one important version, it still remains too wide in scope and should be restricted (Section 8).[4]

2 Some Crucial Passages on the Infallibility of Conscience

Let us begin by looking at three passages in Kant's later work where he explicitly advances the claim that the verdicts of conscience are infallible. The first passage is from the "Concluding Remark" of Kant's 1791 essay *On the Miscarriage of All Philosophical Trials in Theodicy*. Halfway through a series of considerations about truthfulness and conscientiousness, Kant writes:

> (A) Moralists speak of an 'erring conscience'. But an erring conscience is an absurdity; and, if there were such a thing, then we could never be certain we have acted rightly, since even the judge in the last instance can still be in error. I can indeed err in the judgment *in which I believe* to be right, for this belongs to the understanding which alone judges objectively (rightly or wrongly); but in the judgment *whether I in fact believe* to be right (or merely pretend it) I absolutely cannot be mistaken, for this judgment – or rather this proposition – merely says that I judge the object in such-and-such a way. (MpVT 8:268)

The second passage is reported in Johann Friedrich Vigilantius' transcript of Kant's 1793/4 lectures on *Metaphysics of Morals*. In the course of a discussion of conscience, Kant explicitly attacks one of the 'moralists' who admit the possibility of an erring conscience, namely Alexander Baumgarten:

> (B) [T]he judgment founded on examination of the *factum* does not, by itself, constitute conscience, and [...] indeed this judgment may be an error, whereas conscience can never be that, whence the division *inter conscientiam erroneam et rectam* is totally false and unthinkable. Baumgarten locates conscience merely in the *subsumptio factorum nostrorum sub lege*. This amounts, therefore, to equating it with the soul's faculty of judgment [*Urtheilskraft*], whereby the *facta judicantis* would be subjected to the rules of the understanding. (V-MS/Vigil [1793/4] 27:615 f.)

4 Parts of Sections 4, 6 and 8 draw on arguments in Knappik/Mayr 2013.

The third passage is found in the Introduction to the Doctrine of Virtue, the second part of Kant's 1797 *Metaphysics of Morals*. In section XII of this Introduction, Kant discusses conscience alongside with 'moral feeling', 'love of human beings', and 'respect', and claims the following:

> (C) [A]n *erring* conscience is an absurdity. For while I can indeed be mistaken at times in my objective judgment as to whether something is a duty or not, I cannot be mistaken in my subjective judgment as to whether I have submitted it to my practical reason (here in its role as judge) for such a judgment; for if I could be mistaken in that, I would have made no practical judgment at all, and in that case there would be neither truth nor error. (MS 6:401)

Despite differences in emphasis and formulation, all three passages clearly make two crucial claims:

First, the possibility of an erring conscience is rejected in very strong terms – as an "absurdity" (*Theodicy*, *Metaphysics of Morals*) and as "totally false and unthinkable" (*Metaphysics of Morals Vigilantius*). So, not content with claiming merely special reliability for the verdicts of conscience, Kant commits himself to the much stronger view that an error of conscience is *strictly impossible*.

Second, in all passages Kant takes pains to distinguish the verdict of conscience from other kinds of judgments that *do* allow for error – namely, the judgments by which we directly evaluate actions.[5] This other, "objective", or first-order judgment about the rightness of an action is assigned to another faculty (variously identified as Understanding [*Theodicy*], Judgment [*Metaphysics of Morals Vigilantius*], or Practical Reason "in its role as judge" [*Metaphysics of Morals*]).

Moreover, both the passage from the *Theodicy* and the one from the *Metaphysics of Morals* provide a positive specification of the function of conscience that sets it apart from the second, fallible judgment. The verdict of conscience, according to these passages, is a "subjective" judgment (*Metaphysics of Morals*) insofar as it is a *reflective* or *second-order judgment*.[6] The idea of conscience as

5 *Theodicy*: "judgment *in which I believe* to be right", judgment by which I judge "objectively (rightly or wrongly)"; *Metaphysics of Morals Vigilantius*: "judgment founded on examination of the *factum*" (where "*factum*" stands for the deed, as is clear from the context in Baumgarten's *Ethica* (§ 177 with § 175, 27:780 f.) that Kant is referring to here); *Metaphysics of Morals*: "objective judgment as to whether something is a duty or not".

6 In Kant's words: a judgment regarding the question "*whether I in fact believe* to be right (or merely pretend it)" (*Theodicy*), whether "I judge the object in such-and-such a way" (*Theodicy*), or "whether I have submitted" a case "to my practical reason (here in its role as judge)" for its "objective" judgment (*Metaphysics of Morals*). – In the section "Aesthetic Preconditions of Receptivity to Duty" in the *Metaphysics of Morals*, Kant specifies the claim that conscience is not directed towards the "object" but "merely to the subject" (MS 6:400) by adding: "to affect moral feeling by its [sc. practical reason's] act [*das moralische Gefühl durch ihren Act zu afficiren*]" (MS

passing a second-order rather than a first-order judgment is spelt out in more detail in a further important passage from Kant's 1794 *Religion within the Boundaries of Pure Reason*:

> (D) Conscience could also be defined as *the moral faculty of judgment, passing judgment upon itself* [*die sich selbst richtende moralische Urtheilskraft*], except that this definition would be much in need of prior clarification of the concepts contained in it. Conscience does not pass judgment upon actions as cases that stand under the law, for this is what reason does so far as it is subjectively practical (whence the *casus conscientiae* and casuistry, as a kind of dialectic of conscience). Rather, here reason judges itself, whether it has actually undertaken, with all diligence, that examination of actions (whether they are right or wrong), and it calls upon the human being himself to witness *for* or *against* himself whether this has taken place or not. (RGV 6:186)

While Kant does not claim infallibility for conscience in this passage, it is parallel to passages (A) to (C) in that it distinguishes the verdict of conscience from the moral evaluation of actions (which is described here as a task of "reason [...] so far as it is subjectively practical") and explicitly assigns a reflective status to the verdicts of conscience. Moreover, Kant further specifies the object of that verdict: conscience assesses whether we have *diligently examined* the case in question. In view of the obvious parallels to passages (A) to (C), one can hardly avoid the conclusion that Kant also claims that we are infallible with regard to the question of whether we have diligently examined a given action.[7]

The passages we have cited not only make pressing the question of how exactly the infallibility of conscience can be made sense of within Kant's moral psychology. They also strongly suggest that the answer to this question will crucially depend on how precisely Kant conceives of conscience in these passages,

---

6:400). This might be seen as speaking against the idea that the second-order (or subject-directed) character of Kantian conscience is a matter of issuing a (second-order) judgment at all. Rather, the phrase seems to point to a *motivational* function of conscience. (Thus, Guyer 2012, 143, interprets this passage in the sense that conscience "causes or stimulates [...] some moral feeling that is, presumably, a trigger to an action".) While our focus in this article is on the cognitive function of conscience, the account of the second-order cognitive role of conscience that we will give is compatible with the idea that conscience also plays a related motivational role: cf. footnote 52. (Regarding the cognitive dimension of Kantian conscience, Guyer seems to hold a purely first-order reading – for instance, when he proposes to see conscience as "the empirical voice that informs us of our *specific obligations*" in a given situation (145), that is, of the appropriate maxims for that situation (144). He is silent both about the evidence for a second-order function, and about Kant's infallibility claim.)

7 Of course, it is far from clear how Kant can treat such different questions as (a) whether we have judged an action at all, and (b) whether we have diligently examined it, as both providing characterizations of the subject-matter of conscience. We will return to this problem in Section 6.

and this is itself far from clear. For one thing, the claim that conscience issues "subjective" or second-order judgments is quite surprising, to say the least. After all, it is entirely natural to assume that our conscience directly assesses our actions, or courses of action. Even worse, Kant's characterization of conscience as a second-order capacity conflicts with other passages in which he describes conscience as being concerned with our actions, and hence, with "objective", or first-order, questions:

(1) In his earlier lectures on moral philosophy, Kant clearly holds that conscience provides first-order evaluations of actions.[8]
(2) In an earlier section of *Metaphysics of Morals Vigilantius*, Kant claims that "conscience consists in the ability to impute one's own *factum* [i. e., one's deed] to oneself, through the law itself" (27:575). This clearly presupposes that the verdict of conscience is a verdict about one's actions, not about one's judgments.
(3) In the same context of *Metaphysics of Morals* in which the above-cited passage on the infallibility of conscience occurs, Kant also describes conscience as "*die dem Menschen in jedem Fall eines Gesetzes seine Pflicht zum Lossprechen oder Verurtheilen vorhaltende praktische Vernunft*" (MS 6:400), which Gregor translates as "practical reason holding the human being's duty before him for his acquittal or condemnation in every case that comes under a law". At least when understood in this way, this passage, too, suggests that conscience deals with our actions, not with our judgments about actions.
(4) Finally, in another section of *Metaphysics of Morals* – "On the Human Being's Duty to Himself as His Own Innate Judge" – Kant discusses conscience in terms of an "internal court" and assumes that what is judged by this court are our actions themselves. This becomes clear, for instance, when he claims that "a human being's conscience will, accordingly, have to think of *someone other* than himself [...] as the judge *of his actions*" (MS 6:438, second emphasis added).

Any plausible account of Kant's views on conscience and its infallibility will have to address the question of how these passages can be squared with Kant's remarks on the "subjective", second-order character of the verdicts of conscience in *Theodicy* and *Metaphysics of Morals*. In the next section, we will examine how other commentators have dealt with these passages, and how they have related Kant's view on the content of the verdicts of conscience to the explanation of his infallibility thesis.

8 We will come back to the details of Kant's discussion in these lectures in Section 4.

3 Alternative Readings of Kant's Views on Conscience

As we have seen in the last section, two diverging claims about the role of conscience can be found in Kant's texts: (a) the claim that conscience evaluates our actions (thus issuing first-order verdicts) and (b) the claim that conscience deals with the judgments that Practical Reason passes on our actions (and therefore issues second-order verdicts). The approach to Kant's account of conscience that is shared by most commentators,[9] and that we will therefore call the *standard approach*, responds to this divergence by interpreting Kant as *combining* both claims. On this approach, conscience assesses *both* our actions (thus operating at the first-order level), and our first-level moral judgments (thus operating at the second-order level, too). Regarding the first-order function of conscience, proponents of the standard approach agree that *the verdict of conscience compares a concrete action with a first-order moral judgment* issued by Practical Reason.[10] Regarding Kant's remarks about the *second-order* role of conscience, some authors have considered the function of comparing actions to judgments to be already sufficient to cover this role.[11] Others see the second-order function of conscience as a distinct and additional function of conscience.[12] In either case,

9 Moyar 2006 is a notable exception. We discuss his interpretation in footnote 53 below.

10 The idea of a *comparison* between the action and the first-order moral judgment is supposed to take into account Kant's distinction between the (infallible) verdict of conscience and the (fallible) moral judgment of Practical Reason: on this reading, we may be mistaken in our assessment of what is morally permissible in a given situation (the judgment of Practical Reason), but we can infallibly tell whether our concrete action lives up to that assessment or not (the judgment of conscience). – Thus, Thomas Hill Jr. locates the essential function of Kantian conscience in that of "'an inner judge' that condemns (or acquits) one for inadequate (or adequate) effort to live according to one's best possible, though fallible, judgments about what (objectively) one ought to do" (Hill 2002a, 280). And Allen Wood holds that "[c]onscience is the process of moral reflection that makes use of such [sc. first-order] moral judgments in delivering on myself a verdict of guilt or acquittal for some action I have done or am contemplating" (Wood 2008, 190). Cf. also Timmermann 2006, esp. 295, 297, 303; Esser 2013, esp. 280; Ware 2009, 619 f.; Hoffmann 2002, 439.

11 For instance, Wood claims that the "process of moral reflection" in which we compare actions to moral judgments leads to verdicts to the effect "that I have applied the standards of moral judgment to myself (whether or not I have rendered the right substantive judgment in doing so)" (Wood 2008, 190). – Cf. also Paton 1979, 241–243, and Esser 2013, 280.

12 Thus, Hill has suggested seeing this function as a *special case* of the function of an 'inner judge' – the case in which we examine whether we have lived up to a specific duty, namely, the "second-order duty of due care in scrutinizing and appraising our acts diligently (by 'holding them up' to our judgment of the first-order duties)" (Hill 2002a, 303.) – Owen Ware, by contrast, thinks that Kant assigns two *distinct* tasks to conscience: first, to compare our actions to our

the second-order role is seen only as a *partial* aspect of conscience, which Kant adds to an essentially traditional, first-order view of conscience as concerned with assessing our actions.

It should be noted at this point already that the standard approach stands in considerable tension with Kant's texts from the outset. For passage (D) explicitly *restricts* conscience to an exclusive second-order function. And passages (A) and (C) both argue that conscience as such cannot err *because*, unlike first-order evaluations of actions, a certain kind of second-order judgment cannot be mistaken. It is at least very hard to make sense of these passages if conscience is seen as *also* having a first-order function. Therefore, it is quite unclear from the outset how the standard approach can possibly do justice to Kant's own formulations.

The difficulties for the standard approach increase further when we turn to the question of how its proponents have dealt with the issue of infallibility. There are in principle three possibilities here, all of which have been explored in the literature. (1) The first possibility is to hold that for Kant, we are actually infallible regarding the question whether in a given action, we live up to our (fallible) first-order moral judgment about what is morally required in the relevant situation.[13] However, it is hard to see how such a claim on Kant's part could be justified. Knowledge about whether one's action complies with one's relevant moral judgment presupposes both knowledge about what one's relevant moral judgment is, and knowledge about what one's action is or was. If the agent can err about one or both of these factors, his judgment about their conformity cannot be infallible, either. But it seems quite clear that we can be mistaken about both our first-order moral judgments and about our actions. Errors of memory can occur with regard to past judgments and actions, and our judgment about what we have done or are doing is open to factual errors. More importantly, there is a form of self-deception about our past and present actions that is particularly important for Kant: namely, self-deception about our motivation. Such self-deception causes errors about one's motivation that can impugn one's judgment about whether one's action conforms to one's moral judgment, too. To see this, imagine the following case: a friend of mine asks my advice about a paper he has written, which I believe is no good. If I tell him my negative assessment because I want to help him (e. g., to avoid embarassment by publishing the paper) and

moral judgments (a first-order function), and second, to issue a "higher-order judgment of the care the agent applies (or fails to apply) in the act of examining what action she ought or ought not take" (Ware 2009, 693). Hill makes a similar claim in 2002b, 348.

13 Thus, Timmermann writes: "[A]gents [sc. according to Kant] *cannot* be mistaken in their 'subjective judgment' as to whether they have acted conscientiously, that is, in accordance with the decree that they took to be the command of practical reason" (Timmermann 2006, 303).

judge that it is morally good to do so, my action is morally unexceptionable. By contrast, if I tell him my assessment in order to hurt him, because I want to enjoy his discomfiture, my action is morally wrong. Hence, a moral judgment comes, in many cases, with a specification of the reasons for which I ought to (or am permitted to) act. As a consequence, the verdict of conscience, if understood in first-order terms, often has to check whether the action *with its actual motivation* complies with our judgment about what we ought to (or may) do *for what reason*. But Kant is famously pessimistic about our ability to detect the actual motivations of our actions.[14] Hence, there will often be cases in which we think that we have acted in accordance with a moral judgment, but have actually failed to do so, due to acting for the wrong reason. We therefore should conclude that the question whether we have lived up to our own moral judgments *cannot* be what conscience is infallible about for Kant.

By contrast, other advocates of the standard approach have proceeded from the assumption that Kant has in mind a more *restricted* form of infallibility. There are in principle two options for qualifying the infallibility claim. (2) One option here is to make the infallibility in question relative to the subject's epistemic access to her action and to her moral judgment. Thus, Hill suggests that conscience compares not our action, but our "conception of our act"[15], with our moral judgment. That only our "conception of our act" is relevant for the verdict of conscience, makes errors about what our action is irrelevant for the truth of this verdict. And regarding the moral judgment, Hill argues that in cases where we are mistaken about our moral judgment due to self-deception, failure of memory etc., the error is not due to conscience, but due to self-deception, memory, etc.[16] The main problem with this approach is that it fails to do justice to the passages from Kant's texts which we have quoted at the beginning. While the first part of this strategy – the restriction to the agent's conception of the action – may be seen as compatible with what Kant says, the second part clearly is not. In passages (A) to (C), Kant does not merely claim that conscience cannot *cause* an error: he also claims that *the judgments of conscience cannot possibly be mistaken, without qualifying this as to the source of error.* The infallibility claim, as put forward by him, thus entails that it is not possible that the verdict of conscience be false, *even* if other faculties, such as memory or perception, rather than conscience itself would be responsible for the falsity of the verdict.[17]

14 E. g., GMS 4:407; 4:419; RGV 6:38; 6:63.

15 Hill 2002a, 303 n.

16 Hill 2002a, 303 n.

17 We assume that in order to be infallible about the comparison between one's judgment and one's conception of one's action, one has to be infallible about one's judgment in the first place.

(3) Alternatively, proponents of the standard approach can restrict Kant's infallibility claim to the second-order operation that they ascribe to conscience. Thus, while admitting that we can be mistaken in the first-order exercise of conscience, Ware argues that "at least pertaining to its higher-order function, Kant is right: an erring conscience *is* an absurdity, for the simple reason that an agent can't critically assess her duties unconsciously".[18]

This strategy not only makes errors about our action irrelevant to the verdict of conscience that is supposed to be infallible. It also does much greater justice than the other two interpretations to the fact that Kant himself explicitly links his infallibility claim to the second-order characterization of conscience. Nevertheless, even this proposal faces decisive difficulties. First, it gives raise to *new* sources of potential error: we may very well be mistaken about whether we are critically assessing our duty, or have done so.[19] Second, infallibility goes beyond the connection which Ware defends. In order for our judgments about our careful assessment of duties to be infallible, it must (also) be the case that *if* we believe that we are carefully assessing our duties, this belief is true. The connection that Ware argues for does not yet entail this latter conditional. It can at best establish a dependence in the opposite direction, since that connection only entails that *if* I make a careful assessment I will also be conscious of it. Finally, despite its advantages over the first two options, this reading, too, faces a problem of textual adequacy. For in the passages (A) to (C), in which Kant claims infallibility for conscience, he does not restrict this infallibility to some *partial* function of conscience; he claims that conscience *as such* is infallible.

Thus, all extant variants of the standard approach are plagued by considerable difficulties. In the following sections, we will therefore develop an alternative reading which both better fits the passages we have discussed and will eventually allow us to mount a defence of Kant's infallibility claim.

18 Ware 2009, 693. – Similarly, Wood points out that "what Kant might mean in denying an erring conscience is [...] that if we do in fact genuinely submit ourselves to the judgment of conscience, then we cannot fail to be aware of doing so [...]" (Wood 2008, 191). Cf. also Hill 2002a, 303 n.

19 For instance, it might be that my actual deliberation is superficial and leaves out important considerations, but I successfully talk myself into believing that I *am* carefully examining my duties (cf. Wood 2008, 191). In the retrospective case, bad memory about my past deliberation is a further source of error.

4 A Comparison of Kant's Later and Earlier Views on Conscience

In this section, we will argue that Kant's views on conscience have significantly *developed* over time in a way that has not yet been sufficiently appreciated in the literature. Once this development is properly understood, it becomes possible to abandon the standard approach and to interpret Kant's account in the *late* writings – the only texts in which Kant claims infallibility for conscience – in *exclusively* second-order terms.

The first statement of the infallibility claim we have looked at in Section 2 came from Kant's 1791 essay *On the Miscarriage of All Philosophical Trials in Theodicy*. Before this essay, Kant was remarkably silent about conscience in his published writings on moral philosophy.[20] Nevertheless, transcripts from Kant's lectures on moral philosophy throughout the 1770s and 1780s show that these lectures included substantial discussions of conscience. In these discussions, Kant closely followed the treatment of conscience in the textbooks that he used in his lectures, Baumgarten's *Initia philosophiae practicae primae* and *Ethica philosophica*.[21] Baumgarten defines conscience as "act, or faculty, or habit of imputing deeds to oneself, and applying laws to them".[22] This account of conscience is squarely rooted in traditional medieval and early modern views of conscience. Such views (as found in authors like Bonaventure,[23] Aquinas,[24] Duns Scotus,[25] Butler,[26] Pufendorf,[27] and Wolff[28]) all agree that conscience is concerned with our

20 Conscience only receives a brief treatment in the second Critique (KpV 5:98 f.). Possible reasons why Kant grants conscience no important role in the *Groundwork for the Metaphysics of Morals* and the *Critique of Practical Reason* are discussed by Hoffmann 2002, 425 f., 435, and Timmermann 2006, 296.

21 Both works were first published in 1740. Kant uses the third edition (1760) of the *Initia* and both the second (1751) and the third edition (1763) of the *Ethica*: cf. Schneewind 1997, xxi. – The fullest discussion of Kant's engagement with Baumgarten's moral philosophy is still Schmucker 1961, 278–363.

22 "[*A*]*ctus, vel facultas, vel habitus facta sibi imputandi, et his leges applicandi*" (*Initia* § 200 (19:89); our translation). Very similar accounts of conscience are given by Samuel Pufendorf (*De jure naturae* 38 (I 3, § 4)) and Christian Wolff (*Vernünfftige Gedancken* 76 (§ 73)).

23 *Commentary on the Sentences*, Book II, distinction 39, in Potts 1980, 111.

24 Summa theologiae Ia, q79a13; De veritate q17a1.

25 *Ordinatio* II, dist. 39, in *On the Will* 164 f.

26 E. g., *Sermons* 49 (Sermon III).

27 See footnote 22.

28 See footnote 22.

(past, present, or future) *actions*. In the terms that we have been using so far, conscience is interpreted as issuing a *first-order* verdict.

In addition, Baumgarten describes conscience, thus understood, as an "inner tribunal" (*forum internum*).[29] Importantly, the verdict issued by this tribunal, i. e., the judgment of conscience, is fallible. For as Baumgarten argues, it rests on a syllogism, and syllogisms can be accurate or not: they can be formally valid or invalid, and they can involve true or false premises (more specifically, a major premise stating the law, and a minor premise describing the action; both premises can be false).[30] Hence, the judgment of conscience can be true or false, and it can be warranted or unwarranted. Again, this feature of Baumgarten's view is in thorough agreement with earlier accounts. Medieval authors often distinguish conscience as applied to concrete actions from a related capacity (sometimes considered a part of '*conscientia*', sometimes referred to as '*synderesis*') that allows us to cognize general moral truths. Even though medieval thinkers usually consider this latter capacity to be infallible,[31] they agree that error is possible in the application of general moral knowledge to the particular case.[32] The reasons for this are essentially the same as those indicated by Baumgarten.

When dealing with conscience in his lectures on moral philosophy throughout the 1770s and 1780s,[33] Kant adopts the main features of Baumgarten's traditional account. First, he presents conscience as a capacity of evaluating particular actions when he defines it as an "instinct for us to judge and pass sentence on our actions".[34] Second, he characterizes conscience as constituting a "forum

29 *Initia* § 182 (19:83). The distinction between *forum internum* (the tribunal of conscience) and *forum externum* (the legal tribunal) goes back to medieval canonical law (cf. Goering 2004).

30 *Ethica* § 177 (27:781).

31 See, for instance, Aquinas, *De veritate*, q16a2, q17a2, and Duns Scotus, *Ordinatio* II, dist. 39, in *On the Will* 164 ("*synderesis*" as infallible knowledge of general moral truths); Bonaventure, *Commentary on the Sentences*, Book II, distinction 39, in Potts 1980, 113 f., 120 ("*conscientia*" as infallible with regard to general moral truths).

32 E. g., Bonaventure, *Commentary on the Sentences*, Book II, distinction 39, in Potts 1980, 114, 120; Aquinas, *De veritate*, q17a2.

33 Of course, it is not always possible to ascribe a view to Kant on the basis of what he says in his lectures. But as we will see in the following, Kant does not limit himself in his lectures of the 1770s and 1780s to expounding Baumgarten's textbook view on conscience. Rather, he modifies this view by adding further points to it (which seem to be inspired by Crusius: cf. footnote 41). It is therefore fair to assume that Kant actually endorsed the resulting view.

34 V-Mo/Collins [1774/5] 27:296 f.; Cf. Me [1774/5] 161. (On the role of Kant's talk of an "instinct" and a "sentence" in this context, see below.)

internum".[35] And third, Kant takes over Baumgarten's view that conscience is fallible when he discusses different ways in which conscience can err.[36]

At the same time, Kant points out that further characteristics need to be added in order to distinguish conscience from our ordinary capacities for moral reasoning and to guarantee a distinctive function and phenomenology for it.[37] Kant argues, on the one hand, that there is an important distinction to be made between *evaluating* (*beurteilen*) and *sentencing* (*richten*) actions and claims that it is the task of conscience to sentence, and not merely to evaluate, our conduct.[38] Unlike the mere evaluation of an action as good or bad (which could be merely prudential),[39] the verdict of conscience makes a crucial difference for our moral self-esteem, issuing either in qualms of conscience or in relief.[40]

On the other hand, Kant thinks that while we can choose to perform or not to perform evaluations at will, a *sentence* cannot simply be issued or prevented at will. For the culprit, in particular, it is an unavoidable occurrence that is not under his control, and, once issued, is definitive and legally binding. Kant expresses this point by saying that conscience is not strictly speaking a *faculty* (for this would require wilful control), but rather an *instinct* that we cannot (at least not directly) control.[41]

---

35 Me [1774/5] 82 f.; V-Mo/Collins [1774/5] 27:296 f.

36 Thus, he explains in his 1774/5 lectures on moral philosophy: "The difference between the correct and the errant conscience lies in this, that error of conscience takes two forms, *error facti* and *error legis*" (V-Mo/Collins [1774/5] 27:354; cf. Me [1774/5] 165 f.). And similarly, in 1782/3: "All errors of conscience are either moral or logical. They are either located in morality or in understanding [...]" (V-PP/Powalski [1782/3] 27:197 f.).

37 Cf., e. g., V-Mo/Collins [1774/5] 27:296 f.

38 Me 161; V-PP/Powalski [1782/3] 27:197; cf. Hoffmann 2002, 435.

39 Me [1774/5] 162.

40 Me [1774/5] 161.

41 Me[1774/5] 83, 163; V-Mo/Collins [1774/5] 27:297; V-PP/Powalski [1782/3] 27:162, 197; R7181, 19:266; cf. Hoffmann 2002, 434. – Without explicitly referring to him in this context, Kant follows in both points Christian August Crusius' discussion of conscience in the latter's *Anweisung vernünftig zu leben* (first published in 1744; Kant owned the second edition of 1751, cf. Anonymous 1808, 10). Crusius explicitly postulates an innate "instinct of conscience" (*Gewissenstrieb*) (*Anweisung* 177 (§ 132)), and takes pains to distinguish this instinct from simple moral evaluations and prudential regret (*Anweisung* 177 f. (§ 132)). Nevertheless, Crusius agrees with more traditional views that conscience assesses the moral quality of actions, and that it is fallible in its application to concrete actions: *Anweisung* 189 (§ 138). (On Crusius' role in the development of Kant's moral philosophy, see Schmucker 1961, 81–87; for an account of Crusius' moral philosophy in general, see Schneewind 1998, 445–456.) – In addition, Lehmann 1980, 33, mentions Rousseau's *Émile* as likely source for Kant's characterization of conscience as "instinct".

So on the picture that emerges from Kant's lectures in the 1770s and 1780s, Kant's account of conscience in this period adopts the basic tenets of the traditional view – conscience as *first-order* capacity, and as *fallible* with regard to the evaluation of concrete actions – while adding qualifications in order to capture the distinctive character of conscience. Now, as we have already seen, Kant is at pains in his later statements on conscience, too, to set conscience apart from the faculty for moral judgment (see passages (A) to (C) quoted in Section I). But comparison with the earlier texts shows that in doing so he now pursues an entirely different, radically innovative strategy. For Kant now (in passage (B)) explicitly *rejects* Baumgarten's account of conscience: "[T]he division *inter conscientiam erroneam et rectam* is totally false and unthinkable. Baumgarten locates conscience merely in the *subsumptio factorum nostrorum sub lege*. This amounts, therefore, to equating it with the soul's faculty of judgment, whereby the *facta judicantis* would be subjected to the rules of the understanding" (V-MS/Vigil [1793/4] 27:615).[42] We therefore have to conclude that by 1791, Kant had come to believe that the modified Baumgartian view of his lectures does not suffice to do justice to the distinctive nature and character of conscience, but threatens to ultimately conflate it with the general capacity of moral judgment.

As a consequence, Kant's remarks on the second-order character of conscience in passages (A), (C), and (D) must not be read as merely introducing some new sub-function of conscience, which could co-exist with the traditional first-order function of conscience (as the standard approach assumes). Rather, Kant introduces a much more radical change in these passages, assigning to the verdict of conscience an entirely new *object*. On Kant's new account, conscience does not, properly speaking, judge our *actions* at all (this judgment is the task of Understanding, or Practical Reason). Rather, conscience only issues an infallible *second-order judgment* whose object is the first-order moral judgment of understanding.

To this change in object corresponds a crucial second change in the *standard* of judgment. On the earlier account – as on all traditional accounts – conscience judges actions with regard to their moral permissibility or goodness. By contrast, according to passage (A), conscience checks *whether there has been a first-order judgment at all*; according to passage (D), conscience checks *whether we have diligently examined* a case.

It is true that Kant does not always stick to this revisionary account of conscience and continues to use first-order characterizations of conscience in some

42 His related remarks about "moralists" in passage (A) – "Moralists speak of an 'erring conscience.' But an erring conscience is an absurdity [...]" (MpVT 8:268) – can be read, too, as rejecting the traditional view of conscience held by Baumgarten and, for that matter, by Crusius.

passages even after 1791. However, we believe that if the passages in question are examined in their context, it can be shown that they do not constitute any decisive counter-evidence against the proposed reading. We have mentioned the relevant passages already in Section 2: the first of them occurs in *Metaphysics of Morals Vigilantius* (1793/4), where Kant claims that "conscience consists in the ability to impute one's own *factum* [i. e., one's deed] to oneself, through the law itself" (V-MS/Vigil [1793/4] 27:575). While this suggests that conscience itself is still concerned with judging actions, in a passage later in the same series of lectures Kant makes the opposed claim that "conscience *also takes into account* a valid imputation of our actions. All this, however, belongs to practical reason" (V-MS/Vigil [1793/4] 27:616, emphasis added).[43] Thus, the self-imputation of an action is now described as the task of Practical Reason, not, strictly speaking, of conscience. Moreover, there are further passages in *Metaphysics of Morals Vigilantius* where Kant explicitly advances a second-order reading of conscience.[44] Given that Kant often expounds textbook views in his lectures even where they contradict his own position,[45] it should not be surprising that tensions can occur within these lectures – even more so within transcripts written by students.

More serious problems for our reading arise from the two passages in the *Metaphysics of Morals* that seem to use first-order characterizations of conscience. One of them describes conscience in terms of an 'internal court', which assesses our *actions*, not our judgments. Thus, Kant points out that through conscience, the agent imagines another person as "judge *of his actions*" (MS 6:438, emphasis added). We believe that this passage constitutes a relic of Kant's own earlier position.[46] As Kant in this context does not give a full-blown specification of the structure and function of conscience, and is silent about the issue of fallibility vs. infallibility, we may conclude that this passage does not voice Kant's considered account of conscience at this time.[47]

43 By "all this", Kant refers to imputation together with further first-order functions that have been traditionally ascribed to conscience (namely, those of a moral legislature, judiciary, and executive: V-MS Vigil [1793/4] 27:616).

44 Cf. V-MS Vigil [1793/4] 27:614 f. (quoted at length in Section 7 below), and V-MS Vigil [1793/4] 27: 616.

45 Cf. Lehmann 1980, 28.

46 Since judicial metaphors are crucial to Kant's earlier account (remember his idea of conscience as passing a 'sentence' on an action), it would have been natural for him to take up his earlier characterizations of conscience in a context that deals specifically with conscience as a 'forum internum'.

47 This view is supported by an independent observation. The discussion in question occurs in a part of the *Metaphysics of Morals* (§§ 13–15) which closely follows Baumgarten's *Ethica*. In *Ethica* §§ 150–190 (27:909–919), Baumgarten discusses self-cognition, self-judging, and conscien-

Finally, at MS 6:400, Kant describes conscience as "*die dem Menschen in jedem Fall eines Gesetzes seine Pflicht zum Lossprechen oder Verurtheilen vorhaltende praktische Vernunft*". If the phrase is parsed as in Gregor's translation – "practical reason holding the human being's duty before him for his acquittal or condemnation in every case that comes under a law" – it does suggest a first-order account of conscience.[48] This would be a substantial obstacle to our interpretation, as the passage occurs in the very same context in which Kant also characterizes conscience in second-order terms, and ascribes infallibility to it (passage (C) in Section 2). However, Gregor's way of reading the phrase is far from mandatory. It is equally natural to read "*zum Lossprechen oder Verurtheilen*" as specifying the content of the duty in question (rather than as a complement of "*vorhaltende*"). On this alternative reading, the right translation of the phrase would characterize conscience as "practical reason holding the human being's *duty to acquit or condemn* before him in every case that comes under a law" (emphasis added). Thus understood, the passage fits well with a second-order reading: for as we will see in more detail in the next section, Kant treats the question whether we have applied due diligence in the evaluation of our (actual or possible) action as equivalent to the question whether we have passed a moral judgment *at all*.[49]

We therefore take none of the passages we have just discussed to really undermine our contention that Kant, from the early 1790s onward, attributed to conscience an exclusively second-order function. But, independently from the textual evidence, one might have quite another worry about our proposed reading: if Kant, in his later writings, had indeed seen conscience as an exclusively second-order capacity, this would have been completely at variance with the normal usage of 'conscience', and would have made it quite mysterious why acting in accord with one's conscience should have any particular normative importance. After all, why should Kant have begun to use the term 'conscience' to

---

tiousness as duties against oneself; the same topics are discussed, under the same title of 'duties against oneself', by Kant in §§ 13–15 of *Metaphysics of Morals*. The conformity with Baumgarten in this context goes so far that Kant in § 13 treats conscience under the (Baumgarten-inspired) heading of "the Human Being's Duty to Himself as His Own Innate Judge" (MS 6:437), even though Kant had explicitly denied earlier in the *Metaphysics of Morals* (MS 6:400 f.) that there is a duty to provide ourselves with, or to have, a conscience (MS 6:400) (cf. also Esser 2013, 273 and 275). Hence, the notion of conscience as a duty provides a further instance of how Kant, in §§ 13–15 of the *Metaphysics of Morals*, follows Baumgarten's model to the extent that he directly contradicts what he says about conscience elsewhere in the *Metaphysics of Morals*.

48 Timmermann 2006, 296.

49 We think that Kant's formulation "in every case that comes under a law" does refer to (actual or possible) actions that are subject to moral evaluation, but read this as specifying the occasion at which conscience becomes active, rather than the subject-matter of its verdict.

refer to an entirely different capacity than the one that is traditionally designated by it? And while he clearly considered it to be a fundamental moral flaw for an action to go against the verdict of conscience, how could the moral value of an action crucially depend on its accord with the latter, unless conscience assessed the action's permissibility or obligatoriness?

A full response to these questions lies beyond the scope of this article. It must suffice here to briefly make two points. First, the German adjective '*gewissenhaft*', like the English 'conscientious', is both used to describe an agent who acts in accordance with the verdict of his or her conscience, and to characterize someone who acts with diligence and care. Far from being detached from ordinary usage, Kant's late theory of conscience can be seen as being inspired by this feature in the semantics of '*gewissenhaft*' (which is documented already for the 17th century: cf. Grimm & Grimm 1854, s. v.), and as making sense of it: from the viewpoint of Kant's theory, this feature appears not as an arbitrary ambiguity, but rather as pointing to a direct link between diligence and conscience.

Second, Kant does assign a considerable normative impact to the second-order examination of one's moral judgment in his late writings. In particular, he emphasizes that we must not undertake morally relevant actions unless we are, upon careful examination, entirely sure that these actions are morally allowed. As he points out, conscience is "*a consciousness which is of itself a duty*" in the sense that "we *ought to venture nothing where there is danger that it might be wrong (quod dubitas, ne feceris!* Pliny)" (RGV 6:185 f.). Hence, if we follow a first-order judgment that does *not* bear the scrutiny of conscience, and act in this sense "unconscientiously" (e. g., RGV 6:187), our action is *ipso facto* morally wrong. Kant further stresses that such unconscientious actions result from "dishonesty" (*Unredlichkeit*, RGV 6:188) or "untruthfulness" (*Unwahrhaftigkeit*, RGV 6:187) – a vice that is "in itself damnable" (RGV 6:187).[50] Hence, even though not itself concerned with the first-order permissibility of a course of action, the verdict of conscience still has important consequences for an action's moral value.

In addition, it is important to note that, despite its exclusively second-order function, the relevance of conscience, for Kant, is not restricted to the very rare occasions where we engage in abstract self-reflective assessment of our practices of moral deliberation. On the contrary, due to the normative significance of conscience we have just pointed out, it is implicated in ordinary moral deliberation.

50 In Knappik/Mayr 2013, we argue on the basis of these and similar passages that the shift in Kant's conception of conscience should be seen in the context of a new tendency in Kant's late writings to treat untruthfulness and dishonesty ("the radical evil": R 8103 [after 1789], 19:646, our translation) as fundamental phenomena of immorality.

The exercise of conscience and the exercise of our ability for first-order moral evaluation are tightly interwoven,[51] as long as the agent is sensitive to the normative demands of conscience.

Put very briefly, the interplay between conscience and first-order moral judgment can be understood as follows. At the first-order level, we aim to decide whether a particular course of action is morally permissible or not. While according to Kantian moral theory, universalizability provides a clear-cut criterion for moral permissibility, the application of this criterion to concrete situations is notoriously problematic. In particular, agents have to find out (a) what features of the scrutinized course of action and of its circumstances are relevant to the formulation of a corresponding candidate maxim and (b) whether the resulting candidate maxim is actually universalizable or not. Resolving each of these questions can require a significant amount of reasoning, and competing considerations can speak for and against particular answers to each question. It is the task of moral deliberation to examine these considerations. If this examination leads to an unambiguous result, the process of deliberation should lead to the adoption of an affirmative attitude towards a proposition that assigns a moral evaluation (permissible/not permissible) to the course of action under scrutiny; otherwise, the agent has to leave the issue suspended.

Since Kant also thinks, as we have seen, that it is morally forbidden to perform an action of which we are not entirely sure that it is morally permissible, it is crucial that the process of first-order deliberation is *diligent*, i. e., that the agent takes into full account all available considerations that may speak for or against the permissibility of the action in question. While this diligent examination of the case is part of the first-order process of moral evaluation, it is the task of conscience to examine, in its turn, whether such diligent first-order examination has actually taken place (or is taking place), and to make sure that the process of deliberation is not concluded, and no definitive evaluation of the action in question is adopted, *before* the case has been examined with the due care.[52] Before this second-order judgment on whether we have carried out a diligent examination has been made (which, for reasons we will discuss in the next sections, is tantamount to the assessment of whether we are in a position to issue a genuinely

51 This view is supported by the fact that Kant sometimes describes conscience as a reflexive self-application of the *same* faculty that issues the first-order judgment (e. g., in quotation (D) from Section 2, and at MS 6:400).

52 It is natural to assume that in addition to its cognitive role, conscience has also a motivational function in this regard. *Pace* Guyer 2012, 143 (see footnote 6), this would be a motivational function that intervenes in the process of moral deliberation, rather than providing for the transition between a moral judgment and an action which is based on it.

moral judgment or not), we *cannot* bring the first-order process of moral deliberation to a close without being liable to the criticism of lacking conscientiousness. Hence, in agents who are conscientious, the exercise of conscience must directly accompany, and intervene in, the process of first-order moral deliberation.

5 A New Account of Kant's Infallibility Claim

On the reading that we have introduced and defended so far, the key change that leads to Kant's late view of conscience consists in replacing actions with judgments as the objects of the examination of conscience. In addition, we have also seen that from 1791 onwards, Kant comes to claim that conscience is infallible, while his earlier view took conscience to be capable of error. If the above reading is sound, it is very natural to see this further change of mind as a consequence of Kant's radical re-interpretation of conscience. Apparently, Kant took his new, second-order account of conscience to rule out any potential source of error for the verdicts of conscience.

It is indeed easy to see why the possibilities of error that had been acknowledged by Kant's earlier account, as well as by traditional accounts of conscience, no longer apply on Kant's new view: if conscience does not assess actions at all, neither misdescriptions of actions nor invalid applications of general norms to concrete actions can lead to errors in the verdict of conscience anymore. But of course, this does not entail that there could not be *new* sources of error that might lead to mistaken second-order verdicts of conscience. We have seen already in our discussion of Ware's interpretation in Section 3 that it is far from trivial how infallibility can be claimed for conscience *even* if it is understood in second-order terms.[53] In the next two sections, we will discuss in detail why Kant can nevertheless claim infallibility for conscience on his new conception.

53 Moyar's sophisticated second-order interpretation of Kant's account of conscience (in Moyar 2006) ultimately faces similar problems as Ware's reading. According to Moyar, Kant takes conscience to be responsible for (1) an act by which we judge that our deliberation has been sound and complete, and thereby close the deliberation (343), and (2) a practical apperception or self-consciousness that accompanies all practical deliberation and imputation. This latter point is based on Kant's characterization of conscience as apperception in V-MS/Vigil 27:613 f.: "*Conscientia*, taken generally, is the consciousness of our self, like *apperceptio*; in specie it involves consciousness of my will, my disposition to do right, or that the action be right, and thus equals a consciousness of what duty is, for itself". (For Moyar, the notion of conscience "in specie" corresponds to the first of the two functions he ascribes to Kantian conscience: Moyar 2006, 350.) Since Moyar thinks that the second-order acts of conscience in its first function collapse into (fal-

Before we turn to this discussion, some preliminary clarifications are in place. First, we will focus in the next two sections on infallibility with regard to one's *present* moral judgments. We will argue later (in Section 8) that this form of infallibility covers only cases in which the verdict of conscience precedes or accompanies the corresponding action, not cases in which conscience issues its verdict *after* the action. However, we postpone discussion of this ex-post situation until Section 8, as it creates additional problems that we wish to bracket for the moment.

Second, Kant's notion of "judgment" is notoriously ambiguous (among other uses) between an episodic *act* through which we adopt a propositional attitude, on the one hand, and a *standing propositional attitude*, on the other hand.[54] For

lible) first-order judgments (345–347), he argues that Kant's infallibility claim refers to the role of conscience as practical self-consciousness (351 f.). Moyar holds that the basis for infallibility lies in a constitutive relation between apperceptive conscience and first-order judgments (352), assuming that apperceptive conscience is *necessary* (but not sufficient) for first-order judgments. But as we have already objected against Ware's reading, infallibility would require that the converse conditional obtains, too – that if I believe that I make a practical judgment *p*, this belief is true; and Moyar's reading fails to explain why Kant thought he was entitled to this further claim. – Regarding Moyar's reading in general, we agree that conscience is a presupposition for moral judgment and, hence, deliberation; however, as we shall argue in Section 7 below, this is just a consequence of its role as capacity for particular higher-order judgments. And as these judgments on our reading do not themselves close the deliberation, they do not collapse into first-order judgments. As to Kant's remark on conscience as practical apperception in *Metaphysics of Morals Vigilantius*, we see this passage as reflecting on the fact that '*conscientia*' in Early Modern usage had become ambiguous between 'conscience' and 'consciousness' – a fact that had already been commented upon by Crusius (*Anweisung*, 177 f. (§ 132)). Therefore, the passage from *Metaphysics of Morals Vigilantius* might merely try out one way of disambiguating between both notions in terms of a genus/species-ordering.

54 Cf. Chignell 2007, 35, for still more senses of the term in Kant. – A clear instance of the attitudinal usage (which is not mentioned by Chignell) is found at Log 9:65 f.: "[...] the judgment through which something is *represented* as true, the relation to an understanding and thus to a particular subject, is, *subjectively*, assent [*Fürwahrhalten*]" – where Kant's further remarks on assent (Log 9:66, KrV B850) show that by this term, he refers to a class of doxastic attitudes. (We follow Young's change of punctuation for the passage at Log 9:65 f. vis-à-vis the Akademieausgabe, as the full sentence makes sense only if parsed in the way proposed by Young.) Regarding the subject-matter of conscience, an attitudinal reading is suggested, e. g., by Kant's claim that "the formal conscientiousness which is the ground of truthfulness consists precisely in the *care in becoming conscious* of this belief (or unbelief) [sc. the first-order (un)belief "to be right"] and not pretending to hold anything as true we are not conscious of holding as true" (MpVT 8:268). In this passage, Kant characterizes the subject-matter of conscience as an instance of belief and of holding-true, and hence, as an attitude. In the same context (namely, in the last sentence of passage (A) from Section 2), he treats the second-order claim "that I judge the object in such-and-such a way" (a claim about my first-order *judgment*) as equivalent to a second-order claim

the sake of simplicity, we will formulate our interpretation of Kant's infallibility claim in the next two sections for judgments in the *attitudinal* sense. This interpretation can be extended easily to judgments-qua-acts. For a capacity to know our present propositional attitudes at the same time enables us to detect present changes in those attitudes, and hence also to know when a judgment-qua-act (i. e., an act in which we adopt a new judgment-qua-attitude) occurs. We will use "judgment$_{ATT}$" for the standing-attitude reading of Kant's "judgment". In order to express that someone *has* a relevant propositional attitude, we will say that he *holds the judgment*$_{ATT}$ in question.[55]

Given these clarifications, we now have to ask how Kant could claim that conscience provides us with infallible second-order judgments$_{ATT}$ on the questions that were covered by Kant's infallibility claim in the passages we have quoted in Section 2, i. e., on (a) whether we presently hold a first-order moral judgment$_{ATT}$ regarding the moral quality of a given action and (b) whether this first-order judgment$_{ATT}$ is based on careful deliberation.[56]

6 The Specific Certainty Required by Moral Judgments

It may be tempting to understand Kant's claim that we cannot be in error about our moral judgments$_{ATT}$ as following from a general point that our judgments$_{ATT}$ (and presumably judgments-qua-acts, too) are unproblematically accessible to us. Indeed, Kant seems to have assumed that whenever we hold a judgment$_{ATT}$, we are in a position to know that we do, and that, *vice versa*, our self-ascriptions of

about the question "*whether I in fact believe* to be right" (a claim about my first-order *attitude*). By contrast, a reading in terms of judgments-qua-acts is suggested when Kant characterizes, in passage (C) from Section 2, the subject-matter of conscience in terms of the question "whether I have submitted it [sc. the object of the first-order judgment] to my practical reason (here in its role as judge) for such a judgment", and he adds: "if I could be mistaken in that, I would have made no practical judgment at all [*praktisch gar nicht geurtheilt haben würde*]" (MS 6:401). This formulation suggests that Kant thinks here of judgment as a dated event, rather than as a standing attitude.

55 A judgment$_{ATT}$ that p in this sense is roughly what we would nowadays normally call a *belief* that p.

56 For the sake of the following discussion, we will assume that a subject A is infallible about a proposition p iff the following four conditionals hold: (1) (A believes that p) → p; (2) (A believes that ~p) → ~p; (3) p → (A believes, or forms the belief if prompted, that p); (4) ~p → (A believes, or forms the belief if prompted, that ~p).

judgments$_{ATT}$ are always true.[57] As we will see in the next section, such a general claim about our knowledge of our own judgments$_{ATT}$ must indeed play *some* role in a full account of the reasoning behind Kant's infallibility claim. However, it cannot on its own justify the infallibility claim. The reasons for this are connected to particular features of Kant's technical notion of a moral judgment. We will examine these features in some detail, as they will also be important for our positive account.

Remember that Kant had used diverging ways of specifying the subject-matter of the examination of conscience. Sometimes, he had done so by means of the question (1) whether we hold a first-order moral judgment$_{ATT}$. On other occasions, he had specified that subject-matter in terms of the different question (2) whether we have performed a "diligent examination" of the case. Nowhere does Kant indicate that those questions give only *partial* specifications of the subject-matter of conscience. At the same time, Kant can hardly have held that both questions are equivalent, either for judgments$_{ATT}$ in general, or even for moral judgments$_{ATT}$ in particular. For we can perform a diligent examination of the case without actually holding a judgment$_{ATT}$. For instance, we might withhold judgment about *p* at the end of our examination because we realize that the available considerations are insufficient to warrant either a judgment that *p* or a judgment that non-*p*. And

57 Thus, Kant argues in the *Theodicy* essay that we have an "immediate consciousness" of whether our assertoric utterances are *truthful*: "[...] one can and must stand by the truthfulness of one's declaration or confession, because one has immediate consciousness of this". For "where we declare what we hold as true, we compare what we say with the subject (before conscience). Were we to make our declaration with respect to the former without being conscious of the latter, then we lie, since we pretend something else than what we are conscious of" (MpVT 8:267). What we are "conscious of" in this picture can only be our judgments$_{ATT}$, or "what we hold as true" – for it is these attitudes that we express, or fail to express, in our assertoric utterances. Our "immediate consciousness" of the truthfulness of our utterances presupposes that at least upon reflection ("before conscience"), all our judgments$_{ATT}$ are conscious, and all consciousness of judgments$_{ATT}$ is veridical (otherwise, Kant would have to allow for the possibility that one unknowingly makes a wrong statement about one's attitudes). It follows that in general, Kant assumes an unproblematic safe knowledge about our judgments$_{ATT}$. – It is difficult, though, to make out an argument for this specific assumption, or an account of the knowledge in question, in Kant's writings (cf. also footnote 62 below). It is tempting to read Kant's claim that "The *I think* must be able to accompany all my representations" (KrV B131) in the B-deduction (and his theory of apperception more generally) as implying that whenever I have a particular representation (such as a judgment$_{ATT}$), I am able to tell that I have this representation (cf., for instance, Heidemann 2012, 51, about the relation between the 'I think' and cognitive access to our representations). But even on such a reading, no infallibility would result, since infallibility would also require the converse connection (whenever I believe that I have a particular judgment$_{ATT}$, that belief is true), and Kant's theory of apperception does not yield this further connection.

conversely, we clearly seem to be able to form at least *some* judgments$_{ATT}$ without diligent examination. How can Kant then use both questions (1) and (2) to specify the subject-matter of conscience?

The most charitable way to solve this puzzle, we propose, is to assume that for Kant, diligent examination is a *necessary* (though not sufficient) condition for holding a genuinely *moral* judgment$_{ATT}$. Kant would then be able to use both questions (1) and (2) to specify the subject-matter of conscience because in order to know the answer to (1), we have to know the answer to (2). In addition, it should be noted that such a picture only makes sense if it is also assumed that, in an actual instance of moral judging, there has to be a *connection* between the antecedent examination of the case (which will take the form of moral deliberation) and the moral judgment$_{ATT}$: the moral judgment$_{ATT}$ has to conform with, and be based on, this examination.[58]

If we take these observations together, Kant seems to subscribe to the following claim:

(α) For all propositional attitudes *x* of a subject *A*: *x* is a moral judgment$_{ATT}$ → *x* conforms with, and is based on, a diligent examination of the case on *A*'s part.

This means that Kant's notion of a moral judgment$_{ATT}$ is surprisingly demanding. In order to hold a moral judgment$_{ATT}$ at all, it is not enough to hold some judgment$_{ATT}$ with a moral content. Rather, particular requirements have to be met which do not apply in the case of judgments$_{ATT}$ in general – since there is no reason to exclude that some judgments$_{ATT}$ are formed rashly and without sufficient reasons. Due to these special conditions, we cannot expect to be able to explain the supposed infallibility of conscience merely on the basis of the general point about the accessibility of our judgments$_{ATT}$ we have mentioned earlier. For on the emerging picture, it is possible that we hold a judgment$_{ATT}$ regarding the moral permissibility of an action, without thereby already counting as holding a *moral* judgment$_{ATT}$.[59]

This conclusion is reinforced by Kant's further claim that moral judgments$_{ATT}$ must be entirely *certain*. Thus, in the *Logik Jäsche*, Kant states that in moral matters, "[o]ne has to be completely certain: whether something is righteous or

58 Kant does not seem to mind the consequence that actions can be based on moral judgments only when they are preceded by actual deliberation.

59 Note that exactly the same problem arises if the subject-matter of conscience is understood in terms of moral judgments-qua-acts. Moral judgments in *this* sense are acts in which we adopt a moral judgment$_{ATT}$. Therefore, a judgment-qua-act cannot count as a *moral* judgment-qua-act unless the attitude that we adopt in this act satisfies the conditions for being a moral judgment$_{ATT}$.

not, in accordance with duty or not, permissible or not" (Log 9:70). Similarly, he writes: "[W]here does mere opining really occur [*findet statt*]? Not in any sciences that contain cognitions *a priori*, hence neither in mathematics nor in metaphysics nor in morals, but merely in *empirical* cognitions: in physics, psychology, etc." (Log 9:67). As this latter passage shows, Kant is not merely making a normative point here about what features our moral judgments *ought* to have. He makes a claim about conditions that a judgment *must* fulfil in order to count as a moral judgment (although we will see below that this claim *also* has normative implications). Hence, we can conclude that unless one holds a judgment$_{ATT}$ with complete certainty, it cannot be, strictly speaking, a *moral* judgment$_{ATT}$ at all:

(β) For all propositional attitudes x of a subject A: x is a moral judgment$_{ATT}$ → A holds x with complete certainty.

This claim can seem at first quite bewildering. As our subsequent argument will heavily draw on it, it will be helpful to make the claim intuitively plausible by considering an analogous case. In the *Critique of Pure Reason*, Kant claims that mathematical judgments$_{ATT}$, too, require complete certainty.[60] In that case, his view is directly plausible, as we can see from the following example. Imagine someone who is committed to the truth of '20+20=40' only because he has 'inductively checked' this sum by counting different collections of objects, and thinks it is likely to hold for other collections as well. This person would not hold a *mathematical* judgment$_{ATT}$ at all. For the latter requires a certainty that probabilistic assessments lack.

What would an analogous case, in which the certainty that is required for Kantian *moral* judgments$_{ATT}$ is undermined, look like? Given what we have said above about the relation between such moral judgments$_{ATT}$ and previous deliberation, it is fair to assume that the requisite certainty would be undermined, e. g., if one were aware of there being considerations which speak against the permissibility of a course of action and which one has not ruled out yet. Imagine that a soldier has been ordered by his commanding officer to put some civilians who have not participated in any fighting into prison. Let us assume for the sake of argument that the officer should (morally) normally obey his officer's commands and is aware of that. Still, the soldier is also aware that the officer may lack the relevant authority to order the imprisonment of non-combatants in cases like the present ones, because it is a grave breach of international conventions to do the latter, and that the imprisonment will cause very serious harm. Should he there-

60 KrV B 851; see also Log 9:69.

fore refrain from following the order? Or should he simply act on it, because he is permitted to trust the officer's authority and the harm which will be caused is irrelevant when he is obeying orders? As long as the soldier is aware that the proposed course of action raises such questions, but has not really thought through and settled them, any judgment he may form about the matter would not count as *moral* judgment$_{ATT}$ in Kant's sense.

Thus, Kantian moral judgments$_{ATT}$ require not only previous careful examination of the case, but also a certainty that one would lack if one were aware of considerations to the effect that the judgment$_{ATT}$ in question can be wrong. This makes it even clearer that no general point about our epistemic access to our judgments$_{ATT}$ can suffice to account for infallibility about whether one holds a moral judgment$_{ATT}$, since clearly many judgments$_{ATT}$ are held while we lack such certainty.

7 Reconstructing Kant's Infallibility Claim

We have seen in the previous section that moral judgments$_{ATT}$ have to be based on a "diligent examination", and require a specific certainty. If anything, this raises the bar for providing a satisfactory explanation of how we could be infallible about whether we actually hold a moral judgment$_{ATT}$ even higher. For could we not be mistaken or self-deceived about whether we have really conscientiously examined the case, or about whether our judgment is sufficiently certain?[61]

Nevertheless, we do believe that Kant's infallibility claim is based on an argument which is compelling – given some of Kant's other background assumptions – and which can be reconstructed once the special features of Kant's technical notion of a moral judgment$_{ATT}$ are better understood. To see what issues such a reconstruction has to tackle, remember that, according to what we have said so far about Kant's technical notion of a moral judgment$_{ATT}$, three factors are individually necessary for a moral judgment$_{ATT}$: (1) the agent needs to hold a judgment$_{ATT}$ with a relevant content (such as the moral permissibility of a particular action); (2) the judgment$_{ATT}$ must come with the right kind of certainty; and (3) the judgment$_{ATT}$ must conform with, and be based on, diligent examination of the case. In addition, Kant seems to hold that these conditions are jointly sufficient, too, for a moral judgment$_{ATT}$.

61 For example, it might be thought that I can successfully deceive myself, at times, about whether I have taken all the relevant considerations into account, when I have really ignored objections in order to persuade myself that what I want to do is permissible (cf. Wood 2008, 191).

A full account of Kant's infallibility claim therefore has to explain how we can infallibly know whether the conjunction of (1), (2), and (3) obtains or not. This does not necessarily require infallibility about each element, taken on its own. As we are going to argue, Kant can make a case that we infallibly know whether each of elements (1) and (2), taken in isolation, obtains. By contrast, we will leave open whether we are actually infallible about element (3), when taken in isolation. However, we will argue that element (2) – certainty – entails element (3), or diligence. As we will see, this suffices to account for infallibility about whether *the whole package* of elements (1) to (3) obtains or not. Let us go through each of the elements in turn.

(1) Regarding the obtaining of a judgment$_{ATT}$ with a relevant content, we have already seen in the previous section that Kant assumes unproblematic infallible knowledge here.[62] While such knowledge about our judgments$_{ATT}$ in general does not on its own enable us to tell whether we have actually formed a specifically *moral* judgment, it at least enables us to tell whether we hold *some* judgment$_{ATT}$ about the moral quality of the course of action at hand. In order to decide whether this judgment$_{ATT}$ *is* actually a moral judgment$_{ATT}$, we must, in addition, be able to tell whether conditions (2) and (3) also obtain. But *if* this further knowledge can be accounted for, then infallibility about *moral* judgments can be explained by simply adding this further account to the assumption of unproblematic infallibility about judgments$_{ATT}$ in general.

(2) We next turn to infallible knowledge about the second factor that is required for Kantian moral judgments$_{ATT}$, *certainty*. Before we can ask how such knowledge is possible, we first have to get clearer about the *precise kind of certainty* that is needed for moral judgments$_{ATT}$. A good place to start for this question is the section "On Having an Opinion, Knowing, and Believing" in the first Critique, where Kant famously distinguishes between three different modes of assent: opinion, belief, and knowledge. In that context, he identifies the *certainty* that is required both for mathematical and moral judgments with the "objective sufficiency" of one's

62 While such an assumption was common in Kant's days, it has, of course, come under massive attack since. To address the question whether Kant has any resources for a defence of the view against its modern critics would by far go beyond the scope of this paper. It must suffice to note here that Kant's brief discussion in the *Theodicy* essay (see footnote 57) can be read as drawing on normative considerations: we ought to be truthful, and hence must have unproblematic access to our own judgments$_{ATT}$. The most elaborate contemporary defence of the infallibility claim about beliefs, Bilgrami's (avowedly Kantian) account in Bilgrami 2006, is based on related considerations, and therefore provides a congenial extension of Kant's position.

grounds for judgment. At the same time, Kant defines *knowledge* as the combination of such objective sufficiency or certainty, and "conviction", or the "subjective sufficiency" of one's grounds of judgment.[63] Kant seems to assume that certainty or objective sufficiency presuppose conviction or subjective certainty: the case of certainty *without* conviction does not occur in his taxonomy. But if conviction plus certainty are sufficient for knowledge, and certainty itself is already sufficient for conviction, it follows that whenever we have certainty, we also have knowledge. As a consequence, whenever we hold a genuine moral judgment$_{ATT}$, and hence possess the certainty that is required for a moral judgment$_{ATT}$, we *also* possess knowledge. And, indeed, Kant declares that it would be "absurd to have an opinion in pure mathematics: one must know, or else refrain from all judgment. It is just the same with the principles of morality, since one must not venture an action on the mere opinion that something is *allowed*, but must know this" (KrV B851).[64]

At first sight, this seems to make Kant's position even more puzzling. For knowledge, in the ordinary understanding, requires true belief. But if moral judgment$_{ATT}$ requires knowledge, and hence, truth, how can we be infallible about whether we hold a moral judgment$_{ATT}$? We are clearly fallible in these latter judgments – as Kant himself explicitly admits (even in his later writings, where he defends the infallibility claim for conscience). Thus, we certainly can *think* that some attitude is a case of knowledge, and that it comes with the kind of certainty required for this, while it actually is merely false belief or opinion.[65] But if the relevant kind of certainty is connected to knowledge, how could we then be infallible about whether this certainty is present or not?

However, there is reason to believe that Kant is actually relying here on quite *different* notions of knowledge and certainty than it appears at first sight. First, it is possible to read the relevant notion of *knowledge* in this context as *non-factive*, that is, as not entailing the truth of the relevant proposition. Such a reading would make Kant's formulation in the passage from B 851 compatible with his admission of the possibility of false moral judgments – a possibility that would be excluded by the account of certainty in "On Having an Opinion, Knowing, and Believing", if the notion of knowledge in that section were factive. In addition,

63 KrV B 850 f.

64 In similar passages, Kant explicitly extends the point to moral judgments (as opposed to actions): cf. V-MS Vigil 27:615; RGV 6:186; Log 9:67. – We had already seen above that Kant is not merely making a normative point here.

65 A similar problem arises for Chignell's recent interpretation, according to which the "objective sufficiency" of assent consists in possession of a psychological state (or relevant kind of access to external states) that makes the proposition that is assented to *objectively* likely (Chignell 2007, 42).

there is a further point in favour of a non-factive reading of "knowledge" in this context. As we have argued before, Kant's claim that moral judgments have to be certain, and instances of knowledge, is a claim about the conditions that a judgment has to fulfil in order to count as a moral judgment. But at the same time, this claim has normative implications: we are normatively required to *treat*, or propose as, moral judgments only judgments that fulfil the demanding conditions in question (if only because we have a *duty* to judge conscientiously in moral matters). Now this normative requirement presupposes itself that we are able to *tell* whether those conditions are fulfilled. Since these conditions, according to the passage from B 851, include that the judgment has to be a case of *knowledge*, it follows that we have to be able to tell whether a candidate judgment is a case of *knowledge* or not. But this would hardly be a reasonable presupposition if "knowledge" were understood in a factive sense. For as we have already pointed out, it seems clear that there are many cases in which we merely *think* that we know something, while actually having a false belief – that is, cases in which we are *not* able to detect whether we are in a situation of knowledge (in the factive sense) or not.[66] By contrast, Kant's claims in this context make perfect sense if the relevant notion of knowledge is given a non-factive reading.[67]

Acknowledging this point is necessary for an adequate understanding of Kant's notion of *certainty* in this context, too. For once the non-factive reading of "knowledge" is accepted for this context, it follows that certainty cannot require truth here, either.[68] What is needed instead becomes clearer in several passages in which Kant relates his threefold distinction of degrees of assent – opinion, belief and knowledge – to the *modalities of judgment*. In these passages, assent in the case of *knowledge* – and hence in the case that comes with the certainty required for moral judgments – is called *apodictic* assent. Thus, Kant writes:

66 Note that in order to defend the factive reading of "knowledge" in this context, it would not suffice to ascribe to Kant the view that we sometimes know that we know that *p*, or even the view that whenever we know that *p*, we know that we know that *p*. Rather, Kant would have to be read as holding the even stronger view that (at least with regard to moral judgments) if someone believes that *p* but this belief is *not* a case of knowledge, he knows this to be so. In that case, there could be no warranted false beliefs, and it would be difficult to allow for any false beliefs at all – for one could then reach infallibility (also at the level of *first-order* judgments) simply by avoiding beliefs of which one knows that they are not cases of knowledge.

67 It should be noted that parallel points apply to mathematical judgments. For mathematical reasoning is certainly not immune to error, while, at the same time, Kant's discussion of mathematical judgments at B 851 clearly implies that there is a (non-moral) requirement to only adopt judgments as mathematical ones when the conditions for knowledge are met.

68 Nor can it require a strong *objective* likelihood of truth, given one's psychological states (as Chignell would hold: see footnote 65).

> [K]nowing is *apodictic* judging. For [...] what I *know* [...] I hold to be *apodictically certain*, i. e., to be universally and objectively necessary (holding for all), even granted that the object to which this certain holding-to-be-true relates should be merely an empirical truth. (Log 9:66)

And again:

> What I know: [sc. I judge] as apodictic according to laws of understanding; even if the truth is only empirical, the holding-to-be-true (relation to the ground of cognition) is apodictic, i. e. universally necessary (holds for all). (R 2474, 16:385)[69]

Thus, knowledge in the relevant sense requires *apodictic* assent. In R 2474, such apodictic assent is explained, in its turn, as assent that is itself *necessary* for everyone. Similarly, Kant points out in his discussions of the modalities of judgment that judgments are apodictic if *we have a* "*consciousness*" *of the* "*necessity*" *of judging* (Log 9:108, emphasis added), or if "assertion or denial [...] is seen as *necessary*" (KrV B 100). The relevant necessity is most plausibly read as a matter of a *rational obligation* to endorse the relevant judgment.[70]

We can thus put Kant's point at Log 9:66 and in R 2474 in the following way: In the case of knowledge and certainty, we must always be aware of being rationally compelled to assent (since the relevant certainty just consists in this awareness). Importantly, assent can be apodictic in this sense even if the proposition in question is actually false. Kant's rationale for relying on this notion of certainty for certain kinds of judgments seems to be the following: while there are subject-matters where we are permitted to assent to a proposition even if a suspension of judgment would be equally warranted in the light of our deliberation, we can hold *moral* judgments$_{ATT}$ – just as mathematical ones – only if we are aware of being rationally *obliged* to judge in that way.

69 Our translation.

70 In this respect, we follow Mattey 1986, 426–8, who explicitly reads the apodicticity of a judgment in the sense that one has a consciousness of being rationally obliged to endorse it. Leech 2012, 279, criticizes this interpretation, and argues instead that the modalities of Kantian judgments are merely a matter of their position in a given piece of reasoning. A judgment would then be apodictic if it figures as conclusion of an inference. But at least when Kant relates the modalities of judgment to the different degrees of assent, he has to presuppose that the modality of a judgment is evaluated with regard to its place in our *overall* system of propositional attitudes. Otherwise, it would be very easy to turn mere opinion into knowledge: for we can formulate formally valid arguments for every possible judgment, as long as it is not self-contradictory (cf. Mattey 1986, 273 f.). Hence, at least in this specific context, the apodicticity of a judgment has to require a consciousness of being rationally obliged to endorse the judgment, given one's available reasons and background beliefs.

This account of Kant's notion of certainty will prove absolutely central to the rest of our argument. Before we go on, we should therefore address in some detail an objection that this proposal is likely to provoke: namely the objection that this proposal appears to conflate Kant's notion of *knowledge* with his notion of *belief*. Is the awareness of a rational obligation to assent that we have described not merely a case of *subjective* sufficiency of the grounds of judgment, or a case in which the *subject* is certain that *p*? Would the *objective* sufficiency that, according to Kant, characterizes knowledge not have to require a *factive* form of assent, or require that *it* is certain that *p* (cf. KrV B 857, Log 9:72)?

To these worries we reply, first, that Kant clearly considers the characterizations of knowledge vs. belief in terms of objective vs. merely subjective sufficiency, and in terms of apodictic vs. assertoric assent, as equivalent: in several passages, he uses these two ways of demarcating the two forms of assent interchangeably. For instance, in the *Logic Dohna-Wundlacken*, Kant defines knowing both as "holding-to-be-true that is sufficient both subjectively and objectively", and – in the same passage – as a case in which "the holding-to-be-true is [...] *apodeictic*" (24:732) Similarly, he points out that "in belief the holding-to-be-true is [...] *assertoric*. For I say that it is sufficient for me <subjectively>, but I do not settle whether it is objectively sufficient" (24:732).[71] If Kant treats the characterization of knowledge and belief in terms of the modalities of judgment and in terms of the objective vs. merely subjective sufficiency of the grounds of judgment as equivalent, it follows that a judgment's being objectively sufficient is already guaranteed by its apodictic character. And since Kant, as we have seen, identifies the apodictic character of a judgment with an awareness of a necessity to judge, it follows that all that is needed in order for a judgment to be based on objectively sufficient grounds – and hence, to count as an instance of knowledge – is that it comes with an awareness that one *has* to judge this way.

Furthermore, although on our reading, Kant's notions of knowledge, objective sufficiency and certainty turn out to be more 'subjective' than one might think at first glance, this reading can nevertheless distinguish (non-factive) knowledge from belief in Kant's sense. For on this reading, it is natural to hold that objective sufficiency of the grounds of an apodictic judgment is 'objective' in so far as when such sufficiency obtains, we take ourselves to have *objectively* valid, or truth-conducive, reasons that make it obligatory for us to assent to the proposition in question.[72] By contrast, *mere* subjective sufficiency would obtain if we lacked such

71 Cf. also R 2459 (16:378), R2473 (16:385).

72 Cf. Log 9:70, where Kant talks in a related context of "objective grounds of truth that are independent of the nature and the interest of the subject".

reasons (even by our own lights) and based our judgment on reasons that we do *not* take to be truth-conducive. In the case of theoretical judgments, these would be prudential or moral considerations, as is shown by Kant's discussions of pragmatic and moral belief (KrV B852–857). In the case of judgments with a moral content, prudential reasons – for instance, fear of punishment – might figure as merely subjectively sufficient reasons. On our reading, it is this latter case of an assent based on reasons which we do *not* ourselves take to be truth-conducive that Kant characterizes by saying that, in such cases, *I* am certain that *p*. By contrast, when we take ourselves to have reasons for assent to *p* that actually *show* that *p* must be the case, it is appropriate for us to say that *it* is certain that *p*.

We therefore can conclude that the objection we have addressed can be overcome, and that the technical notion of certainty that is relevant for Kant's account of moral judgments$_{\text{ATT}}$ can indeed be understood in the sense that an agent is aware of being rationally obliged – or, as we will put it in the following, that she is rationally obliged 'as far as she can tell' – to assent to a proposition, independently of whether this proposition is really true or not.

Before we can turn to the question of how we can be infallible about certainty in this sense, we have to clarify two further points regarding the relevant notion of certainty. While the subject-relative qualification in this account of certainty – rational obligation 'as far as the agent can tell' – clearly allows that an agent may have such certainty if the proposition in question is false, the qualification can itself be understood either in a factive or in a non-factive way concerning the presence of *rational obligation*. Understood factively, it implies that the rational obligation in question actually obtains – if not absolutely speaking (i. e., when considering the fact of the matter without taking into account the agent's limited epistemic perspective),[73] at least by the light of the reasons that are *de facto* available to the agent. Understood non-factively, it has no such implication. For instance, there might merely *appear* to be a rational obligation because the agent has made a logical mistake in figuring out what follows from his further commitments. Kant seems to understand certainty in a way that corresponds to the second reading. In his discussion of the modalities of judgment at KrV B100 from which we have quoted above, Kant writes that in the case of apodictic judgment, assertion or denial of a proposition is "*seen*" as necessary, and in the parallel cases of the other modalities, he uses "regards" and "considered" (KrV B100, our emphasis). These are all terms that strongly suggest the non-factive reading. We

73 For it may be impossible that there are compelling reasons from this perspective for a judgment that is false.

therefore should conclude that the condition of being rationally obliged, *as far as one can tell*, should be understood in a non-factive way in the above sense.

Moreover, the subjective qualification 'as far as one can tell' means that the agent has some form of *access* to the (presumed) rational obligation. This access can be provided by two different mental states. One way of having it is by *holding a judgment*$_{ATT}$ that one is rationally obliged to assent to the proposition in question.[74] Another way of having such access is by experiencing, upon considering *p*, an epistemic *feeling of certainty* regarding that proposition. Kant's account of certainty provides no reason to restrict certainty to either of these options. In the following discussion, we will therefore be similarly liberal about this point[75].

Given these clarifications, we can summarize the notion of certainty which we will take to be the relevant one for Kant's theory of moral judgments as follows:

(γ) A proposition *p* is certain for an agent *A* iff *A* is, as far as she can tell, rationally obliged to hold the judgment$_{ATT}$ that *p*.[76]

As we have seen, a proposition can be certain for an agent in this sense regardless of (a) whether *p* is actually true; (b) whether it is de facto rationally obligatory to judge that *p*, absolutely speaking; and (c) whether it is de facto rationally obligatory *for the agent* to judge that *p*, given the reasons that are available to her.

We now can finally begin to see how Kant's view that agents have infallible knowledge of the certainty needed for moral judgment can be defended. To establish such infallibility, we will need to show that our belief about the presence or absence of certainty, on the one hand, and the actual presence or absence of certainty, on the other hand, cannot come apart. Therefore, we have to show that each of the following four conditionals hold:

(a) whenever a proposition *p* is certain for an agent, the agent believes (or forms the belief, if prompted) that *p* is certain for him;

74 It is true that infallible knowledge about the presence of such judgments$_{ATT}$ is already accounted for by the general infallibility about judgments$_{ATT}$ that Kant, as we have seen, assumes. However, this does not suffice to account for infallibility about certainty precisely because the access in question can also take other forms than that of a judgment$_{ATT}$.

75 Of course, this raises the question whether different modes of such access can be in conflict with each other, and if so, how such conflicts can be resolved. We will come back to these questions presently.

76 In some passages – for instance, Log 9:66, quoted above – Kant presupposes that certainty, in addition, requires that one actually assents to the proposition. As infallibility about this additional point is already covered by our discussion of the first factor of moral judgment above, we can abstract from it for the present purpose.

(b) whenever an agent believes that a proposition *p* is certain for him, *p* is certain for him;
(c) whenever a proposition *p* is uncertain (i. e., not certain) for an agent, the agent believes (or forms the belief, if prompted) that *p* is uncertain for him; and
(d) whenever the agent believes that a proposition *p* is uncertain for him, *p* is uncertain for him.

Let us go through each of these conditionals in turn. For the purpose of presentation, we will start with (d). Can an agent believe that *p* is uncertain for him, while at the same time, *p* is actually certain for him? Given our definition of certainty, this would amount to a situation in which it is rationally obligatory for the agent, as far as he can tell, to assent to *p*, while at the same time, the agent believes that it is actually *not* rationally obligatory for him to assent to *p*. Can such a situation arise? Brief reflection shows that it cannot. Imagine, for illustration, that you have rehearsed and examined a complicated piece of reasoning a couple of times, and, when asking yourself whether it is, as far as you can tell, rationally obligatory for you to assent to *p*, you conclude that it is not. In that case, it seems that it cannot *still* be obligatory for you, *as far as you can tell*, to assent to *p*.

Why is the negative belief 'As far as I can tell, it is not rationally obligatory for me to assent to *p*' self-validating in this way? Certainty would require that the agent is aware of being rationally compelled, which is incompatible with his being, at the same time, aware of reasons which make the possibility of non-*p* a genuine option. Now a belief that it is not rationally obligatory, as far as one can tell, to assent to *p* is simply *one form* of being aware of the existence of such reasons. Having such a belief implies that, as far as one can tell, there are considerations against *p* that one has not ruled out yet, or that one has doubts or hesitations about *p*. (If one did not believe some of the latter options, one would take one's own higher-order judgment about the lack of rational obligation to be unwarranted, and would normally abandon it immediately.) As a consequence, the higher-order judgment is itself sufficient to establish uncertainty – a condition in which it is not rationally obligatory, *as far as the agent can tell*, to assent to *p*.

Given our earlier point that there can be different modes of access to a presumed obligation to assent to a proposition *p*, one may wonder, though, whether these modes cannot be in conflict with each other. In particular, could there not be a situation in which the agent believes that there are reasons which speak against *p*, while at the same time, she experiences a strong sense of certainty regarding *p* while deliberating about that proposition? In that case, the agent would be at the same time *also* aware of certainty regarding *p*. Would this not render the higher-order belief about the uncertainty of *p* false, after all?

We think that such internal conflicts – where the agent feels, as it where, 'torn about whether she is certain' – can indeed arise, at least transitorily. But even during such conflicts, the agent is in an unambiguous *overall* state of *un*certainty. When the agent experiences a sense of certainty while thinking about *p*, and still believes that there are reasons against *p*, she is, as far as she can tell, *not* obliged to assent that *p*. She cannot then take the feeling she has to be a genuine feeling of rational compulsion to judge that *p*. The awareness of doubts must therefore override, as it were, the original awareness of rational obligation.

We can therefore conclude that conditional (d) is indeed true: a belief that one is not rationally obliged, as far as one can tell, to assent to a proposition *p* is sufficient for one not being rationally obliged, as far as one can tell, to assent to *p*. For any doubts one may have about being rationally obliged, as far as one can tell, are self-validating in the way we have described.

In addition, the issue of conflicts between different modes of access to a presumed obligation allows us to make a stronger point, which we will rely on at a later stage of our argument. Not only does doubt in the mode of *belief* override a feeling of certainty. There can also be cases in which a belief that *p* is certain co-occurs with a feeling of doubt regarding *p*. In such cases, the agent feels that something is wrong with *p*, and this means that as far as she can tell, it is not obligatory for her to assent to *p*. Hence, in such cases, too, the feeling of doubt will override the awareness of certainty, and the subject will be in an overall state of uncertainty. It follows that the overriding force of doubt does not depend on the mode in which the doubt occurs. Rather, any awareness of doubt necessarily 'trumps' a co-occurrent feeling or belief to the contrary: being in *some* way (belief, feeling) uncertain about *p* is sufficient for an overall state of *uncertainty* with regard to *p*. For an overall state of *certainty* requires a total *absence* of any form of awareness of doubts and persisting objections. (Notice that this has the consequence that in cases of such conflict, the belief that *p* is certain turns out to be wrong. We will therefore argue below in discussing conditional (b) that in rational thinkers conflicts of this kind, too, can only occur transitorily.)

Let us turn, next, to conditional (a) – the claim that whenever a proposition *p* is certain for an agent, the agent believes (or forms the belief, if prompted) that *p* is certain for him. Whenever I am, as far as I can tell, rationally obliged to assent to *p*, and I am confronted with the question *whether* I am thus obliged, there are, prima facie, two possible scenarios.[77] First, I can have (or adopt) the correct belief that I am, as far as I can tell, rationally obliged to assent to *p* – in

77 Provided that one possesses the requisite conceptual and reflective abilities to consider this question.

that case (a) is true, anyway. Second, I can fail to have the correct belief about my obligation, either because I have the *wrong* belief that I am *not*, as far as I can tell, rationally obliged to assent to *p*, or because I find myself undecided or uncertain about whether there is such an obligation for me. However, it seems that neither of these two situations can actually obtain. That the first scenario is not possible follows directly from our above argument for conditional (d): as we have tried to show, the negative belief that as far as I can tell, I am not rationally obliged to assent to *p*, is self-validating. As we want to show now, the second scenario is not possible, either, as higher-order uncertainty or ambivalence, too, undermines first-order certainty. Imagine, again, that you have rehearsed and examined a complicated piece of reasoning a couple of times, and as a result, you are unsure about whether you are rationally obliged, as far as you can tell, to endorse its conclusion. This situation, it seems, is just tantamount to a situation in which you fail to be, as far as you can tell, rationally obliged.

To see why this is so, consider the following. In thinking about a given proposition, it seems that one is *either* aware of a rational compulsion to assent, *or* one is positively aware of factors that make assent to *p* optional or even forbidden – and hence, of doubts about *p* (or about one's ability to assess *p*). Such factors can take different forms: they can consist in reasons that speak against *p*, or in a lack of strong reasons in favour of *p*, or in conditions which impede my clear judgment at the moment (e. g., being tired), or in feelings of uncertainty or hesitation regarding *p*. If I am not aware of any such factor of uncertainty at all in considering *p*, then *p* will seem evident to me – I will be aware of being rationally compelled to endorse *p*. In other words, awareness of rational compulsion and awareness of a lack of rational compulsion are exclusive alternatives in considering a given proposition: one cannot simply remain indifferent, and fail both to be aware of rational compulsion, *and* of a lack of rational compulsion. It is a consequence of this point that if an agent, upon considering *p*, fails to conclude that assent to *p is*, as far as he can tell, rationally obligatory, he *ipso facto* is in some way aware of doubts about *p* (or about his own ability to assess whether *p*). Hence, higher-order uncertainty undermines, and therefore cannot co-exist with, first-order certainty. Taken together with the earlier points, it follows that conditional (a) is true, too.

Before proceeding in our argument, we should notice a consequence of this last point that we will come back to at a later stage of the argument. We have argued that with regard to the question whether *p* is certain or not, one cannot remain indifferent – one either is aware of reasons that speak against *p* (or against one's judging that *p*), or one finds *p* compelling. From this, it follows not only that higher-order uncertainty undermines first-order certainty. In some sense, higher-order uncertainty about whether *p* is certain for me also undermines itself.

For it is plausible to assume that an agent who is capable of rational deliberation needs to be at least implicitly aware of the point that awareness of rational obligation and awareness of a doubt, or of a lack of rational obligation, are exclusive alternatives. If this is so, an agent who is undecided about whether *p* is certain for him is *ipso facto* not only in a state of uncertainty, but also in a state in which he is *able to tell* that *p* is uncertain for him: this simply follows from the fact that he does not find himself aware of a rational obligation to assent to *p*. Hence, he is able to abandon his higher-order uncertainty, and instead form a justified and correct belief that *p* is uncertain for him. In this sense, higher-order doubt is not a stable condition: the presence of such higher-order doubt *ipso facto* enables the agent to move to a negative belief about the absence of certainty. As we said, we will come back to this implication at a later point of our argument.

Let us take conditional (b) next – the claim that whenever an agent believes that she is, as far as she can tell, rationally obliged to assent to *p*, the agent is actually thus obliged, as far as she can tell. A first point to be made here is that, as stated earlier, the higher-order belief about certainty itself amounts to a way in which the agent has access to a (presumed) obligation. However, taken on its own, this point does not suffice to show that the positive higher-order belief is self-validating in the same way as the negative higher-order belief and higher-order doubt turned out to be. For we have also seen that different forms of access to a (lack of) rational obligation can be in conflict, and that in the case of such a conflict, the overall state of (un)certainty depends on whether *some* form of doubt is present or not. And the existence of a positive higher-order belief ('It is rationally obligatory for me, as far as I can tell, to assent to *p*') does not itself rule out the simultaneous presence of a form of doubt in another mode, e. g., by feeling a lack of confidence about *p*.

Yet while a conflict between a belief that *p* is certain and a simultaneous awareness of doubt that occurs in another mode (such as an epistemic feeling) is certainly possible, it is equally important to notice that this type of conflict cannot stably persist – at least not as long as the agent possesses the capacities that are required for rational deliberation. For not only does the conflict in question amount to a conflict between two inconsistent mental states (one of which affirms, while the other denies, the certainty of *p*). It also concerns mental states that stand in particularly tight relations to each other: for the accuracy of one state is affected by the presence of the other state. Take the case where a subject believes that *p* is certain for her, while at the same time she feels a sense of doubt regarding *p*. Not only does the content of the doubt contradict the content of the belief; as we have seen earlier, the very existence of the doubt makes it the case that *p* is uncertain for the agent, and hence, makes the belief '*p* is certain for me' wrong. A rational thinker who is capable of deliberation and of assessing the

certainty or uncertainty of propositions needs to grasp, at least implicitly, such relations, and to be able to resolve conflicts of this kind. Otherwise, his thought will suffer from a massive disintegration.

We can therefore conclude that in a subject capable of rational thought and deliberation, conflicts between a positive belief about certainty ('I am rationally obliged, as far as I can tell, to endorse *p*') and an awareness of uncertainty in another mode (e. g., a feeling of uncertainty) can occur only transitorily, and will be immediately resolved (either because the belief is abandoned, or because the conflicting awareness of uncertainty ceases to exist – e. g., if a feeling of doubt occurs only for a moment and does not return). As a consequence, *persisting* positive higher-order beliefs about certainty will be guaranteed to be correct, too: for they cannot, in a subject capable of rational thought, stably coexist with a conflicting awareness of uncertainty; rather, in the case of conflict, one of the two elements will immediately cease to exist. Thus, while a transitory dissociation between positive beliefs about certainty and actual certainty for the subject is possible, a *stable* belief of an agent that *p* is certain for her is indeed sufficient for *p* being actually certain for *the* agent. With this minor qualification, conditional (b) is correct, too.

We can finally turn now to conditional (c) – the claim that whenever *p* is uncertain for an agent, she also *believes* that *p* is uncertain for her. If *p* is uncertain for the agent, and the agent is asked whether *p* is certain for her or not, there are, again, prima facie three possibilities. First, the agent can already have, or form, the belief that assent to *p* is not rationally obligatory for her, as far as she can tell – in which case (c) is true. Second, she can wrongly believe she is rationally obligated to judge that *p*. But it directly follows from conditional (b), for which we have argued above, that there can be no such case, at least not stably: whenever the agent has the belief that *p* is certain for her, *p* is (transitory conflicts apart) certain for her. So the second possibility can be excluded as not a real one.

Third, the agent could be in doubt about whether assent to *p* is obligatory or not. But in the case of such doubts, a point from our earlier argument for conditional (a) applies. As we argued there, when considering the question whether assent to a given proposition *p* is, as far as they can tell, rationally obligatory, agents cannot remain indifferent; and they have to know that this is so. It follows that whenever they realize that they fail to believe that assent to *p* is, as far as they can tell, rationally obligatory for them, they can move to the belief that assent to *p* is *not* thus obligatory. Hence, the state of higher-order doubt about the (un)certainty of a proposition is not a stable condition – it *ipso facto* enables the agent to abandon the higher-order doubt, and to form a justified true belief that *p* is uncertain for her. Therefore, the case in which *p* is uncertain, but the agent finds herself undecided about whether *p* is uncertain or not, collapses into the first case, in

which the agent has the correct belief that *p* is uncertain. Thus, conditional (c) holds under this condition, too.

Taken together, the above arguments establish the four conditionals (a) to (d). They thus show that it is actually plausible to assume that (with the above qualification regarding stable positive beliefs) we can infallibly tell whether a proposition is certain for us – that is, whether we are, as far as we can tell, rationally obliged to assent to it or not.[78]

(3) So far, we have examined infallibility regarding the first two factors of a Kantian moral judgment$_{ATT}$: the obtaining of *some* relevant judgment$_{ATT}$, and the requisite form of certainty. What we still lack now is an understanding of how conscience assesses the third condition for a moral judgment$_{ATT}$ – the fact that a judgment$_{ATT}$ conforms with, and is based on, diligent examination of the case. (Call this condition 'diligence'.)

As a first step, we must note that diligence could either require that you have *de facto* taken into account all considerations that are actually relevant, and drawn the right conclusions from them. Or, alternatively, it could mean that, *for all you can tell*, you have done so. Since Kant clearly wants diligence to provide a

78 In particular, the reading that we have suggested excludes that we can be self-deceived about whether a given proposition is certain for us (in the proposed sense). This leaves open the possibility that the certainty itself might in some cases result from self-deception: it is conceivable that someone manipulates himself into being certain that *p* by gradually silencing all doubts about *p*, to the point where they completely vanish. Regarding Kant's views on this point, the discussion of the inquisitor at RGV 6:186 f. is directly relevant. The inquisitor is firmly convinced that he is obliged by divine will to sentence a heretic to death. Kant points out that this conviction is merely based on a particular interpretation of historical events (revelation), a kind of evidence which can always turn out to be deceptive. Therefore, he explains, the inquisitor cannot really be certain in his moral evaluation: "[...] we can always tell him outright [*auf den Kopf zusagen*] that in such a situation he could not have been entirely certain that he was not perhaps doing wrong" (RGV 6:186). As a consequence, Kant concludes that the inquisitor "risk[s] the danger of doing something which would be to the highest degree wrong, and on this score he acts unconscientiously" (RGV 6:187). Hence, Kant seems to exclude here a seemingly natural option: that the inquisitor has talked himself, or has been indoctrinated, into being so convinced of the rightfulness of the inquisition that he is simply not aware anymore of how shaky the grounds for his conviction actually are. Kant's position in this context can be given a charitable interpretation if Kant is read here as making a claim about the *psychological* limits of possible self-deception (or indoctrination). This claim would be that in situations of such enormous moral weight as that of the inquisitor's decision, it is simply not credible that an agent could lack even the slightest doubts about the justification for his intended action. Acknowledgement of this point is compatible with the idea that in less dramatic situations, people *are* capable of suppressing their doubts, and of thereby manipulating themselves into a state of certainty that they did not previously possess.

normative standard – something that one ought to attain, and for whose lack one can be fairly blamed – the second, subject-related reading is clearly preferable. For you can be unaware of relevant considerations through no fault of your own. We therefore adopt the following reading of diligence:

(α*) For all propositional attitudes x of a subject A: x is a moral judgment$_{ATT}$ → *For all that A can tell*, x conforms with, and is based on, a diligent examination of the case on A's part.

On the basis of (α*), we can make a crucial further step. Imagine an agent who holds a judgment$_{ATT}$ that p with certainty (in the sense of the above condition (2)). In that case, the agent is aware of being rationally compelled to judge that p. The important thing to notice is that this awareness of rational compulsion is incompatible with having any remaining *doubts* that some consideration upon closer examination would show that there is another rationally permitted option in the given situation (such as suspension of judgment). Any such doubts would undermine the awareness of rational compulsion. But the absence of doubts of this kind is just tantamount to diligence in the sense of (α*) – to an awareness of having diligently examined (as far as one can tell) the case. For if no doubts remain, any further examination of the case is unnecessary, as far as the agent can tell. We thus discover that *certainty entails diligence*: whenever condition (2) is fulfilled, condition (3) has to be present, too. Moreover, since we must understand that certainty entails diligence in order to be capable of conducting our moral deliberation properly, we will have to be at least implicitly aware of this connection.

This result enables us to complete the argument in favour of Kant's overall infallibility claim – on our reconstruction, the claim that conscience is infallible about whether the conjunction of conditions (1), (2), and (3) obtains or not. For remember that we have already accepted infallibility about whether each of conditions (1) and (2), taken in isolation, is present or not. Given that certainty entails diligence, it follows that whenever conditions (1) and (2) are fulfilled, diligence – condition (3) – is given, too. Therefore, conscience is infallible about the positive case, in which the conjunction of (1), (2), and (3) obtains.

What about the negative case, in which the conjunction of conditions (1), (2), and (3) does *not* obtain? From the previous parts of our reconstruction, it follows that conscience is infallible about cases in which (1) or (2) are lacking. Therefore, the only remaining case about which conscience might be in error or ignorance is that in which the conjunction does not obtain because condition (3), diligence, is not given, while conditions (1) and (2) *are* fulfilled. Yet it is a direct consequence of the above argument that this case cannot possibly occur: for certainty (condi-

tion (2)) entails diligence (condition (3)). It follows that conscience is infallible, too, about the case in which the conjunction of conditions (1), (2), and (3) does *not* obtain. This completes our account for Kant's infallibility claim regarding the verdicts of conscience about moral judgments.[79]

8 The Need to Restrict Kant's Claim

By way of conclusion, we wish to point to a limitation of Kant's view. As announced in Section 5, the preceding discussion has focused on infallible knowledge of one's *present* moral judgments$_{ATT}$. This is clearly the form of infallibility that is relevant for situations where the verdict of conscience takes place in advance of, or simultaneously with, the action. But what about verdicts of conscience that are issued *after* the action?

Imagine a case where *A* has ϕed at *t*, without being at that time certain that he was morally permitted to ϕ. His conscience awakens a year later. Meanwhile, he has acquired new background beliefs that make it rationally compulsory for him to judge *now* that ϕing at *t* was morally permitted. In such a case, it cannot be right that the subsequent change in *A* should lead to an assessment of his action as having been conscientious. So conscience has to verify whether *A* held a moral judgment$_{ATT}$ *at the time of the action*, that is, at *t*.

It follows that in order to issue an ex-post verdict, conscience has to assess the agent's *past* judgment that was responsible for the action. However, such an assessment has to rely on memory, which is obviously fallible. Thus, Kant's infallibility claim cannot be defended for the ex-post conscience, even by Kant's own lights. But we still take it to be a substantial result in its own right that Kant's account can be made plausible at least for ex-ante and simultaneous cases, given Kant's background assumption that we can infallibly know whether or not we

79 It should be noted that while this interpretation defends a substantial form of infallibility for the verdict of conscience, it is – unlike some versions of the standard approach that we have discussed earlier – fully compatible with a variety of forms of error and self-deception in the context of moral judgment and agency. In particular, infallibility about the conjunction of (1), (2) and (3) is compatible with both factual error and self-deception regarding (a) the commands of duty, both in general and as applied to concrete situations (cf. Kant's emphasis on the fallibility of moral judgment in passages (A), (B), and (C) from Section 2, and on our tendency to rationalize (*vernünfteln*) against our moral obligations, e. g. GMS 4:405); (b) features of our actions, including their motivation (see subsection (1) of Section 2 above); (c) the question whether our moral judgments$_{ATT}$ are really rationally obligatory for us; and (d) whether we have really diligently examined all available considerations.

judge that *p*. This last assumption is widely rejected today. Yet even if one does not accept it, there remains a crucial point from Kant's argument that is both original and, as we have tried to show, convincing: that you cannot err about whether you are certain, or whether you judge conscientiously. If Kant is right, our infallibility in this respect is tied up with our capacity for rational judgment itself.[80]

KrV	Immanuel Kant: *Kritik der reinen Vernunft*. Vol. 3–4 of *Kants Gesammelte Schriften*. Ed. Königlich-Preussische Akademie der Wissenschaften. 29 vols. Berlin 1900-. Translation: *Critique of Pure Reason*. Trans./ed. P. Guyer/ A. Wood. Cambridge 1998.)
KpV	Immanuel Kant: *Kritik der praktischen Vernunft*. Vol. 5 of *Kants Gesammelte Schriften*. Ed. Königlich-Preussische Akademie der Wissenschaften. 29 vols. Berlin 1900-. Translation in: *Immanuel Kant: Practical Philosophy*. Trans./ed. M. Gregor. Cambridge 1996, 193–272.
Log	Immanuel Kant: *Logik Jäsche*. Vol. 9 of *Kants Gesammelte Schriften*. Ed. Königlich-Preussische Akademie der Wissenschaften. 29 vols. Berlin 1900-, 1–150. Translation in: Immanuel Kant: *Lectures on Logic*. Trans./ed. J. M. Young. Cambridge 1992, 521–642.
Me	P. Menzer (ed.): *Eine Vorlesung Kants über Ethik*. 2nd ed. Berlin-Charlottenburg 1925 (=Moral Brauer [1774/5]).
MpVT	Immanuel Kant: "Über das Mißlingen aller philosophischen Versuche in der Theodicee." Vol. 8 of *Kants Gesammelte Schriften*. Ed. Königlich-Preussische Akademie der Wissenschaften. 29 vols. Berlin 1900-, 254–272. Translation in: Immanuel Kant: *Religion and Rational Theology*. Trans./ed. A. Wood/G. di Giovanni. Cambridge 1996, 19–38.
MS	Immanuel Kant: *Metaphysik der Sitten*. Vol. 6 of *Kants Gesammelte Schriften*. Ed. Königlich-Preussische Akademie der Wissenschaften. 29 vols. Berlin 1900–, 203–494. Translation in: *Immanuel Kant: Practical Philosophy*. Trans./ ed. M. Gregor. Cambridge 1996, 353–604.
RGV	Immanuel *Kant: Die Religion innerhalb der Grenzen der bloßen Vernunft*. Vol. 6 of *Kants Gesammelte Schriften*. Ed. Königlich-Preussische Akademie der Wissenschaften. 29 vols. Berlin 1900-, 1–202. Translation in: *Immanuel Kant: Religion and Rational Theology*. Trans./ed. A. Wood/G. di Giovanni. Cambridge 1996, 39–216.

80 We have presented earlier versions of parts of this article in Axel Hutter's colloquium at Ludwig Maximilians University, Munich, in 2010, in Tobias Rosefeldt's colloquium at Humboldt University, Berlin, in 2013, and at the Oxford conference 'Conscience and moral consciousness' in 2014. We thank the audiences at these occasions, as well as Thomas Höwing, Michael Oberst, and two anonymous reviewers of this journal, for precious feedback, and Axel Hutter for having drawn our attention to the topic. Special thanks go to Andrew Chignell for a very helpful set of written comments on an earlier version of this article.

V-PP/Powalski *Praktische Philosophie Powalski* [1782/3]. Vol. 27 of *Kants Gesammelte Schriften*. Ed. Königlich-Preussische Akademie der Wissenschaften. 29 vols. Berlin 1900–, 91–235.

V-Mo/Collins *Moralphilosophie Collins* [1774/5]. Vol. 27 of *Kants Gesammelte Schriften*. Ed. Königlich-Preussische Akademie der Wissenschaften. 29 vols. Berlin 1900–, 242–471. Translation in: *Immanuel Kant: Lectures on Ethics*. Ed. P. Heath/J. B. Schneewind. Trans. P. Heath. Cambridge 1997, 41–222.

V-MS/Vigil *Metaphysik der Sitten Vigilantius* [1793/4]. Vol. 27 of *Kants Gesammelte Schriften*. Ed. Königlich-Preussische Akademie der Wissenschaften. 29 vols. Berlin 1900-, 479–732. Translation in: *Immanuel Kant: Lectures on Ethics*. Ed. P. Heath/J. B. Schneewind. Trans. P. Heath. Cambridge 1997, 251–452.

Anonymous. 1808. *Verzeichniß der Bücher des verstorbenen Professor Johann Friedrich Gensichen, wozu auch die demselben zufallene Bücher des Professor Kant gehören […]*. Königsberg.

Aquinas, Th. 1970–76. *Quaestiones disputatae de veritate*. In Sancti Thomae de Aquino opera omnia iussu Leonis XIII P.M. edita. Vol. 22. Rome.

–. 188–1906. *Summa theologiae*. In Sancti Thomae de Aquino opera omnia iussu Leonis XIII P.M. edita. Vols. 4–12. Rome.

Baumgarten, A. G. 1760. *Initia philosophiae practicae primae*. 3rd ed. Halle. Repr. in vol. 19 of Immanuel Kant: Gesammelte Schriften. Edited by Königlich-Preussische Akademie der Wissenschaften. 29 vols. Berlin 1900–.

–. 1763. *Ethica philosophica*. 3rd ed. Magdeburg/Halle. Repr. in vol. 27.2.2 of Immanuel Kant: Gesammelte Schriften. Ed. Königlich-Preussische Akademie der Wissenschaften. 29 vols. Berlin 1900–.

Bilgrami, A. 2006. *Self-Knowledge and Resentment*. Cambridge, Mass./London.

Butler, J. 1726. *Fifteen Sermons preached at the Rolls Chapel*. London.

Chignell, A. 2007. "Kant's Concepts of Justification." *Nous* 41, 323–360.

Crusius, C. A. 1751. *Anweisung vernünftig zu leben* […]. 2nd ed. Leipzig.

Esser, A. 2013. "The Inner Court of Conscience, Moral Self-Knowledge, and the Proper Object of Duty (TL 6, 437–444)." In *Kant's "Tugendlehre". A Comprehensive Commentary*. Ed. A. Trampota et al. Berlin/Boston, 269–292.

Frank, W. (ed). 1997. *Duns Scotus on the Will and Morality*. Selected/trans. a. Wolter, O.F.M. 2nd ed. Washington, D.C.

Goering, J. 2004. "The Internal Forum and the Literature of Penance and Confession." *Traditio* 59, 175–227.

Grimm, J./Grimm, W. 1854–1961. *Deutsches Wörterbuch*. 33 vols. Leipzig.

Guyer, P. 2012. "Moral Feelings in the *Metaphysics of Morals*." In *Kant's Metaphysics of Morals. A Critical Guide*. Ed. L. Denis. Cambridge, 130–151.

Heidemann, D. 2012. "The 'I Think' Must Be Able to Accompany All my Representations. Unconscious Representations and Self-Consciousness in Kant." In *Kant's Philosophy of the Unconscious*. Ed. P. Giordanetti *et al.* Berlin/Boston, 37–59.

Hill, Jr., T. 2002a. "Four Conceptions of Conscience." In *Human Welfare and Moral Worth: Kantian Perspectives*. Oxford, 277–309.

–. 2002b. "Punishment, Conscience, and Moral Worth." In *Human Welfare and Moral Worth: Kantian Perspectives*. Oxford, 340–361.

Hoffmann, T. S. 2002. “Gewissen als praktische Apperzeption. Zur Lehre vom Gewissen in Kants Ethik-Vorlesungen.” *Kant-Studien* 93, 424–443.
Knappik, F./Mayr, E. 2013. “Gewissen und Gewissenhaftigkeit beim späten Kant.” In *Kant und die Philosophie in weltbürgerlicher Absicht. Akten des XI. Kant-Kongresses 2010*. Ed. S. Bacin et al., 5 vols., Berlin/Boston, vol. 3, 329–342.
Leech, J. 2012. “Kant’s Modalities of Judgment”. *European Journal of Philosophy* 20, 260–284.
Lehmann, G. 1980. “Zur Analyse des Gewissens in Kants Vorlesungen über Moralphilosophe.” In G. Lehmann, *Kants Tugenden. Neue Beiträge zur Geschichte und Interpretation der Philosophie Kants*. Berlin/New York, 27–58.
Mattey, G. J. 1986. “Kant’s Theory of Propositional Attitudes”. *Kant-Studien* 77, 423–440.
Moyar, D. 2006. “Unstable Autonomy: Conscience and Judgment in Kant’s Moral Philosophy.” *Journal of Moral Philosophy* 5, 327–360.
Paton, H. J. 1979. “Conscience and Kant.” *Kant-Studien* 70, 239–251.
Potts, T. 1980. *Conscience in Medieval Philosophy*. Cambridge.
Pufendorf, S. 1672. *De jure naturae et gentium libri octo*. Lund.
Schmucker, J. 1961. *Die Ursprünge der Ethik Kants in seinen vorkritischen Schriften und Reflektionen*. Meisenheim am Glain.
Schneewind, J. B. 1997. Introduction to *Immanuel Kant: Lectures on Ethics*. Ed. P. Heath/J. B. Schneewind. Trans. P. Heath. Cambridge 1997, xiii–xxvii.
–. 1998. *The Invention of Autonomy: A History of Modern Moral Philosophy*. Cambridge.
Timmermann, J. 2006. “Kant on Conscience, ‘Indirect’ Duty, and Moral Error.” *International Philosophical Quarterly* 46, 293–308.
Ware, O. 2009. “The Duty of Self-Knowledge.” *Philosophy and Phenomenological Research* 79, 671–698.
Wolff, C. 1968. *Vernünfftige Gedancken von der Menschen Thun und Lassen, zu Beförderung ihrer Glückseligkeit*. Ed. H. W. Arndt. *Gesammelte Werke*, vol. I.4. Hildesheim et al. Repr. of 4th edition, Frankfurt/Leipzig 1733.
Wood, A. 2008. *Kantian Ethics*. Cambridge.