Sticker, M. (2017). When the Reflective Watch-Dog Barks – Conscience and Self-Deception in Kant. *Journal of Value Inquiry*, 51(1), 85-104. <https://doi.org/10.1007/s10790-016-9559-4>

Peer reviewed version

License (if available):
CC BY

Link to published version (if available):
[10.1007/s10790-016-9559-4](https://doi.org/10.1007/s10790-016-9559-4)

[Link to publication record in Explore Bristol Research](#)
PDF-document

This is the accepted author manuscript (AAM). The final published version (version of record) is available online via Springer at <https://doi.org/10.1007/s10790-016-9559-4> . Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
<http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

When the Reflective Watch-Dog Barks

Conscience and Self-Deception in Kant¹

I discuss Kant's conception of conscience in the context of his conceptions of two phenomena central to Kant's practical philosophy: ordinary moral-cognition and self-deception. Conscience in Kant has the role of a *reflective watch-dog*: It watches over or reflects about how *cautiously* an agent uses her rational capacities when she reasons about concrete matters of duty. In my paper I discuss a problem that has not yet received its due attention. Conscience can warn agents every time they are about to transgress the moral law and it retrospectively reproaches agents for transgressions, yet Kant believes that there is also a natural propensity to self-deception. Self-deception, however, is only possible if agents can successfully hide from themselves the moral implications of some of their actions or present them as morally innocent. I argue that Kant's model of an internal court of conscience is inadequate to account for the pervasive threat self-deception poses. I propose a modified model instead, which I call "internal-panel-model".

Kant's conception of conscience has recently become a matter of intense interest and debate in the Kant literature.² I will focus on a problem that has not yet received its due attention. Conscience can warn agents every time they are about to transgress the moral law and it retrospectively reproaches agents for transgressions, yet Kant believes that there is also a natural propensity to self-deception. Self-deception, however, is only possible if agents can successfully hide from themselves the moral implications of some of their actions or present them as morally innocent. I will begin by outlining Kant's conception of ordinary cognition of duty and of self-deception or rationalizing (sec.1). I will then provide a brief overview of possible functions of conscience in Kant and explain Kant's conception of conscience as a *reflective watch-dog*: Conscience does not judge directly about our duty in concrete cases but it watches over or reflects about how *cautiously* an agent uses her rational capacities when she reasons about concrete matters of duty (sec.2). Finally, I will argue that Kant's model of an internal court of

¹ This paper is indebted to discussion with and feedback from Jens Timmerman, Claudia Blöser, Marcia Baron, Oliver Sensen, Paul Guyer, Nora Kassan, Adrian Piper, Joe Saunders, Alix Cohen, Sophie Møller and Samuel Kahn. I am grateful to the *UK Kant Society*, the *Institute for Advanced Study in the Humanities* in Essen, and to the *British Society for the History of Philosophy* for the opportunity to present my material, and to the philosophy departments of the universities of St Andrews, Stirling and Göttingen for supporting my research.

² For older pieces on conscience see Funke (1971), Paton (1979), Heubült (1980), and Hill's essays reprinted in Hill (2002, ch.9, 11). For recent literature see Velleman (1999), Hoffmann (2002), Timmermann (2006), Moya (2008), Wood (2008, ch.10), Ware (2009), Guyer (2010), Esser (2013), Vujošević (2014), Schmidt, Schönecker (2014a, 2014b) Sensen (2015), Kahn (forthcoming).

conscience is inadequate to account for the pervasive threat self-deception poses. I will propose a modified model instead, which I call “internal-panel-model” (sec.3).

My account of conscience differs from other accounts advanced in the Kant literature in that I believe that we can understand the specific role conscience plays in Kant only if we pay attention to how some of the functions commonly ascribed to conscience are in Kant’s ethics fulfilled by conceptions other than conscience and how conscience relates to these functions. Furthermore, I believe that we can understand the function of conscience better if we pay attention to cases in which something goes wrong in an agent’s critical self-assessment. I therefore analyse conscience in the context of two phenomena central to Kant’s practical philosophy: ordinary moral-cognition and self-deception. My paper thus stands in a recent tradition of shedding light on conscience via its relation to other elements of Kant’s practical philosophy, such as indirect duty (Timmermann 2006), the Fact of Reason (Moyar 2008), the duty to self-knowledge (Ware 2009), and the Aesthetic Preliminary Concepts (Guyer 2010).

Two notes before we begin. Firstly, I will assume that there is sufficient unity in Kant’s conception of conscience in the 1790s for his writings on conscience from this decade to shed light on each other.³ Furthermore, I will occasionally make reference to pre-critical (as well as critical) lecture notes. Reference to pre-critical lecture notes, however, always comes with the implicit caveat that Kant’s practical philosophy, at the time, might have been different from his critical conception and cannot be used as strong support for an interpretation of his mature conception of conscience.

Secondly, I will make use of the distinction between first-order and second-order judgement. By a “first-order judgement” I mean a judgement about what my duty is in a particular situation (for instance: “I ought not to make a promise to this particular person to pay back money she might lend me, if I do not have the intention of paying her back.”). Kant sometimes calls these judgements “objective”, since they have an object other than the agent, namely, actions (VI:401.5). By a “second-order judgement” or “reflective judgement” I mean a judgement about the reasoning that leads to first-order judgements (for instance: “I have reasoned about my options for how to extricate myself from financial predicaments with due caution.”). Kant sometimes calls these judgements

³ Moyar (2008, 328), by contrast, believes that Kant’s conception of conscience between 1791 and 1797 “shows a remarkable lack of uniformity”. The most notable tension in Kant’s conception of conscience in the 1790s is due to claims Kant makes about conscience as an Aesthetical Preliminary Concept in the Introduction to the *Doctrine of Virtue*.

“subjective” (VI:401.7). By this he does not mean that these judgements concern pleasure and pain or discretionary matters, but the *agent* or subject herself, as opposed to her actions.

1. Universalization and Self-Deception

In this section, I will lay out those elements of Kant’s philosophy without which we cannot properly understand conscience. These elements are the capacity of every rational agent to obtain moral knowledge via a simple universalization procedure (1.1), and a propensity to self-deception (1.2)

1.1 The Common Universalization Test

Kant explicitly states that “it is not conscience, which judges whether an action is in general right or wrong” (VI:186.2-3, see also VI:400.27-31). In fact, conscience presupposes the ability to tell right from wrong (XXVII:576.1-6, 617.10-5). First-order moral cognition is a matter of the understanding (ibid.614.31-615.5).⁴ Kant believes that every rational human being can find out what ought to be done in relatively easy ways. In the *Groundwork* and Second Critique, Kant describes moral cognition as an agent *asking* herself questions pertaining to the universalisability of her intended course of action.⁵ The moral significance of universalization is something all agents “always actually have before their eyes and use as the standard of judging” (IV:403.35-7). Even without philosophical training an agent can ask herself “would I actually be content that my maxim [...] should hold as a universal law” (IV:403.5-8, see also XXIX:628.32-4)? Common agents in their everyday moral evaluations can make use of pre-philosophical, *common universalisation tests* to check whether a maxim is morally permissible.⁶

According to Kant, the common universalisation test can be applied “in the very quickest and yet undeceptive way” (IV:403.4). It is within the power of every rational

⁴ There is consensus in the Kant literature that conscience is not primarily concerned with first-order matters. A notable exception is Guyer (2010, 145-6) who sees conscience as a capacity “to seek and listen to the kind of moral law that applies to particular cases, namely, particular maxims”, i.e., to apply general moral rules to particular cases (seeking) and being motivated to obey these rules (listening). Esser (2013, 277fn.16) convincingly argues that the idea that conscience is a form of power of judgment, which applies general rules to particular cases, rests on a misinterpretation of a passage in the *Religion* (VI:186.10-1).

⁵ “I ask myself” (IV:403.3-6), “I just ask myself” (ibid.21), “he still has enough conscience to ask himself” (IV:422.19), “[y]et he still asks himself” (IV:423.2-5).

⁶ See also IV:421.24-423.35, V:27.22, 36.4-6, 44.2-3, 69.20-70.9.

human agent and does not require a philosophical background, “wide-ranging acuteness” (IV:403.19) or experience “with regard to the course of the world” (IV:403.19-20). Kant sometimes overstates this when he says that an agent is “certain on the spot” (VIII:287.16-7), and “without hesitation” (V:36.28-9) what her duty is. We will see below that it would be in contradiction to what Kant says about the function of conscience as a watch-dog to interpret these statements as saying that finding out what ought to be done is never difficult.⁷

In the *Groundwork* and the Second Critique, Kant is optimistic that common universalization tests are sufficient to distinguish between right and wrong in concrete cases. Agents might only need an occasional reminder to make use of this test, or a reproach if they fail to do so. Accordingly, conscience is only mentioned three times in the two works (IV:404.22, 422.19, V:98.14). I will argue below that conscience observes the way agents apply common universalization tests to concrete cases. Conscience is therefore a matter of Kant’s more applied ethics, particularly of the *Metaphysics of Morals*. Once Kant focuses on concrete moral questions, and outright difficult ones such as the casuistical cases, he acknowledges that conscience has to watch over an agent’s reasoning, since some moral questions require cautious reasoning. That Kant discussed conscience in lectures that pre-date the *Religion* (see for instance XXVII:196-200, 296-8, 351-7) indicates that even pre-1790s he is aware of the importance of conscience. He brackets conscience in his published works in the 1780s, because it plays no role for the *Groundwork’s* and Second Critique’s foundational enterprise.⁸

1.2 Rationalizing

Given the alleged quickness and undeceptiveness of the common universalization test (see IV:403.4) we may wonder: How is it possible that agents get their judgement about duty wrong? Kant addresses this issue in his discussion of rationalizing [Vernünfteln].⁹ Since Kant’s thoughts on rationalizing are scattered throughout his works, interpreters often fail to realize the full extent to which Kant discusses the different forms and

⁷ The previous two paragraphs are a condensed version of Sticker (2015b). There I also discuss the moral-epistemological function of humanity or rational nature, which I have to bracket here.

⁸ This is the explanation advanced by Timmermann (2006).

⁹ I will follow Timmermann (2011) and use “rationalizing” as a translation of “Vernünfteln”.

aspects of this form of self-deception.¹⁰

Kant's most prominent discussion of rationalizing is the penultimate paragraph of *Groundwork* I:

Innocence is a glorious [1] thing, but then again it is very sad that it is so hard to preserve and so **easily seduced [2]**. Because of this even wisdom – which otherwise probably consists more in behaviour than in knowledge – yet needs science too, not in order to learn from it, but to obtain access and durability for its prescription. The human being feels within himself **a powerful counterweight to all the commands of duty [3]** – which reason represents to him as so worthy of the highest respect – in his needs and inclinations, the entire satisfaction of which he sums up under the name of happiness. Now reason issues its prescriptions unrelentingly, yet without promising anything to the inclinations, and hence, as it were, with reproach and disrespect for those claims, which are so vehement and yet seem so reasonable (and will not be eliminated by any command). But from this there arises a *natural dialectic* [4], **i.e. a propensity to rationalize against those strict laws of duty [5]**, and **to cast doubt [6]** on their validity, or at least their **purity [7]** and **strictness [8]**, and, where possible, to make them better suited to our wishes and inclinations, **i.e. fundamentally to corrupt them and deprive them of their entire dignity [9]**, something that in the end even common practical reason cannot endorse. (IV:404.37-405.19)

The common agent without philosophical training is in a state of “innocence” [1]. She has a correct understanding of what matters morally, namely, universality, and she can make use of this notion to discover her duties in concrete cases. Innocence, however, is difficult to preserve and the common agent is “easily seduced” [2]. The tension between unconditional commands of duty and happiness, which constitutes “a powerful counterweight to all the commands of duty” [3], gives rise to a “*natural dialectic*” [4]. This dialectic is *natural* in the sense that our rational and our sensuous natures are in tension with each other, and that this tension is unavoidable for finite rational agents. Kant characterizes the natural dialectical in the *Groundwork* as identical to¹¹ “a propensity to rationalize against those strict laws of duty” [5]. Rationalizing is characterized as the attempt to cast doubt on central elements of morality or to modify one's way of reasoning about duty. Such a modification, Kant fears, would fundamentally

¹⁰ For exemplary cases of rationalizing see for instance VI:70.fn., 77.26-78.2, 168.8-170.11, 192.1-202.5, 430.19-26, VIII:265.28-266.5, LK:154-7, 192-3. There is a growing interest in rationalizing in the Kant literature. See Schönecker (1997), Guyer (2000), Piper (2008), Shell (2009), Sticker (2015a, sec.3), van Ackeren, Sticker (2015, sec.3).

¹¹ Kant uses the strong “i.e.” [d.i.] [5] “this is” to characterize the relation between the natural dialectic and the propensity to rationalize.

“corrupt” the strict laws of duty, “and deprive them of their entire dignity” [9].¹²

Corruption is a twofold problem. Firstly, a corrupted agent *reasons incorrectly* about morality, namely, according to other than the strict and pure standard of universality. Secondly, a corrupted agent’s actions do not express *unconditional commitment to duty*, since the agent might not have acted, had certain grounds of excuses obtained. Rationalizing represents an epistemic as well as a motivational problem. The two main strategies of rationalizing Kant mentions are “casting doubt” [6] on the purity [7] and strictness [8] of duty. *Purity* firstly demands that nothing empirical, i.e., inclinations, an agent’s non-obligatory goals, etc. function as *criteria* for what is moral (the cognitive component). Secondly, purity demands that an agent’s obligatory actions or omissions are not *motivated* by pathological desires, but by respect for the moral law (the motivational component). Agents who reason without due regard for purity believe that other factors than universality, such as consequences or external authorities, can inform their reasoning about duty and provide moral reasons for actions. *Strictness* is the claim that moral commands are always of supreme authority and, at least in the case of perfect duties, never admit of exceptions. Agents who reason without due regard for strictness believe that the normative force of moral commands can, in exceptional circumstances, be outweighed by other considerations.

Kant is adamant that disobedience to the moral law can only mean to “permit ourselves (with all respect for it) just a few exceptions” (IV:424.36-7). Rationalizing only makes sense for agents who are in principle committed to morality and who want to retain their self-esteem as moral agents.¹³ The rationaliser’s commitment to duty, however, is not unconditional. She is committed either unconditionally to an impure conception of morality that combines elements of morality and inclinations (no purity), or committed conditionally to a pure conception of morality, i.e., she believes that there can be exceptions and excuses (no strictness). In the former case the impure elements are part of the agent’s conception of morality itself. In the latter case agents still act from a pure principle when they act from what they take to be the moral option, but their obedience is conditioned on other factors than duty. An agent who has abandoned purity perceives her inclinations and personal goals as having a say in the question of what is permissible or even obligatory. The agent might believe that she is fully morally justified

¹² Again Kant uses the strong “i.e.” [9]. Adapting one’s conception of morality and corruption is identical.

¹³ See also VI:36.1-33, 321.38-9, XXIX:609.34-6, 629.2-5.

in what she is doing, even when she is not. An agent who has abandoned strictness, by contrast, is aware that there are strong moral considerations that speak in favour of disregarding personal happiness for the sake of the moral option, but she believes that there are also grounds of excuses for not heeding moral commands.¹⁴

The distinction between taking oneself to be merely excused and believing to be justified is not always clear-cut. The goal of rationalizing is to boost one's perceived moral standing, while at the same time not having to forego satisfaction of inclinations. Justifications are more potent for this purpose than excuses, since justifications suggest that the agent does not need to morally improve at all. When the agent still has a firm grasp of what is right and wrong, however, the only thing she can do to rationalize is to invent an excuse for an action that she acknowledges as wrong.

In what follows, I will show what role conscience plays for Kant given that for him it does not provide first-order moral insight, and given that Kant stresses the pervasiveness of a human being's propensity to self-deception.

2. Conscience as a Reflective Watch-Dog

In this section, I argue that we should understand conscience as a reflective watch-dog concerned with second-order reflection on the cautiousness of our first-order processes of cognizing our duty (2.1). I will then discuss the alleged infallibility of conscience in contrast to the fallibility of our first-order judgements about duty (2.2). This will lead us to the problem of how there can be self-deception before the internal court.

2.1 The Reflective Function of Conscience

Kant scholars have ascribed to conscience in Kant at least three different functions or combinations thereof: (i) Conscience makes my duty *my* duty parallel to the I think of the First Critique, which makes my thoughts *my* thoughts¹⁵; (ii) conscience has a

¹⁴ Kant claims that the agent who makes an exception for himself, as opposed to the agent who adopts an immoral rule as his maxim, still "can at the same time detest his transgression" (VI:321.38, see also *ibid.*379-380.fn.). I loosely follow Austin's (1961, 176) distinction between justifications and excuses. In the former case someone accepts responsibility but denies that an action was bad, in the latter case someone accepts that an action was bad but denies full, or even any, responsibility.

¹⁵ See XVIII:579.4-5, XXVII:613.37-614.9, Moyer (2008, 350), Vujošević (2014).

motivational function¹⁶; (iii) conscience is *reflective*: it warns an agent before an action that she cannot be certain that what she intends to do is right or it makes itself felt after an action in the form of pangs of conscience.¹⁷

In what follows, I am concerned with the reflective dimension of conscience (iii). Before I begin, let me say a word about the other two functions. Function (i) is, I believe, already fulfilled by the agent's awareness of being under the moral law; an awareness that is discussed extensively and without any appeal to conscience in Kant's Second Critique's discussion of the Fact of Reason (see esp.V:30.32-5). It would be strange if an agent who is subject to the moral law needed an additional capacity on top of her awareness of duty to understand that her duty is not anyone else's but *her* duty. Understanding my duty as obligating *me* is exactly what awareness of being under the moral law means. One might argue that conscience simply *is* my awareness that I am under the moral law. In this case, however, conscience would not serve a function distinctive of the Fact of Reason.¹⁸ However, if conscience presupposes first-order judgments, as we will see it does, it cannot be identical to our awareness of being under the moral law. We could not make first-order judgments without this awareness already in place.

Furthermore, reason when critically reflecting on its own workings can affect an agent's mind or her sensuous side (ii). Its verdicts can be the ground of "repentance" (V:98.28) and of "*pain* one feels from the pangs of conscience" (VI:394.3-5, see also V:98.29-30). According to my reading, the warnings of conscience can lead an agent to refrain from acting when she realizes that she is not in a position to be certain that what

¹⁶ See VI:400.30 and Guyer's (2010, 145-6) reading of conscience as part of an empirically informed theory of moral motivation. Moyer (2008, sec.6), like Guyer, stresses the authority and motivational role of conscience. He, however, adds the caveat that this stress is a Fichte inspired reading of Kant. Vujošević (2014, 450) argues that "conscience itself is not a moral motive", but it "has an inescapable role to play in the process of our motivation to act morally".

¹⁷ See for instance Hill (2002, 280-1). Moyer (2008, 338) believes that conscience is a reflective capacity "directed at my process of judgment and my grounds for thinking that I have done the best that I possibly can in arriving at my judgment". Moyer (ibid.345) also argues that Kant "invests conscience with so much authority that the first-order/second-order distinction becomes completely untenable", or that the question: "Is this the right maxim?" is indistinguishable from: "Do I believe that this is the right action?" (see ibid.346). I do not share this criticism: The two orders only collapse if conscience judges about the certainty of *beliefs*. Conscience, however, judges about an agent's way of reasoning (see below). The question conscience asks is rather: "Did I reason about a situation with due caution?". This is different from the question what maxim or action is right. First- and second-order judgements are about different objects (actions/maxims vs. reasoning) and they employ different methods (universalization tests vs. critical self-examination). See also Grenberg (2013, 180fn.10) who argues convincingly that Kant does not collapse the two questions.

¹⁸ Moyer (2008) has sympathy for this reading. Textual support for such a much more fundamental role of conscience than what the other passages on conscience support is in VI:399.1-403.6.

she is about to do is morally permissible. This is different from being motivated by respect for the moral law. In the case of respect agents act or refrain from acting because they are positively certain that an action is obligatory or forbidden, whereas in the case of conscience agents refrain from acting because they realize that they are not certain that their action would be morally right. Whether the latter is morally worthy or merely a prudential means to avoid pangs of conscience is a question I cannot answer here. The motivational function of conscience presupposes a verdict of conscience or the outcome of the reflective function, which I will discuss for the rest of this paper.

Kant characterizes conscience as “practical reason holding the human being’s duty before him for his acquittal or condemnation in every case that comes under a law” (VI:400.27-8). Conscience is concerned with concrete cases and with the moral status of the agent herself (her “acquittal or condemnation”), and it comes into effect in *every* morally relevant case (“every case that comes under a law”). Furthermore, conscience is *reflective*: a “moral faculty of judgement, passing judgement upon itself” (VI:186.10-11).¹⁹ It is important that Kant does not claim that conscience passes judgement on the *agent*, but on an agent’s moral faculty of judgement (“itself”) of which conscience is a part (see VI:400.27-8). Conscience critically scrutinizes first-order judgments and the reasoning that leads to these judgments.

Kant explicitly acknowledges that an agent can err in her “objective judgement as to whether something is a duty or not” (VI:401.5-6). He contrasts the fallibility of first-order moral judgments with the infallibility of conscience (VIII:268.13-8, XXVII:615.32-6 – see my sec.2.2).²⁰ Conscience judges whether the agent is or was *warranted to believe* that something is or is not her duty. Conscience is reason when reason “judges itself”,

¹⁹ Timmermann (2006, 295) notes that this reflective function of conscience is coherent with the standard Protestant conception of conscience at the time as something more narrow than an agent’s sense of right and wrong. He (ibid.295) argues that conscience is “the power [...] – within *every* human agent – that acknowledges the need to conform to moral standards”. The moral standards themselves are provided by reason, not conscience.

²⁰ Timmermann (2006, 306) believes that the “possibility of objective errors [about my duty in concrete cases] [...] seems to undermine the very foundation of Kantian ethics”, namely, the idea that every rational human agent has insight into what she is morally required to do. On Timmermann’s conception agents can fail to act on what they are aware of as obligatory, but they cannot have false beliefs about what their duty is in a concrete situation. My interpretation of conscience, in contrast to Timmermann’s, accommodates mistakes in objective judgements without undermining the foundation of Kant’s ethics. Rational agents always have what it takes to reach the correct judgment in moral matters, namely, the ability to make use of the standard of universality (see sec.1.1). If they reason incorrectly they can be held morally accountable for it, since it would have been possible and sufficiently easy for them to judge correctly.

according to the question of “whether it has actually undertaken, with all caution [Behutsamkeit]” the “examination of actions” (VI:186.17-8).²¹ Caution is a property of a *process*.²² We can reason cautiously, or with lack of caution. Conscience warns agents not to perform a concrete action when, due to lack of a cautious application of common universalization tests, an agent cannot be certain of the moral permissibility of her action.²³

The best illustration for the function of conscience in Kant is the *Religion's* inquisitor who condemns a heretic to death. The inquisitor is “firm in the belief” (VI:186.30) that he is permitted, maybe even morally required to put to death a heretic. The inquisitor does what he believes is right, but, according to Kant, he acts against his conscience, since he “could not have been entirely certain” (ibid.29) that putting to death the heretic is right.²⁴ There is no sound deliberative route that leads to the belief that the heretic ought to die, since, firstly, it is absolutely certain that taking a human life because of a person's faith is wrong (VI:186.36-7, see also IX:69.13-70.8), and, secondly, everything that tells otherwise is “a matter of historical documentation and never apodictically certain” (VI:187.3-4, see also XXVII:614.11-30). Only an insufficient use of rational capacities, which ignores that which is absolutely certain and which is based on questionable evidence, lets the inquisitor reach his firmly held but mistaken conviction. Conscience is supposed to warn agents of unwarranted moral beliefs.²⁵

²¹ See also VIII:268.7-8. Wood, Di Giovanni (1996, 203) translate “Behutsamkeit” as “diligence”. “Caution” better preserves the original meaning of the word: to be aware of or to protect oneself of danger (see *Adelung* vol.1, col.817-8). I am grateful to Jens Timmermann for suggesting the term “caution”.

²² The same is true of the term “conscientious” and the German “gewissenhaft”. Both aptly describe the working of conscience as concerned with the question as to whether agents reached their moral convictions in a reliable way. Sensen (2015, 127-8) argues plausibly that being conscientious for Kant means that a moral judgment was “cautiously examined” [sorgfältig geprüft] by the agent (see also ibid.130-1).

²³ In the *Vigilantius* lecture notes from 1793/4 conscientiousness appears as a matter of a “sincere examination” [aufrichtige Prüfung] of what we take to be certain (XXVII:615.19). It is the role of conscience “to obtain consciousness that one has performed the examination with great thought” (XXVII:619.27-8). See also Vujošević (2014, 465) who believes that conscience assesses “our way of judging whether an action is right or wrong”. Vujošević, however, is not always clear about the object of assessment. She claims that conscience judges our “maxim-formation and adoption” (ibid.) or incentives (ibid.470). The latter seems very different from our way of reasoning.

²⁴ In *Vigilantius* the inquisitor is said to “take himself to be justified” in his deeds (XXVII:615.29).

²⁵ Esser (2013, 283-4) thinks that conscience is not supposed to scrutinize the use of rational capacities, but the *motives* of actions (see also Vujošević 2014, 461). This interpretation is contradicted by the inquisitor case. The inquisitor got wrong the question what his *duty* is. He does not merely lack the right motivation. Schmidt, Schönecker (2014a, 283) make a very convincing case against the idea that the scrutiny of conscience extends to motivation: If conscience pronounced a verdict about the motives of actions, it should be possible to know whether our actions are morally worthy, something Kant strenuously denies (A/B:278/334, 551/579, VI: 25.5-6, 38.7-12, 51.7-21, 70.1-71.20, 75.8-76.1, VI.451.21-36).

Conscience is thus a reflective Watch-Dog: It critically “observes” (VI:438.14) that agents think about the universalizability of a maxim with due caution, and when it detects mistakes it barks in order to warn the agent (before the action) or to reproach her if she ignored the warning (after the action). Furthermore, agents can presumably reason with so little caution or be so convinced of their views that they do not have the time or do not see the need for critical reflection before an action. In this case conscience will condemn the action once agents critically reflect about the action in retrospection. Conscience can never precede moral reasoning. It presupposes moral reasoning as the object of its reflection. Conscience can, however, precede *action*. Agents can notice insufficiencies in their reasoning before they act, during their action, as well as in retrospection.²⁶

Kant’s idea that conscience is concerned with second-order or subjective judgements is already anticipated in his legal philosophy. We will discuss below in detail the juridical metaphor Kant uses to explain the proceedings of conscience: the court of law. In the *Doctrine of Right*, Kant claims that it is a “common fault (*vitium subreptionis*) of experts on right” to mix up the rightful principle a court is “bound to adopt for its own use (hence for a subjective purpose)” with an “objective” principle of what is right in itself. Both principles are “very different” (VI:297.21-9). Kant here already distinguishes between a subjective principle that structures the court proceedings for the purpose of judging agents, and an objective principle of right. Whilst this characterization leaves open that juridical courts are deficient in their subjective judgements, since they do not base their verdicts on what is right in itself, we will see below that in the case of an internal court there is no such deficiency. The internal court is infallible in its subjective judgements. Its judgements are about a completely different (a reflexive) matter than objective judgments of moral agents.

Owen Ware (2009) in a paper on conscience and the duty to self-knowledge, comes closest to the conception of conscience I advocated so far. He believes that while conscience is also concerned with the question of “whether the agent’s actions really do line up with the judgements of practical understanding” (ibid.693) or whether we did what we believed to be our duty, another function of conscience is to judge “the care the agent applies (or fails to apply) in the act of examining what action she ought or ought

²⁶ See explicitly “when it comes, *or has come*, to a deed, conscience speaks involuntary and unavoidably” (VI:401.15-6). See also XXVII:198.17-24.

not take” (ibid.). The main difference between our respective accounts is that Ware seems to believe that the infallibility of conscience makes self-deception impossible or at least that self-deception can only be successful until conscience becomes active. Furthermore, he only discusses self-deception as a form of opacity about our motives. I, by contrast, believe that Kant’s conception of self-deception is much richer and more substantial, for instance pertaining to general structures of one’s conception of morality, such as strictness and purity (see my sec.1.2). Kant’s optimism about the accuracy and force of conscience must be reconciled with his pessimism about how dishonest agents tend to be with themselves.

2.2 *The Infallibility of Conscience*

After we have established that conscience critically watches over an agent’s moral reasoning, we now turn to Kant’s claim that in her reflective judgement about whether the agent reasoned cautiously about her duty an agent “cannot err” (ibid.401.8, see also VIII:268.10-3, XXVII:615.32-6).²⁷

The claim that conscience is infallible – and even stronger that an erring conscience would be an “absurdity” (VI:401.5, VIII:268.11) – is an enduring source of puzzlement for Kant scholars. The claim becomes much easier to stomach if we bear in mind that the infallibility of conscience does not imply that our first-order moral judgments about duty are infallible.²⁸ Kant explains the infallibility of conscience as follows: “I cannot be mistaken in my subjective judgement as to whether I have submitted it to my practical reason (here in its role as judgement) for such a judgement” (VI:401.6-8). The “it” here is most naturally understood to refer back to the objective or first-order judgment, which Kant discusses immediately before the passage quoted. Infallibility of conscience then does not actually pertain to the verdict of conscience itself, but to the judgment *that* I did critically scrutinize my first-order judgment or *that* I submitted it to the process of the internal court.²⁹

²⁷ The standard conception of conscience at the time was that conscience *can* err. See Baumgarten reprinted in Kant’s *Academy Edition* (XXVII:781) and Crusius (1767, §138-41). In pre-critical lectures Kant himself thought of conscience as fallible (XXVII:42.32-7, 197.36-198.9, 354.39-355.27).

²⁸ Paton (1979) is very skeptical of Kant’s infallibility claim and argues, correctly I believe, that we can only make sense of it, if we understand conscience as being about second-order matters.

²⁹ Hill (2002, 302-3.fn.50) suggests that conscience judges always correctly, but we might fail to submit an action to the scrutiny of conscience. The claim that conscience cannot err is therefore somewhat “exaggerated” (Hill 2002, 241, see also ibid.348). I agree that the claim is exaggerated, but for almost the

Kant's almost pompous claim that an erring conscience is an absurdity strongly contrasts with the rather deflationary explanation he provides for this statement in the *Metaphysics of Morals*. In the *Theodicy* essay he explains conscience is infallible because I cannot err in my belief that I am convinced that something is right, if I am in fact convinced that it is right (VIII:268.15-8). According to this notion, infallibility almost sounds like a tautology. Kant probably thought that on top of infallibility about whether I did submit my judgment to critical scrutiny and that I cannot err in my belief that I am convinced that something is right, the process of conscience itself is set up in such a way that without interference by rationalizing conscience would not err. It is only due to the agent being corrupted and buying into a transformed conception of morality that the verdict of conscience errs. Conscience viewed in isolation of corrupting influences might be infallible indeed – after all, why would it fail?

According to my conception, conscience is in an ideal position to get it right every time, but it is not actually infallible. I admit that this is an attempt to explain away many of the problematic aspects of Kant's claim that conscience is infallible. This, however, is the only appropriate reaction to this claim if we keep in mind that Kant also stresses the propensity to self-deception as an element of the human condition. We must understand the infallibility claim such that it is still intelligible why agents are constantly tempted to engage in self-deception, and this would not be the case if self-deception, due to the infallibility of conscience, was impossible.³⁰

Kant's discussion of conscience aims to capture an agent's way of critically reflecting on her way of approaching moral questions. It works on a different framework than first-order reasoning about concrete moral matters, and it hence offers a partly external perspective on this reasoning; a perspective which can have diagnostic functions. This is a deflationary understanding of conscience, but I believe that it captures Kant's intentions in so far as they can be defended and insofar as they fit with the rest of his philosophy. We must now turn directly to the question of how critical self-evaluation sits with the pervasiveness of self-deception.

opposite reason as Hill: The verdict of conscience can err, but an agent cannot be mistaken as to whether she submitted her moral judgement to the scrutiny of conscience

³⁰ Other deflationary readings who rather aim to explain the claim away are Wood (2008, 190-1), Ware (2009, 692-5). Wood explicitly acknowledges the difficulty of understanding the infallibility claim in the context of self-deception.

3. Internal Courts and Internal Panels

Agents may “artificialize” [künsteln] as much as they want to represent moral transgressions as “unintentional fault” or mere “oversight”; they cannot “reduce to silence the prosecutor within” (V:98.13-21) or “bribe reason” (VI:77.27). If conscience cannot be silenced or be partial (bribed) and if it becomes active in *every* morally relevant case (VI:400.27-8), then how could agents ever escape its reprimands other than through strict obedience to the law? If they could not, it would be inexplicable how agents can be in constant danger of engaging in rationalizing. They might attempt it a few times and realize that it leads them nowhere and then either reason about their duty according to the strictness and purity of the moral law and act on the outcome of this reasoning or suffer pangs of conscience. This, however, would be in tension with Kant’s notion that rationalizing is the actualization of a propensity rooted in human nature, and thus something that does not simply go away after a while (see my sec.1.2 [4-5]). In addition, Kant believes that the transition from common rational cognition of duty to practical philosophy is necessitated “on practical grounds” by the danger of rationalizing (IV:405.20-35). If this danger were only transitory, then Kant’s critical practical philosophy would not serve an important practical function. The philosopher could simply wait until agents discover that all rationalizing is in vain.

There must be something agents can expect to get out of their rationalizing. Indeed, Kant admits that there can be “opium” administered to conscience (VI:78.32-5). Conscience can be “stunned” or “put to sleep” temporarily, though even the most thorough self-deceiver cannot avoid “waking up from time to time” and hear the “fearful voice” of conscience (VI:438.13-23).³¹ These metaphors are best understood as referring to *excuses* for actions, which the agent acknowledges as wrong. Once conscience has pronounced an agent guilty, the agent can wonder about whether she is fully responsible for her action, i.e., whether there are excuses or mitigating circumstances. Kant says that whilst no finite agent endowed with practical reason is without conscience (VI:400.23-25), agents may pay no heed to its verdicts, or try to ignore them.³² The possibility of finding and even of making up excuses does in no way infringe upon the impartiality of conscience, since excuses only enter *after* the verdict of conscience and help agents

³¹ Wanting to serve and praise God without moral sincerity is an “opiate for conscience” and “a cushion” on which conscience is meant to “sleep quietly” (IX:495.13-5). In pre-critical lecture notes, Kant speaks of “numbing” [übertäuben] and “giving leave” [Ferien geben] to our internal judge (LK:135, see also XXVII:357.4-5).

³² See VI:400.31-33, 401.10-11, 438.13-23.

ignore this verdict.

Agents, however, often look for more than excuses. They also look for *justifications*, in fact, as pointed out above (sec.1.2), agents normally wonder whether what they are doing is right or justified, not merely whether they are excused. If it is possible to falsely deem oneself justified, not merely excused, it must be possible that the verdict of conscience itself can be interfered with. Indeed, Kant clearly states that there can be undeserved “peace of conscience” [Gewissensruhe], i.e., conscience itself can be silent when it should warn or condemn (VI:38.12-23). An example that we already came across is the inquisitor (see my sec.2.1). The inquisitor’s reasoning about duty is not guided by the notion of universality. He draws on impure (morally irrelevant) considerations, such as “historical documentation” (VI:187.3). He thinks that scripture and tradition can inform him about duty and he is convinced that he is justified in his doings.³³

We need to gain a better understanding of how Kant thinks conscience proceeds in the critical assessment of an agent’s reasoning. Kant elaborates on his conception of self-assessment via his metaphor of an “*internal court*” (VI:438.11).³⁴ Kant’s metaphorical use of a court to make a philosophical point is not limited to conscience. For instance, Kant prominently conceives of his First Critique as “the true court of justice for all controversies of pure reason” (A/B:751/779).³⁵ In his Introduction to the *Metaphysics of Morals*, Kant calls a judge or court of law a “(natural or moral) person that is authorized to impute with rightful force” (VI:227.27-9). Kant here does not distinguish between the judge and the court in its entirety (see also VI:317.24-5). What matters here for his treatment of conscience is that a court or the judge is the rightful and authorized source of imputation. There would be no legal imputation (and hence no legitimate punishment) without a court of law. This translates to the moral sphere where conscience is presented as the source of imputing actions to oneself and as the capacity to critically assess these actions. Furthermore, the idea that conscience is a *forum internum* is prominent in Alexander Gottlieb Baumgarten’s *Initia philosophia practicae primae*³⁶, which had great influence on Kant and was used by Kant as a text-book for his

³³ See also VI:38.12-7, 174.27-30, 333.33-4, VIII:268.26-269.1 for examples of agents who falsely deem themselves morally justified. In the *Vigilantius* lecture notes, Kant calls the attempt to directly deceive conscience “sophistry” and gives a number of short examples (XXVII:619.36-620.8).

³⁴ See also V:98.13-28, XXVII:197.19-36, 295.12-36, 351.22-357.5, 572.13-573.24.

³⁵ See Møller (2013) for a recent discussion of judicial metaphors in the First Critique.

³⁶ See XIX:83.11-35. Baumgarten’s *Initia* is reprinted in volume XIX of the *Academy Edition*.

lectures.³⁷

The central idea of the court metaphor for understanding conscience is that the accused and the judge cannot be the same person, otherwise the prosecutor would always lose (VI:438.29-33). To allow for an impartial evaluation of an agent, Kant introduces two distinctions. Firstly, he distinguishes between an agent's rational self, the *homo noumenon*, who is the "prosecutor" (ibid.439fn.34) of the internal court, and the "human being as a sensible being endowed with reason" (ibid.30-1). The latter is the accused. Interestingly, Kant here does not mention the *homo phaenomenon*, the usual counterpart of the *homo noumenon*. This suggests that the accused is not merely the agent as an instrumentally rational animal (*homo phaenomenon*), but the agent as a finite *moral* being or the agent as a whole. This makes sense, since it is the agent endowed with empirical practical reason and under the moral law who can transgress moral laws and hold herself responsible for it.

Kant distinguishes between "the human being as a moral being (*homo noumenon*)" and "as a natural being (*homo phaenomenon*)" (VI:430.14-5).³⁸ Kant emphasizes that the distinction is one between different ways of regarding "the same human being" (VI:418.17, see also ibid.239.27-8), and that this distinction is only made "in a practical respect" (ibid.439.31).³⁹ Kant draws on the distinction between the *hominis* to explain how we can have duties to ourselves (ibid.418.5-23, XXVII:539.1-16, 579.8-23), and to remove a supposed contradiction between self-determination and natural causality (XXVII:505.31-506.3). Whilst Kant is explicit that the *homo phaenomenon* is an animal or our animal self (VI:418.7), this animal is still rational, albeit only instrumentally rational (ibid.5-13). The *homo phaenomenon* is not a rational *moral* agent, for him "the concept of obligation does not come into consideration" (ibid.16-7). He can only be moved by hypothetical imperative, not by categorical ones which require a source of motivation independently of inclinations.⁴⁰ The *homo phaenomenon* therefore does not have what it

³⁷ See Kuehn (2010, sec.3) for more on Baumgarten's influence on Kant.

³⁸ Kant elaborates on his distinction between the *hominis* elsewhere in the *Metaphysics of Morals* and in his lectures of the 1790s. See VI:239.21-30, 335.17-22, 418.18, 423.5, 434.22-435.5, 439.30-1, XXIII:398.11-2, XXVII:579.29-30, 593.6-22. Byrd, Hruschka (2010, ch.14) point to Linné's *Systema Naturae* as an important scientific influence on Kant for this distinction, as well as to the traditional Roman law distinction between a thing and a person.

³⁹ See VI:239.26, 418.5, 434.32, XXVII:504.29, 505.5, 39, 510.5, 579.17-8. Reath (2013, 269) argues with reference to IV:440, VI:418, XXVII:593, 610, 626 that the *homo noumenon* "is not an obscure metaphysical being, but what a human being conceives himself to be according to the Categorical Imperative".

⁴⁰ See Byrd, Hruschka (2010, 281-2): "As a thing, the *homo phaenomenon* has no duties, which in turn means that he has no rights or moral faculties" (ibid.281). However, in the *Vigilantius* lecture notes, the

takes to be the accused before a moral court.⁴¹ We should therefore not think of the accused before the internal court as the *homo phaenomenon* but as the agent in the most comprehensive sense of the term (animal self, instrumental rationality, subjection to the moral law and capability to act from pure practical reason).

Secondly, Kant distinguishes between the agent or the accused and “someone other” (ibid.438.34), whom the agent has to envisage as the “authorised judge of conscience” (ibid.439.4): “**God**” (ibid.13). In §13 of the *Doctrine of Virtue*, Kant clarifies the role of God for the internal court: God is only a matter of “subjective practical reason”. Nothing follows from this practical presupposition for our theoretical use of reason (ibid.439.20-440.1). A conception of God serves as an “*analogy*” (ibid.440.2) for us to a holy, i.e., morally perfect being. A morally perfect being can serve as a standard, which is *ideal* and *independent* of the agent’s first-order reasoning. Kant’s hope is that even an agent who lacks caution in her exercise of reason will reproach herself, when she considers her use of reason from the perspective of a perfect being that never accepts pseudo-justifications (see VI:76.25-34, 140.12-26, XXVII:197.24-6). I will argue below that we should understand God for the purpose of conscience not as a judge but rather as the *standard* according to which an agent judges herself. An ideal and independent standard is necessary for conscience because conscience could not be effective if the agent judges her use of rational capacities according to the same standards of rigor that she applies in determining what her duty is. If these standards are low, her conscience would never issue guilty verdicts, if they are high, conscience would be redundant.

We also have a “legal adviser (defence counsel)” before the internal court.⁴² In the *Kaehler Lecture Notes*, the internal advocate is called a “twister of the law” [Rechtsverdreher] (LK:201, see also XXVII:359.14). “Rechtsverdreher” is a derogatory German term for a lawyer. Rechtsverdreher are members of the legal profession who work within the legal framework, but who readily make use of any loophole and

human being in appearance or *homo phaenomenon* seems to be able to be morally obligated by one’s rational self (XXVII:579.8-580.22, 593.1-16, 601.30-5). It certainly is an intricate issue to discern how to understand Kant’s distinctions between the *homines* in various places. I will focus on Kant’s official account in the *Metaphysics of Morals*. I am grateful to an anonymous referee for pressing me on this issue.

⁴¹ Sensen (2011, 130) emphasizes that the *homo phaenomenon* is “the human being as observable in experience” and as such entirely a creature of the phenomenal world. The *homo noumenon*, by contrast, is “a moral ideal that might not be observed in experience, but that is prescribed by reason”. For the difficult (causal) relationship between the *homini* when it comes to moral actions see Byrdd, Hruschka (2010, 288-9).

⁴² VI:439.fn.35, see also VI:440.20-4, XXVII:197.30-2, 618.23-4, LK:193, 200. In the *Collins* lecture notes from the 1770s Kant specifies that our “advocate” before the internal court is “self-love”, which “excuses [the agent] and makes many an objection to the accusation” (XXVII:354.20-1).

ambiguity they can find to further the interests of their clients, or to “make use of the letter of the law” to undermine its spirit (LK:201).

It is important that we do not think of the internal court along the lines of the rather silly picture of a number of little agents inhabiting a big agent and arguing with each other. Kant aims to describe different *standpoints* that an agent can take up on herself (see VI:439.fn.). The court metaphor captures central elements of how morality appears to us: Agents find within themselves an ideal or impartial standard of self-assessment, which might be different from how they deliberate about their actions on a day to day basis (the judge). Agents reproach themselves in the light of this standard, as if another person is accusing them (the prosecutor). They can try to come to terms with negative self-assessment either via an earnest resolution to improve (or, in the case of *ex ante* conscience, by not performing the intended action), or by retrospectively finding grounds of excuses, or, more drastically by twisting the law that functions as a standard for self-assessment. Kant’s court metaphor aims to explain how agents can get their objective judgments about their duty wrong and yet still take up another perspective on themselves that lets them cognize that they were mistaken.

There is, however, still the danger that Kant’s metaphor of the internal court fails to do justice to the fact that we often experience critical reflection as a *unity*, not as an interplay between different perspectives. This is particularly so in easy cases when it simply becomes apparent to us that we were wrong. In what follows, I will grant Kant his way of speaking of different perspectives as persons with different roles, but I take issue with the specific metaphor he uses.⁴³ The court metaphor is unhelpful for understanding the process of self-deception about what is justified. It should be clear by now that it is difficult to see how the lawyer’s manoeuvres can be successful before the internal court when an agent engages in rationalizing. The internal court metaphor suggests that the lawyer has to deceive the *judge*, since the judge pronounces the verdict. This might make sense in a real court scenario with a non-ideal judge, but it cannot work if the judge is ideal in the way Kant describes. It would make much more sense if the *defendant*, the finite but rational agent, was the target of the lawyer’s deceptive manoeuvres.

That for rationalizing the focus should be on the defendant is reinforced by a second problem. Kant is explicit that self-deception, which rationalizing is a form of, is

⁴³ Schmidt, Schönecker (2014a, 290-1, 301-11, 2014b, 243) criticise the obscurity of Kant’s judicial metaphors for the functioning of conscience. I am more optimistic than they are that something can be made of Kant’s descriptions if we rationally reconstruct them.

“intentional” [vorsätzlich] (VI:430.12-3) and “deserves the strongest censure” (ibid.35-6). Agents are criticisable for self-deception *itself* not merely for actions committed when self-deceived.⁴⁴ A self-deceiver is not the passive victim of deception, otherwise she would not deserve censure for it. In the court scenario the agent, as the accused, however, is *passively* awaiting a verdict to be passed upon her. It seems that she is doing nothing wrong when her lawyer advances pseudo-arguments and exploits loopholes in order to deceive or sway the judge. After all, we do not blame *defendants* for getting themselves a smart lawyer (though we disapprove of *a legal system* in which the quality of the lawyer impacts the court’s verdict). Rationalizing can only be the agent’s fault if she *chooses* to listen to the lawyer’s pseudo-arguments and gives in willingly to his sophistries. It is noteworthy that this second problem persists even if we are skeptical that agents can really imagine an ideal judge. Even if we admit that the judge of conscience can be non-ideal, we can only explain how self-deception is possible, not how we can hold agents responsible for it, unless we admit that it is the agent who chooses to give in to the more appealing story the lawyer has to offer.

Both problems reveal the underlying issue that Kant’s notion of the internal court cannot adequately account for the possibility of self-deception as something the agent herself is responsible for. Kant, however, might not have intended to do this. It is important that Kant does not say that conscience *is* (identical to) an internal court. He only claims that conscience is the “consciousness” of an internal court (VI:438.10, see also XXVII:197.19-36) or a way to cash out how conscience appears to us. Kant himself might not be unhappy to abandon his model of an internal court for the purpose of talking about self-deception, if there is a better way to understand this process and if this way also preserves the idea that agents can critically reflect on their objective judgements.

Before I suggest such a better way, I will address a potential critical reply to my argument so far. One might think that it is a mistake to take Kant’s metaphor at face value. Maybe we should take it with a grain of salt, not pick it apart the way I did. I think it is important to bear in mind, however, that the metaphor has an important function for Kant. The metaphor of the internal court is supposed to show that and how it is conceivable that agents can get their objective judgments wrong, but yet take up a

⁴⁴ Self-deception is in any case a violation of the “**First Command** of all Duties to oneself [...] *know* (scrutinize, fathom) *yourself*” (VI:441.2-4). See Ware (2009) for more on the duty to know oneself and the danger of self-deception.

perspective on themselves that allows them to become aware of their mistakes and to correct them. Kant must offer a story as to how we can imagine this to happen, otherwise his idea that objective and subjective judgments can diverge seems unwarranted or *ad hoc*. I will now show that there is indeed a way to conceive of self-assessment, according to which agents can critically and impartially assess themselves whilst rationalizing is still possible and a constant threat. I acknowledge that my solution might not satisfy those who are unhappy with Kant's original way of talking about different perspectives to begin with, but I maintain that at least within the Kantian explanatory framework my approach does a better job at accounting for self-deception than Kant's own does.

In an essay on self-deception and the role of philosophy in Kant, Lucas Thorpe (2006, 468) brings up an interesting idea that can be developed into an answer to the question of how rationalizing can be met with success before conscience and how an agent can be responsible for it. Thorpe suggests that the agent might "choose to listen to the defense attorney and try to block out the voice of the judge", and that the voice of the judge might be "drowned out by the pleading voice of the defense attorney" (ibid.). Self-deception here appears as a lack of attention to the incorruptible standard, and undue attention to the wrong source (ibid.469-70). The original standard is still present but, due to interference, eclipsed or bracketed.⁴⁵ Even though Thorpe seems to be unaware of this and he does not develop his thought, he has in mind, I think, rather an *internal panel-discussion* than a court. Rationalizing is the agent's fault, since she herself chose to listen to the lawyer's pseudo-arguments. In an actual court, however, it does not matter for the verdict to whom the defendant pays attention. The notion of an internal panel-discussion seems more fruitful to understand rationalizing than the internal court does.

Self-assessment, according to the panel-discussion-model, means that the judge determines the standard of justification an action has to meet in order for it to be rationally evaluated as morally permissible (or even as obligatory). This standard is the strictness and purity of the moral law. The prosecutor argues that an agent's reasoning about a specific case does or did not proceed with sufficient caution in implementing this standard. The corresponding moral judgement hence is doubtful and a proposed or

⁴⁵ This conception is supported by VI:430.27-35 where self-deception is described not as requiring the active interaction of different faculties of an agent's mind but rather as an agent giving in to biases. In the pre-critical *Kaehler/Collins* lecture notes it is pointed out that human beings have a tendency to rather "listen" to their internal advocate than to their prosecutor (LK:193, XXVII:354.28-9).

already committed action might be wrong. The defence attorney argues against the prosecutor. It is then up to the *agent* to discover whether the prosecutor or the defence attorney present a more compelling case and whether the way she reasoned about an intended or performed action was in line with the standards established by the judge. The responsibility for the final verdict here is shifted from the judge to the accused. *Rationalizing* on the model of an internal panel means that the defence attorney – or, less metaphorically, the agent when taking up this position – questions the standards of the debate and suggests criteria, which are more lenient than those of the judge and more in line with how other agents live their lives, with social norms and customs, religious practices, etc. “Successful” rationalizing means that agents chose to listen to their lawyer more than the lawyer’s arguments warrant, instead of to the prosecutor, and even makes the lawyer’s standards temporarily the standards of her self-assessment. The lawyer here appears as a *sophist* who tries to convince the agent by all means possible to bracket the standards she rationally acknowledges as the right ones.⁴⁶

Kant himself, when he claims in the *Religion* that reason or the inner judge cannot be bribed, adds the caveat that if the agent appeals to “another judge, of whom news will be had through sources of information elsewhere” the agent will have much “to counter the judge’s severity under the pretext of human frailty” and can hope to “mollify” this judge (VI:77.27-35). This represents another sense in which the verdicts of conscience might be infallible. The verdicts of conscience are always accurate according to the standards of self-assessment that prevail at the time. These standards are, however, not always those of the judge. If an agent listens to a less than ideal standard for her self-appraisal, if the agent makes the more convenient standards proposed by her lawyer the standards of her self-assessment, she might believe herself justified before conscience, even when she is not.⁴⁷

Kant’s chief aim in those passages that call into question the success of rationalizing is to stress that rationalizing can never be *completely* successful. An agent can never invent a story that *truly* justifies a moral transgression. Furthermore, an agent can never invent a story that is so persuasive that it numbs the voice of conscience *entirely*, or that the rationalizer is completely and whole-heartedly certain that what she

⁴⁶ In the lecture notes *Powalski*, presumably from around 1782/3, the activity of our internal advocate is explicitly called “sophistry” (XXVII:197.36.).

⁴⁷ According to pre-critical lectures, the judgements of conscience are infallible, unless conscience uses a false standard of judgement (LK:193-4, XXVII:354.26-8).

is doing is right and that she could not have known that it was wrong. The judge or the ideal standard do not disappear, they are merely temporarily silenced by demagoguery. Kant believes that every rational agent, no matter how much she has previously engaged in rationalizing, can be intellectually upright and uncover her mistakes, if she earnestly reflects on her reasoning in the light of purely rational standards. Making a habit of this uprightness is what Kant demands when he claims that while we cannot have a duty to *have* conscience, we do have an indirect duty to *cultivate* conscience (VI:401.19-25).⁴⁸ We have a duty to face pangs of conscience in their full severity and to amend our future reasoning and acting accordingly (see VI:401.19-21) without looking for excuses, let alone foisting biased standards on our self-assessment.

Conclusion

My interpretation of conscience as a reflective watch-dog and internal panel is the attempt to understand conscience based on two important aspects of Kant's philosophy: The way rational agents can come to know of their duty and their propensity to self-deception. There is a tension between Kant's emphasis that self-deception is a pervasive part of human existence and his optimism that conscience can never be fully deceived. My solution is firstly to abandon the notion of an internal court as a model for self-deception, and to rather think of conscience on the model of an internal panel. Secondly, we should understand Kant as maintaining that rational agents always *could have* known that they are mistaken just as they always could have known on a first-order level what their duty was, but that they can also fail to do so. The fundamental fault of the rationalizer is that she does not make the purity and strictness of the moral law the focal point of her deliberation. If she did, she would see through the spuriousness of the considerations her lawyer advances.

My discussion also reveals something important about rationalizing. Rationalizing interferes with first-order reasoning about duty and it also manifests itself as internal sophistry even in moments of calm and critical reflection about one's reasoning. This double function is due to the reflective function of conscience. If conscience does not err unless false standards are assumed and if it can warn every time an agent reasons with lack of caution, the rationalizer must take rationalizing to the next, the reflective, level,

⁴⁸ See also VI:399.4-14. See Timmermann (2006, sec.2-3) for a discussion of the normative status of indirect duty.

since otherwise self-deception could not make the agent feel moral.

There are certainly questions left open by my treatment of conscience and rationalizing. A particularly engaging one is whether it is only the advocate or also the *prosecutor* who can overshoot. After all, some agents might be excessively self-critical and concerned with their moral propriety beyond the ordinary.⁴⁹ Can for such an agent a certain amount of rationalizing sophistry be necessary for a balanced self-assessment? Or, more generally: Just because something is sophistical and biased does not mean that it cannot be right. Is rationalizing *only* a distortion of self-assessment, or can it also sometimes help us discover something about our own reasoning?

Literature

I quote Kant according to *Kant's geammelte Schriften*, edited by the Königlich Preußische Akademie der Wissenschaften, Georg Reimer: Berlin. The First Critique is quoted according to the A/B edition. The *Kaehler Lecture Notes* (LK) are quoted from Stark (2004). Translations from the *Groundwork* are from Timmermann (2011). Other works by Kant are quoted, with occasional modifications, from the *Cambridge Edition of the Works of Immanuel Kant* edited by Paul Guyer and Allan Wood. Translations of untranslated texts (such as LK) are my own.

van Ackeren, Marcel, Sticker, Martin, (2015): "Kant on Moral Demandingness", in: *Ethical Theory and Moral Practice*, vol.18(1), pp.75-89.

Adelung, Johann Christoph (1811): *Grammatisch-kritisches Wörterbuch der hochdeutschen Mundart*, online:
<<http://lexika.digitale-sammlungen.de/adelung/online/angebot>>.

Austin, J. L. (1961): "A plea for excuses", in: Austin (ed.): *Philosophical Papers*, Clarendon Press: Oxford.

Byrd, Sharon, Hruschka, Joachim (2010): *Kant's Doctrine of Right. A Commentary*, Cambridge: CUP.

Crusius, Christian August (1767): *Anweisung vernünftig zu leben, Darinnen nach Erklärung der Natur des Menschlichen Willens die natürlichen Pflichten und allgemeinen Klugheitslehren im richtigen Zusammenhang vorgetragen werden*, Leipzig.

Esser, Andrea (2013): "The Inner Court of Conscience, Moral Self-Knowledge, and the proper Object of Duty (TL 6:437-444)", in: Trampota, Sensen, Timmermann (eds.): *Kant's "Tugendlehre". A Comprehensive Commentary*, de Gruyter: Berlin, Boston, pp.269-293.

⁴⁹ In *Collins* Kant calls a conscience that errs on the side of the prosecutor a "melancholic conscience" [schwermütiges Gewissen] (XXVII:356.38).

Funke, Gerhard (1971): "Gutes Gewissen, falsches bewußtsein, richtende vernunft", in: *Zeitschrift für Philosophische Forschung*, vol.25(2), pp.226-51.

Grenberg, Jeanine (2013): *Kant's Defense of Common Moral Experience. A Phenomenological Account*, CUP: Cambridge.

Guyer, Paul (2000): *Kant on Freedom, Law, and Happiness*, CUP: Cambridge.

– (2010): "Moral feelings in the *Metaphysics of Morals*", in Denis (ed.): *Kant's Metaphysics of Morals. A Critical Guide*, CUP: Cambridge, pp.130-152.

Heubült, Willem (1980): "Gewissen bei Kant", in: *Kant-Studien*, vol.1980, pp.445-54.

Hill, Thomas, JR. (2002): *Human Welfare and Moral Worth. Kantian Perspectives*, OUP: Oxford.

Hoffmann, Thomas S. (2002): "Gewissen als praktische Apperzeption. Zur Lehre vom Gewissen in Kants Ethik-Vorlesungen", in: *Kant-Studien*, vol.93(4), pp.424-43.

Kahn, Samuel (forthcoming): "Kant's Theory of Conscience", in: Muchnik, Thorndike (eds.): *Rethinking Kant*, vol.4, Cambridge Scholars Publishing: Cambridge.

Kuehn, Manfred (2010): "Kant's *Metaphysics of Morals*: the history and significance of its deferral", in: Denis (ed.): *Kant's Metaphysics of Morals. A Critical Guide*, CUP: Cambridge, pp.9-28.

Moyar, Dean (2008): "Unstable Autonomy: Conscience and Judgment in Kant's Moral Philosophy", in: *Journal of Moral Philosophy*, vol.5, pp.327-60.

Møller, Sofie C. (2013): "The Court of Reason in Kant's *Critique of Pure Reason*", in: *Kant-Studien*, vol.104(3), pp.301–20.

Paton, H. J. (1979): "Conscience and Kant", in: *Kant-Studien*, vol.70(3), pp.239-51.

Piper, Adrian (2008): *Rationality and the Structure of the Self. Volume II: A Kantian Conception*, online: <<http://www.adrianpiper.com/rss/index.shtml>>.

Reath, Andrews (2013): "Kant's conception of autonomy of the will", in: Oliver Sensen (ed.): *Kant on Moral Autonomy*, CUP, pp.32-53.

Schmidt, Elke, Schönecker, Dieter (2014a): "Kants Philosophie des Gewissens. Skizze für eine kommentarische Interpretation", in: Egger (ed.), *Philosophie nach Kant. Neue Wege zum Verständnis von Kants Transzendental- und Moralphilosophie*, de Gruyter: Berlin/Boston, pp.279-312.

– (2014b): "Vernunft, Herz und Gewissen. Kants Theorie der Urteilskraft zweiter Stufe als Modell für die Medizinische Ethik", in: Bormann, Wetzstein (eds.): *Gewissen. Dimensionen eines Grundbegriffs medizinischer Ethik*, de Gruyter: Berlin/Boston, pp.229-250.

Schönecker, Dieter (1997): "Gemeine sittliche und philosophische Vernunftserkenntnis.

Zum ersten Übergang in Kants *Grundlegung*", in: *Kant-Studien*, vol.88(3), pp.311-33.

Sensen, Oliver (2011): *Kant on Human Dignity*, de Gruyter: Berlin/Boston.

– (2015): "Kants Begriff des Gewissens", in: Bunke, Mihaylova (eds.): *Gewissen zwischen Gefühl und Vernunft*, Königshausen und Neumann: Würzburg.

Shell, Susan (2009): *Kant and the Limits of Autonomy*, Harvard University Press: Cambridge, Mass.

Stark, Werner (ed.) (2004): *Immanuel Kant. Vorlesung zur Moralphilosophie*, de Gruyter: Berlin, New York.

Sticker, Martin (2015a): "Educating the Common Agent. Kant on the Varieties of Moral Education", in: *Archiv für Geschichte der Philosophie*, vol.97(2), pp.358–87.

– (2015b): "The Rationality and Fallibility of the Common Agent – A Reply to Ido Geiger concerning the Moral-Psychology of Ordinary Agents", in: *British Journal for the History of Philosophy*, vol.23(5), pp.976-89.

Thorpe, Lucas (2006): "The Point of Studying Ethics According to Kant", in: *The Journal of Value Inquiry*, vol.40(4), pp.461-74.

Timmermann, Jens (2006): "Kant on Conscience, 'Indirect' Duty, and Moral Error", in: *International Philosophical Quarterly*, vol.46(3), pp.293-308.

– (2011): *Groundwork of the Metaphysics of Morals* (Translated by Jens Timmermann and Mary Gregor), CUP: Cambridge.

Velleman, David J. (1999): "The Voice of Conscience", in: *Proceedings of the Aristotelian Society*, vol.99, pp.57-76.

Vujošević, Marijana (2014): "The Judge in the Mirror: Kant on Conscience", in: *Kantian Review*, vol.19(3), pp.449-74.

Ware, Owen (2009): "The Duty of Self-Knowledge", in: *Philosophy and Phenomenological Research*, vol.79(3), pp.671-98.

Wood, Allen, Di Giovanni, George (1996): *Religion and Rational Theology*, CUP: Cambridge.

Wood, Allan (2008): *Kantian Ethics*, CUP: Cambridge.