

Why we go wrong: beyond Kant's dichotomy between duty and self-love

Martin Sticker & Joe Saunders

To cite this article: Martin Sticker & Joe Saunders (2022): Why we go wrong: beyond Kant's dichotomy between duty and self-love, Inquiry, DOI: [10.1080/0020174X.2022.2075457](https://doi.org/10.1080/0020174X.2022.2075457)

To link to this article: <https://doi.org/10.1080/0020174X.2022.2075457>



© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 24 May 2022.



Submit your article to this journal [↗](#)




View related articles [↗](#)



View Crossmark data [↗](#)

Why we go wrong: beyond Kant's dichotomy between duty and self-love¹

Martin Sticker^a and Joe Saunders ^b

^aDepartment of Philosophy, University of Bristol, Bristol, UK; ^bPhilosophy, Durham University, Durham, UK

ABSTRACT

Kant holds that whenever we fail to act from duty, we are driven by self-love. In this paper, we argue that there are a variety of different ways in which people go wrong, and we show why it is unsatisfying to reduce all of these to self-love. In doing so, we present Kant with five cases of wrongdoing that are difficult to account for in terms of self-love. We end by suggesting a possible fix for Kant, arguing that he should either accept a pluralistic account of self-love, or move beyond the duty/self-love dichotomy entirely.

ARTICLE HISTORY Received 8 December 2021; Accepted 11 March 2022

KEYWORDS Kant; ethics; wrongdoing; self-love

You are at the supermarket. It's the middle of the pandemic, and stocks are limited. Out of the corner of your eye, you spy a whole box of sea-salted dark chocolate bars. You put one in your trolley, then another, then another. You know that this chocolate is a popular item, and that there is likely not enough for everyone today, but you end up buying all of it. Here you go wrong, taking too much chocolate for yourself. Why? In this case, the answer is simple: You prioritize your wants over the wants of others, and over concerns of fairness.

CONTACT Martin Sticker  martin.sticker@bristol.ac.uk  Department of Philosophy, Cotham House, University of Bristol, Bristol, BS6 6JL

¹The authors are grateful to Jens Timmermann, Robert Stern, Irina Schumski, Lucas Thorpe, Dagmar Wilhelm, and Seiriol Morgan for helpful discussion about this paper. They also want to thank the anonymous referee at *Inquiry* for their comments. We should also thank the audiences at the following workshops and conferences, where we presented earlier versions of this paper: *Perspectives on Evil*, University of Bristol, the *UK Kant Society 2019 Annual Conference*, University of Bristol, the *British Society for the History of Philosophy. Annual Conference*, Kings College London. We are grateful to the audiences and organizers of these events.

© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

This case illustrates wrongdoing driven by self-love in its simplest form: wanting more for yourself at the expense of others (and of fairness). Obviously, many real-life cases are subtler and more significant than this. Moreover, self-love often appears to us not just as a simple urge to hoard chocolate, but rather in the guise of supposedly rational and legitimate claims to resources, recognition, attention, etc. For instance, we might convince ourselves that stocks are only low because everyone who really wants chocolate already bought their provisions and that the rest is thus fair game, or maybe even that, since chocolate is unhealthy, we are taking one for the team and doing others a favour by protecting them from stress eating unhealthy convenience food in the midst of a pandemic.²

In this paper, we take a critical look at a prominent figure who famously holds that all immoral actions are the result of self-love: Immanuel Kant. According to Kant, whenever we fail to act from duty, we are driven by our self-love, where he understands self-love as a principle that unites all our desires and inclinations. Since, according to Kant, all actions are either motivated by duty or self-love, self-love is why we go wrong. In this paper, we argue that there are a variety of different ways that human beings go wrong, and contend that it is unsatisfying to reduce all of these to self-love.

Our argument takes the following structure: In Section 1, we lay out Kant's dichotomy between self-love and morality. In Section 2, we present five cases of wrongdoing that are difficult to account for in terms of self-love. In Section 3, we critically discuss what several contemporary Kantians have said in order to defend Kant's claim that all wrongdoing is motivated by self-love. In Section 4, we revisit the cases outlined in Section 2 and argue that self-love is not always the best explanation for wrongdoing. We maintain that we should be pluralistic about the grounds of wrongdoing. Finally, in Section 5, we end by suggesting a possible fix for Kant on this score, namely, to allow for different grounds of wrongdoing either by admitting of different, irreducible forms of self-love or by moving beyond the self-love/duty dichotomy altogether.

The main upshots of our paper are as follows. Firstly, we confront Kant's framework with a variety of *concrete and everyday* cases of wrongdoing that remain undertheorized in the literature about Kant's

²For more on the role of self-deception in wrongdoing see Papish (2018), Wehofsits (2020), Sticker (2021a).

conception of self-love, wrongdoing, evil and how to react to wrongdoing. This literature often focuses on the infamous murderer at the door case³, and other rather extreme cases such as Eichmann's supposed banal evil alongside his supposed commitment to Kantian morality (see Arendt 1964), and on agents driven by extraordinary vices or 'an evil end, such as hatred, malice or an evil ideology, as an all-encompassing passion' (Formosa 2009, 205). Moreover, discussions of evil in the Kant literature typically focus on more abstract and textual issues such as whether there is a formal proof of evil⁴, how reform of character can take place given that this seems to involve both noumenal as well as phenomenal and temporal aspects⁵, and how to understand Kant's rigorism, according to which a character is either wholly good or entirely evil (VI:20.30-4).⁶ By contrast, we bring Kant's framework to bear on a number of *applied* cases. Secondly, we offer an argument against currently popular so-called 'expansionist' readings of self-love in Kant. Thirdly, we put forward a positive proposal for how Kant could respond to our criticism, suggesting that he can and should accept a range of sensuous incentives that cannot be reduced to one form of self-love. This would allow Kant to account for a wide variety of ways in which we go wrong. It would also involve reconceptualizing his account of our sensuous selves and merely permissible actions for the better, as well as helping us to better understand the burden of proof that is on Kant when he maintains that duty is a motive unlike anything else.

Before we begin, two notes are in order: Firstly, in this paper, we focus on wrongdoing instead of evil. We take wrongdoing to be a more general and broader conception of moral failure than evil.⁷ We are interested in evil (after all, this is a form of wrongdoing), but we are also interested in more mundane wrongs and thus we mainly talk about wrongdoing.⁸ In addition, we distinguish wrongdoing from *radical* evil in the Kantian

³E.g. Korsgaard (1996, ch.5), Varden (2010), Bojanowski (2018).

⁴E.g. Morgan (2005).

⁵E.g. Biss (2014).

⁶E.g. Blöser (2013).

⁷Card (2010, 43) argues that Kant, 'like so many other moral philosophers [...] does not distinguish evils from lesser wrongs'. However, Goldberg (2017) suggests that Kant has the means to distinguish between wrongdoing and evil based on the specific material ends that an agent subverts morality to. Some of these ends involve direct violations of humanity and acting on them is not merely wrong but evil. See also Formosa (2013) for the distinction between evil and moral badness.

⁸See Woods (2021) for a thoughtful account of ordinary wrongdoing, encompassing 'cases of acting from hanger, understandable frustrations, minor but really funny cruelty, crimes of passion and dispassion, and the like' (Woods 2021, 165). However, these are cases of wrongdoing 'not deserving of blame when all relevant information concerning the wronging is known' (Woods 2021, 170). They are thus less severe than the cases we will be addressing.

sense as a universally shared propensity that explains how we can do wrong in the first place, and how this is imputable to us. Radical evil is a separate issue from the concrete instances of wrongdoing we will be talking about.⁹

Secondly, we primarily focus on issues of *individual* wrongdoing. We do so for the sake of simplicity, to get five basic cases of wrongdoing into focus, that do not appear to be motivated by self-love. However, in doing so, we mostly leave unaddressed an important aspect of wrongdoing, namely *collective and social* wrongdoing (rape culture, endemic sexism/racism, etc.).¹⁰

1. Kant's dichotomy between self-love and morality

According to Kant's *Second Critique's* Theorem II, all material i.e. non-moral,¹¹ principles 'come under the general principle of self-love or one's own happiness' (V.22.6-8). This is a strong claim because Kant maintains that this holds for *all* material principles. All actions not motivated by duty fall under the principle of self-love, including breaches of duty as well as merely permissible actions (that are neither obligatory nor forbidden). According to this Theorem, there is just one principle of self-love and all feelings of pleasure that we get from objects that determine the will are 'of the same kind' [von einerlei Art] (V:23.16). This is standardly interpreted as there being a 'common currency' (pleasure or desire-satisfaction) for the non-moral¹², and that all pleasures are commensurate and can be weighed against each other. Furthermore, Kant believes that the dichotomy between material and formal principles is exhaustive.¹³ Thus there is a *single* principle for *every* action not motivated by respect for the moral law: self-love or one's own happiness.

Moreover, Kant has a distinct conception of what happiness is. According to the first *Critique*, happiness is 'the satisfaction of all of our

⁹See Indregard (2020) for a recent discussion and Kohl (2017b, sec.2-5) for critical discussion of the most influential current interpretations of radical evil.

¹⁰See Card (2010, ch.3) for a discussion that brings out the intricacies of collective evil that confronts us with various forms of complicity by group members and institutions and can involve complex interplays between being harmed and harming.

¹¹In this paper, we take 'non-moral' to refer to morally forbidden as well as merely permissible actions. 'Non-moral' does not refer to ethically neutral or indifferent actions (*adiaphora*), the existence of which Kant denies (see VI:22.19-23, and Sticker 2021b, 297-298).

¹²Reath (2006, 50) and Timmermann (2022, sec.1.4) make much of this idea of a common currency.

¹³In addition, Kant assumes that formal principles are exclusively the normative principles of his own theory. This makes it difficult to account for supposedly incorrect but seemingly not-self-love driven principles such as impartial Act-Consequentialism, a principle which can demand great sacrifices from the self-love of an agent (see Sticker, 2020).

inclinations (*extensive*, with regard to their manifoldness, as well as *intensive*, with regard to degree, and also *protensive*, with regard to duration)' (A/B:806/834). Happiness here is presented as nothing else than satisfaction of inclinations (see also IV:405.7-8). In other definitions or explanations of happiness, Kant's focus is less on inclinations and more on positive mental states: '*Happiness* is the state of a rational being in the world in the whole of whose existence *everything goes according to his wish and will*' (V:124.21-3), and '[a] rational being's consciousness of the agreeableness of life uninterruptedly accompanying his whole existence is *happiness*' (V:22.18-9, see also V:23.32-33). These explanations are broader and somewhat less reductive, yet they do make clear that happiness for Kant is not a matter of pursuing objective goods or fulfilling a species function, but instead a matter of positive mental states and satisfying inclinations.¹⁴

For our purpose it does not matter whether the best way to spell out Kant's subjectivist account of happiness is as a form of hedonism or rather as a preference-satisfaction theory.¹⁵ Moreover, on Kant's framework hedonism and preference-satisfaction might ultimately converge. According to Kant, when you act from a principle of self-love the 'determining ground of choice' is 'pleasure in the reality of an object' or the 'feeling of agreeableness that the subject expects from the reality of an object' (V:21-22). Desire satisfaction and expected pleasure are thus intimately connected, and it is difficult to determine whether agents ultimately act for the sake of expected pleasure alone or whether this is just one (necessary) component that drives agents to pursue their ends. In fact, Kant sometimes even indicates that pleasure is the only driving force of *all* actions: 'Every determination of choice proceeds *from the representation of a possible action* to the deed through a feeling of pleasure or displeasure, taking an interest in the action or its effect' (VI:399.21-3, see also V:22.11-4).¹⁶

There are two things worth noting here. Firstly, in the *Religion*, Kant sides with the so-called '*rigorists*' (VI:22.25) who deny the existence of

¹⁴See also Kant's brief explanation of happiness as 'the greatest sum of pleasure' (XX:294.22-23) in his 1793 Prize Essay.

¹⁵Hills (2006) and, to some extent, Reath (2006) read Kant as a preference-satisfaction theorist of happiness. By contrast, Kohl (2017, 519) stresses that 'there is incontrovertible textual evidence that Kant has a hedonistic conception of non-moral motives'. His main text passages for this are A 546/B 578, V:21-25, 62-64, 205-207, VI:215. See also Papish (2018, ch.1).

¹⁶Sticker (2020) argues that Kant was committed to psychological hedonism in the sense that actions without any subjectively motivating component, such as pleasure, are inconceivable for him. Even respect for the moral law therefore can only affect agents through the medium of pleasure and pain (V:73.2-8).

actions or characters that are neither good nor bad or both at the same time ('*adiaphora*' VI:22.20).¹⁷ Kant believes that 'the first subjective ground of the adoption of the maxims, can only be a single one' (VI:25.5-6), i.e. either good or bad, and must be freely chosen, 'for otherwise it could not be imputed' (VI:25.8-9).¹⁸ Kant here indicates that ultimately our most fundamental principle is either the moral law or self-love, and that the other principle is subordinated to our primary or fundamental incentive (VI:36.1-33). Kant further contends that it cannot be the case that some actions are grounded in the moral law and others in self-love without there being a priority of one of the principles over the other within an agent's character. We take this rigorism to be an additional substantive commitment on Kant's part that we cannot discuss any further in this paper. We think that Kant's theory of self-love is beset with problems even if we do not assume his rigorism. However, if Kant's theory of self-love is flawed, in the ways we will go on to suggest, then this could also pose problems for his rigorism.

Secondly, Kant's dichotomy between morality and self-love is not merely accidental, some quirk of his philosophy, but instead has roots in the structure of transcendental idealism. Moral motivation has its ultimate source in the noumenal,¹⁹ whereas self-love has its source in the phenomenal and our human finitude; and there are no other realms or standpoints to ground other forms of motivation.²⁰

However, from transcendental idealism as such it does not follow that each realm or standpoint must have only *one* unifying principle of action or ultimate incentive. Kant is optimistic that morality is without internal contradictions or dilemmas (VI:224.9-26), and that the moral option is thus always (relatively) easy to determine (V:35.13-8).²¹ We can understand why he wants to resist there being different noumenal incentives, as these could pull agents in different directions and make it more

¹⁷Kant is much more open to the possibility of *adiaphora* pertaining to merely permissible *actions* in his discussion of fantastic virtue in the *Metaphysics of Morals* (VI:408.26-409.19).

¹⁸For more discussion of this passage, see Allison's (1990, 35-42; 2011, 114-120) treatment of the Incorporation Thesis; see also Reath (1993).

¹⁹There is debate about whether respect for the moral law should be understood as a form of noumenal causation (see Grenberg 2013, 60-66), or whether the moral law determines the will immediately and respect is merely epiphenomenal (see Guyer 2010). We do not need to settle this debate here. We will briefly come back to this in our final section.

²⁰Transcendental idealism is a complicated issue that we cannot address here. We merely want to note that Kant's dichotomy is not incidental, but something that runs throughout his system as a whole. See Saunders (2016 and 2019) for critical discussion of this dichotomy in the context of Kant's ethics as well as Kahn (2018) who argues against the resulting dichotomy between duty and happiness, through making the case that Kant should accept a duty to promote one's own happiness.

²¹See Timmermann (2013) for discussion of Kant's claim that moral dilemmas are impossible.

difficult, perhaps even impossible, to determine the morally correct option and act on it. This would introduce contingency and uncertainty into moral reasoning. However, Kant thinks that contingency, being pulled in different directions and being unable to determine the best option with certainty is an integral part of the pursuit of happiness (IV:418.1-419.11). In fact, he holds that one major benefit of following moral principles rather than inclinations is that the former allow for certainty (V:35–36), whereas the latter might not even be stable across one and the same agent over time (V:25.25-37). It seems that the prudential sphere, unlike the moral, could allow for different and incommensurable fundamental principles or incentives that do not all come under the same monolithic unifying principle. We will come back to this in our final section.

2. Cases of wrongdoing

We will now present five different cases of wrongdoing, that look like they originate from something other than self-love. Of course, one could dig in their heels and object that, in these examples, the wrongdoing in fact *is* motivated by self-love. We will turn to discuss this in detail in the next two sections, dealing with a variety of general responses that could be made on Kant's behalf (Section 3), and more specific responses to our individual cases (Section 4).

2.1. Doing wrong to benefit others

James is a bad teacher. He doesn't really care about students, and doesn't put any effort into teaching. He is applying for a job, and asks you for a (confidential) letter of recommendation. You really want James to get a job (for his own sake), and so lie about his teaching abilities in your letter.

You do something wrong here, but your (primary) motivation is to benefit someone else, namely James, not yourself.

2.2. Thoughtlessness

Steve is at a crowded gym. He uses the rowing machine for 15 min. He then leaves the machine to do some free weights. Without thinking, he leaves his towel and water bottle on the rowing machine. Peggy was waiting for the rowing machine, but sees that Steve has left his towel and water bottle there, thinks that he might just be taking a short

break and continues to wait. After watching Steve use the free weight for 10 min, she thinks 'what a thoughtless jerk!'.

Steve does something wrong here, but he is not consciously looking to benefit himself, he is merely thoughtless.

2.3. Malpractice and loyalty

Lizzie is a dentist. She has excellent knowledge, and an easy-going nature, which leads to a good rapport with patients. However, she fails to always wash her hands between patients. This is a breach of professional conduct. It is a rule that dentists should wash their hands between patients to prevent infection. Lizzie's behaviour causes harm to some patients.

Moreover, Joe is Lizzie's dental nurse. He thinks Lizzie is an excellent dentist, but notices she doesn't always wash her hands. He feels loyal to Lizzie though. She is a good dentist, and he sees that she cares about patients, so he doesn't flag her lack of hygiene to anyone.

Lizzie and Joe both go wrong here, but neither are motivated by self-love. Lizzie is merely careless, and Joe is too loyal.

2.4. Benevolent paternalism

Martina receives a letter in the mail from Harvard – she got the job! Andrew who knows Martina very well sees the letter before Martina, and knows that she would be very unhappy there. He doesn't want her to be unhappy, so throws away the letter. Andrew doesn't tell Martina about the letter, who assumes she didn't get the job, and doesn't go to Harvard.

In this case, Andrew goes wrong, but not through self-love. It might even be in Andrew's self-interest that Martina goes to Harvard, as he could expect to get invitations from Martina to speak at prestigious Harvard conferences. Instead of self-love what is doing the work here is a desire to benefit Martina without properly respecting her own end-setting or what is commonly referred to as 'autonomy'.²²

2.5. Spite

John buys a beautiful house. The house has large windows, that catch the morning sun. His brother wants to spite him. It turns out he can buy a

²²Kant, of course, has his very own conception of autonomy which is essential to his overall philosophy and differs from our everyday use of the term. See the discussions in Sensen (2012).

small plot of land just next to John's. He buys the land, and builds an awkward ugly house on it, with the sole purpose of blocking out John's morning sun. The house is hideous, and bad to live in. His brother does not enjoy living in it, but does so to spite John.²³

John's brother goes wrong here. He harms John, but in a way that also harms himself (the house is bad to live in), and so this doesn't seem to be an action performed from self-love.

These 5 cases present us with instances of wrongdoing that, at least *prima facie*, do not look like they are motivated by self-love. In the next section (Section 3), we will consider several general responses to this, and in Section 4 we will return to offer critical discussion of each individual case.

Before we turn to that, however, we should note that we are not the first to raise objections against Kant's conception of self-love. Bernard Williams (1985, 64) criticizes Kant for holding 'that all actions except those of moral principle were to be explained not only deterministically but in terms of egoistic hedonism'. In contrast to Kant, Williams (1993, 79–80) thinks that we should not even spell out everything we find of non-moral value as coming under the notion of *happiness*, either understood in a narrow hedonistic sense or even more broadly: 'authenticity [...] submission, trust, uncertainty, risk, even despair and suffering' are things that '[m]en do, as a matter of fact, find value in'. Andrew Reath (2006, 36), who ultimately defends Kant on this issue (see our Section 3), concedes that Kant's dichotomy 'ignores many ordinary activities that give value and substance to life, in which case [his framework] seems radically incomplete. Or it includes them by forcing them into a hedonistic mould that is inappropriate'.²⁴ We are sympathetic to the general gist of these criticisms and concerns, and our five cases spell out and illustrate the

²³This is a real phenomenon called 'spite house'. A google image search for spite houses reveals spectacularly ugly and uncomfortable buildings. Clearly, whoever lives in these houses does not lead an existence in which 'everything goes according to his wish and will' – even if they consciously decided to live in one of these buildings.

²⁴See also Papish (2018, 2–3): 'The worry is that in arguing that all evil arises from self-love or a concern for one's own happiness or pleasure, Kant commits himself to a crude, simplistic, and implausible account of human behavior and motivation. Among the many cases one can put forward as possible counterexamples to the idea that evil involves self-love are: terrorists and religious fanatics; vengeful individuals who do terrible things even though these actions thwart their own happiness; those who appear to do evil at the behest of an authority figure—such as the participants of the Stanley Milgram shock experiments—but not out of self-love; and Adolf Eichmann, particularly as he is immortalized in our public consciousness and through the writings of Hannah Arendt'. Moreover, Stark (1997) argues that Kant's framework is unable to account for important social and moral phenomena such as an overly deferential wife or a person of colour who regards themselves as inferior. Due to their perceived inferiority both of these agents do not always act in their self-interest or follow their self-love and, from a Kantian perspective, their actions of deference to – and sacrifice for – others appear puzzling.

problem with Kant's dichotomy, especially when it comes to the issue of wrongdoing. Our subsequent discussion will evaluate possible responses Kantians have made (Section 3–4), before we put forward a possible fix of our own (Section 5).

That self-love is insufficient to account for all significant cases of moral wrongdoing is also emphasized in contemporary accounts of evil, for instance by Claudia Card (2010, 58) who remarks that '[t]o say only that self-interest is getting priority over morality is to ignore the costs to others of satisfying the prioritized interests and the evils in those interests'. She worries that focus on self-love as a motive puts too much emphasis on the wrongdoer, and overlooks the impact of evil actions *on the victims*. Moreover, Card thinks it matters which specific interests are being prioritized. It makes a great difference to our ethical assessment whether my interest is to exterminate another ethnic group or simply to get by in a world full of evils that I have to deal with as a victim. Both of these can be understood as self-love, but self-love is too coarse to guide ethical assessment of these cases. In our five cases, by contrast, we focus on examples of wrongdoing where it seems that self-love is not what motivates the action at all (regardless of the harm done).

3 General responses available to Kant

We will now rehearse a number of recent sympathetic interpretations of Kant's conception of self-love and discuss their potential to account for the cases in the previous section. In the next section, we will look at the cases again in more detail.

Let us begin with a general reply that could be made on behalf of Kant to all five of our cases. In each of those cases the agents were clearly not forced to do what they did by external powers. They did what they *wanted*, be that spite their brother or help someone get a job or 'save' a friend from a job at Harvard. Even though these actions might not result in pleasure or be in an agent's long-term self-interest, agents seem to act from self-love at least in the sense that they do what they want or desire.

We need to distinguish between two senses of self-love here. On a *broad* conception, assumed in the previous paragraph, acting from self-love just means doing what, at the time of acting, one prefers to do over all other options. To make this a little more concrete, consider a soldier who jumps on a grenade to save your life (for your own sake).

There is a sense in which they have done what they preferred to do, and so in a (very) weak sense what they did was out of self-love. In contrast, if you take all the chocolate from the supermarket during a pandemic, and do so because you want to eat it, and don't care that others won't get any, then you are acting out of self-love in a narrow and substantial sense of the term.

In the broad sense of the term, Kant is right that all immoral actions are from self-love, but this does not tell us anything substantial about these actions. Instead, it is closer to a tautology (more on this shortly). Giving self-love a broad reading has one advantage, though. According to such a reading, Kant's theory of self-love is not a type of crude egoism as one's desires and inclinations can be *other-directed*. Self-love means acting on one's strongest desire(s) and there is no stipulation that the object of one's desire or preferences must be one's own well-being.

On a *narrow* conception of self-love learning that someone performed a non-moral action from self-love would tell us something, because self-love is one possible incentive among a number of other potential candidates for non-moral incentives, such as acting for aesthetic appreciation/enjoyment, acting out of boredom, anxiety, (pathological) sympathy, loyalty, a desire for self-realization, etc. One candidate for such a substantive conception, in ordinary discourse frequently associated with talk about self-interest and self-love, is *selfishness*. If I say that someone is selfish, this tells you something about the person and their actions, namely, that the person prioritizes their own wants, needs, and plans to the detriment of others. We can assume that there are certain activities they will avoid (such as those that would require personal sacrifices for the sake of others) and other activities that we would expect them to engage in (satisfaction of their wants, needs and promotion of their plans at the expense of others).

Importantly though, while selfishness is a substantive and informative conception of self-love that contrasts with other non-moral incentives, it is clearly too narrow to be an adequate conception of self-love in Kant's sense. After all, prioritizing one's wants and needs to the detriment of others implies that we do not accept moral constraints on our actions (or at least accept too few), or that we do not accord others the respect they deserve. This is rather what Kant refers to as 'self-conceit' (V:73.14), which consists of 'claims to esteem for oneself that precede accord with the moral law' and which are 'null and quite unwarranted' (V:73.19-20). We should bear in mind that whilst for Kant all wrongdoing is the result of acting for self-love, not all acting for self-love is selfish and

morally wrong.²⁵ A substantive Kantian conception of self-love should therefore be able to explain *immoral* as well as *merely permissible* actions.

A better candidate for a substantive Kantian conception of self-love than selfishness is hedonism. According to many Kantians, this is indeed the conception of self-love Kant himself subscribes to. Again, this conception helps us understand agents' non-moral (as well as immoral) behaviour to some extent, because it is not tautologically true that all non-moral actions are driven by pursuit of pleasure. If by 'self-love' we mean 'hedonism' then we would expect (somewhat) different behaviour from an agent driven by self-love than from one whose actions are, for instance, to be explained by an objective list theory of happiness instead.²⁶ On such a substantive conception of self-love some non-moral actions might appear puzzling or pointless, and this suggests that these actions were not motivated by self-love. These actions might have other plausible (non-moral) explanations, such as an agent's pride, sense of identity, commitments, etc. What makes a conception of self-love substantive is that self-love does not simply mean that one does what one prefers to do, rather, we gain information if we learn that an action was motivated by self-love.

A number of Kantians have recently endorsed the broad conception of self-love as a reading of Kant's conception of self-love, holding that self-love is quasi-tautological and just means that I do what I desire most.²⁷ However, there has also been criticism against such a broad reading. Laura Papish (2018, 142–143) argues that these so-called 'expansionist' readings of self-love have abandoned Kant's hedonism, because hedonism does not have the same wide scope as the idea that agents do what they want. It is a substantive position, which maintains that, in terms of prudence, agents ultimately act for their own subjective well-being spelled out in terms of positive mental states. Since Papish believes that Kant was clearly a hedonist in this sense, and hedonism is a

²⁵Pure practical reason merely infringes upon self-love, inasmuch as it only restricts it, as natural and active in us even prior to the moral law, to the condition of agreement with this law' (V:73.15–7).

²⁶Of course, the differences would only be partial, as objective list theories tend to include positive mental states (and absence of negative ones) among objective goods.

²⁷See, for instance, O'Neill (2013, 103): 'Acts done to achieve some desired end are all done with the motive of achieving what is desired, i.e. out of self-love'. See also Louden (2000, 138–139), Wood (2010, 145). O'Neill's account would implausibly imply that I act out of self-love if I desire a morally required option, even if my effective motive is duty (see our discussion below for more on this). Reath's (2006, 43) non-hedonist reading of the principle of happiness might also amount to a broad conception of self-love, as on Reath's reading 'one chooses by judging what one will find most satisfying on balance – or what seems equivalent, what one desires most strongly'. This has the additional problem that it is not clear that we would actually have a choice here. After all, it is not a matter of *choice* what we desire most strongly, as Reath himself acknowledges (2006, 55).

substantive conception of self-love, we should reject expansionist readings of self-love.

We find this persuasive. But even without appealing specifically to Kant's hedonism, there are grounds to resist a broad interpretation of self-love as such an interpretation is incompatible with Kant's framework in a way that has been overlooked in the literature so far. If self-love meant *doing what one wants most* then there could not be a duty-self-love dichotomy. After all, if I act from duty, I also do what I want most, otherwise I would not have done what I did. It seems therefore that I acted from self-love, according to the broad reading. But that is not how Kant describes acting from duty. In fact, he is keen to stress that happiness and duty are incentives of radically different kinds (see V:77.19).²⁸

Acting from duty can presumably occur in two ways. Firstly, agents can have strong countervailing inclinations and yet do what is right, because it is right. Here we clearly have a case that counts as acting *from* duty (though agents themselves can never be certain that they acted from duty²⁹) and it would clearly be false to classify this as acting from self-love. After all, if an agent had wanted to give in to inclinations more than acting from duty, they would not have acted from duty. Secondly, there might be cases in which an agent externally conformed to duty and was also motivated by respect, but this is not the only motive. For instance, an agent might help a friend out of duty *and* pathological sympathy, or an agent might have a 'cheerful heart' that can accompany dutiful action (VI:485.5). There is a substantial debate over whether such actions have moral worth.³⁰ It would be unsatisfactory though to settle this debate by merely pointing out that agents did what they wanted and thus necessarily acted from self-love, as this would imply that even actions from duty alone count as acting from self-love.

The broad conception of self-love is thus unhelpful for Kant. In order to make it work in the context of his dichotomy between self-love and duty, he would need to add a clause in order to exclude actions from duty from the scope of self-love. But why would the principle of self-love be: Doing whatever one wants except one's duty? That seems *ad hoc*, especially since at least sometimes agents genuinely *want* to do the right thing

²⁸It is worth noting that Kant himself, at V:22, draws a distinction between the *lower* and *higher* faculty of desire. The lower faculty of desire is what we would typically understand by desire, whereas the higher faculty of desire includes our ability to act for duty's sake. We think that Kant accepts that the lower faculty of desire involves self-love, but the higher faculty of desire does not.

²⁹See VI:38.7-12, 51.7-21, 70.1-71.20, 451.21-36.

³⁰We cannot get into the intricate debate about overdetermination here. See instead Herman (1981) and Ryan Lockhart (2017) specifically for discussion of imperfect duty and acting from respect.

more than anything else (because they have cultivated a virtuous disposition, acquired a cheerful heart, etc.).³¹ The dilemma for a broad conception of self-love is then that either acting from duty turns out to be acting from self-love; or self-love cannot have a very wide scope. But in the latter case it is not clear why only actions from duty should fall outside this scope. Why not other actions (e.g. living in a spite house, being overly deferential to superiors) that represent what an agent wants but that can be explained in terms of self-love only in a tortured manner?

Having put aside this most general response, we now want to think about other strategies available to Kant and Kantians. Doing so will bring out a number of distinctive substantive accounts for why people can go wrong.

Firstly, one explanation of wrongdoing that has support in Kant's text and chimes with the spirit of his philosophy is that wrongdoing is a matter of making *exceptions* for oneself to rules that one wants others to adhere to (see IV:424).³² According to this account, we do not necessarily need to avail ourselves of self-love as an explanatory concept to understand wrongdoing on a Kantian framework, but should instead rather focus on inconsistencies in how agents evaluate themselves and others.

We think that making exceptions is indeed one fruitful way to understand wrongdoing. But there are many interesting cases of wrongdoing that it cannot account for. Two simple cases involve extending who we make exceptions for, such as: making exceptions *for others* ('Yes, he can be a bit of a creep, but ...') and making exceptions for *oneself and others* (Lizzie might not mind if her fellow dentists also do not wash their hands). Moreover, if we ask *why* agents make exceptions for themselves, then the only answer on Kant's framework can be: self-love. Making exceptions certainly tells us something about the structure of some cases of wrongdoing, but it does not provide an alternative *motive*. After all, agents do not usually make exceptions for themselves for the sake of making exceptions for themselves. Often, they make exceptions for the sake of something else, such as chocolate.³³

³¹This *ad-hocness* is, for instance, a problem for Wood (2014, 36) who claims that self-love and inclinations are 'merely placeholders for whatever non-moral incentives might be chosen in preference to those of morality. Kant is not imposing any limits on what one can have an inclination to will'. However, why would self-love only be a placeholder for *non-moral* incentives if there are no limits on the content of what one wills?

³²See *pars pro toto* Sensen (2014).

³³This is not to deny that people sometimes do make exceptions for themselves for the sake of it, for instance, to indulge their sense of entitlement. Morgan emphasizes this aspect of wrongdoing (see below).

Secondly, in the *Religion* (VI:26–27), Kant distinguishes between the predispositions to animality, humanity and personality in human beings, and notes that the distinction between animality and humanity brings with it two different kinds of self-love. The kind of self-love associated with animality he calls ‘*mechanical self-love* [...] for which reason is not required’ (VI:26). This includes self-preservation, the propagation of the species, and the social drive for community with other human beings. As concerns the kind of self-love associated with humanity, Kant sees it as involving comparison with others. This comparison, whilst it can be innocent if agents are just aiming for equality with others, can give rise to what he calls ‘*vices of culture*’, which includes ‘*jealousy, rivalry, envy, ingratitude, joy in others’ misfortunes*, etc’ (VI:27). This is a substantial account of self-love that might enable Kant to accommodate some of our cases. For instance, perhaps in his spite, John’s brother is succumbing to a vice of culture. However, Kant’s account in the *Religion* seems less able to account for cases of *doing wrong to benefit others* or *benevolent paternalism*. In these cases, we are motivated by what we think is good for someone else, not by either mechanical self-love, or jealously, envy, joy in others’ misfortune. In fact, the opposite seems to be the case.

Thirdly, maybe some immoral acts are *motivated by respect* for the moral law (not self-love).³⁴ Textual evidence for this is Kant’s discussion of ‘two crimes deserving death’, infanticide and killing in a duel. In both cases, it is the attempt to preserve one’s honour that leads to a crime. Honour, Kant maintains, ‘is incumbent as a duty’ (VI:335.36–336.6). Agents might act immorally to protect the special status they enjoy as rational agents.³⁵ Kant here does not say explicitly that these actions are motivated by respect for the moral law. Rather, what the agent does is morally wrong, but Kant concedes that the agent is driven to this by sense of honour, and maintaining one’s honour is a duty. Most likely, what Kant has in mind here is an agent who misunderstands what would be required to maintain honour, albeit the misunderstanding is an innocent one, in the sense that it is not driven by self-love but by genuine confusion, confusion created by social expectations, and the complexity and high-pressure nature of the situation.

³⁴Kerstein (2002) has argued that we can act from duty against duty.

³⁵See also a long reflection from the mid to late 1790s in which Kant discusses political revolution. Kant argues that agents will make use of violent and morally prohibited means, such as revolutions, to secure their ‘innate rights’ against political authorities. Such a ‘breach of law’ is rooted in ‘moral propensities’ (XIX:611.12–25).

At least some cases of spite can plausibly be understood as an agent, for instance, John's brother, believing himself to be seriously wronged and acting from a sense of justice, willingly accepting adverse effects on his self-love in order to restore the supposed moral balance of the universe. In this case, there would be immoral actions (if John's brother's spiteful reaction is excessive or unwarranted) not motivated by self-love, but Kant could potentially account for them within his system, if he concedes that *perceived* duty (as opposed to what is actually one's duty) can motivate. In addition, we can see how Andrew's benevolent paternalism might be motivated by a (false) sense of duty towards the wellbeing of Martina.

However, this strategy does not help us to account for cases such as Steve's thoughtlessness or Lizzie's unprofessional neglect of hygiene, since they do not act from perceived duty. Steve and Lizzie, insofar as they reflect about their actions, might believe that their behaviour is permissible and they might tell themselves all kinds of stories to present them as legitimate to themselves, but that a course of action is supposedly permissible is not sufficient to motivate this action. After all, there are usually plenty of different permissible courses of action open to us at any given time. Which of these we take is, at least on Kant's framework, a matter of our inclinations which supposedly all fall under self-love. For the majority of problem cases we presented we are therefore back to the question of how self-love can account for them.

Fourthly, Andrews Reath (2006, ch.6) thinks Kant maintains that we take interest in something *initially* because of the pleasure it yields or promises to yield, but once we do take an interest, it can develop a life of its own in the sense that we can go on to become attached to the objects of our interest for reasons other than pleasure. According to Reath, pleasure plays a causal role in the formation of desires and inclinations, but their objects do not have to be mere pleasure. The principle of self-love or happiness should be understood as a *deliberative procedure* to choose between different options the one that, on reflection and balance, is the most satisfying. This is a reason-guided activity and not a form of hedonism in a problematic sense.

In a similar way, Laura Papish (2018, ch.1) aims to defend Kant's 'cornerstone' claim 'that there are two, and only two, incentives structuring human action' (Papish 2018, 11). She argues that whilst our inclinations are shaped by memories or expectations of pleasure, sometimes beliefs devised to facilitate pursuit of one's self-love go on to develop their own logic in the sense that they might motivate actions even if the interests that initially lead agents to adopt them are no longer present:

[D]evotion to self-love becomes entrenched insofar as self-deception enables self-love to stake out new territory that it did not previously have and that outstrips our initial commitment to securing more banal and immediate objects of desire (Papish 2018, 110).

She (Papish 2018, 109) provides an instructive example for this:

Much like how racial ideologies undergo a metamorphosis in which beliefs that initially promoted material pleasure can outstrip their initial purposes, so too can we develop passions whose connection to our more basic material interests becomes increasingly attenuated.

The example here is of a white supremacist ideology supposed to justify slavery, which was at some point in the material interests of slaveholders and benefited their self-love in a direct way. Once the institution of slavery is abolished (and thus white supremacy cannot serve the function it originally did), agents might still hold on to their racist beliefs, because they have become an integral part of their identity. These beliefs might now come to serve as motivations in their own right. There can be a complex interplay between self-interest and the stories we tell ourselves to justify our pursuit of self-interest.³⁶

Reath's and Papish's accounts share an important general idea, namely that pleasure is important to get acting started in the first place, but once pleasure directs us to an end, activity or belief, we can go on to develop commitments and interests that are not straightforwardly hedonistic.³⁷ We do think that the observation that pleasure or expectation thereof can lead agents to develop commitments that become independent of any (expectation of) pleasure is correct and important. However, we also think that it does not vindicate the claim that *all* action should be understood as motivated by self-love. Human beings seem to be motivated by a variety of things, as our previous examples suggest.

³⁶Papish (2018, ch.1) also argues that we should understand pleasure as a *mediator* between our actions and what ultimately drives them. Hedonism thus functions as a principle of choice and pleasure is supposed to *indicate* well-being but does not constitute it (or not solely so). Papish proposes a number of candidates for what ultimately drives our actions: 'vital force' (Papish 2018, 20), 'welfare' (Papish 2018, 20), 'sense of self' (Papish 2018, 2324), 'the terms [an agent] has laid out' (Papish 2018), and 'egoism' (Papish 2018). However, firstly, the differences between these concepts remain unclear in her account. Secondly, Papish seems to intend them to constitute *objective* standards of well-being, but this contradicts the many passages in which Kant commits himself to a *subjectivist* theory of happiness. Kant makes no mention of a vital force when he spells out his conception of happiness, and the most pertinent passages strongly indicate that agents pursue positive mental states for their own sake (see esp. A/B:806/834, XX:294.22–23).

³⁷The main difference between their accounts is that Reath maintains that Kant's account ultimately amounts to a preference-satisfaction account, whereas Papish maintains that Kant was a hedonist (see Papish 2018, 13–15). However, in a postscript to his paper Reath acknowledges that Kant, on his reading, might still be a hedonist and he is rather concerned that Kant does not turn out to be a *problematic* hedonist in the sense that his conception of happiness casts doubts on his overall theory.

Perhaps Reath and Papish are right that, at some previous point in time, (expectation of) pleasure did motivate us to adopt an end or to pursue an activity that we now pursue for the sake of something other than pleasure. Whether or not this is right is an empirical question. But even if they were right, that would not show that we *currently* act out of self-love if we pursue this end or activity. For instance, perhaps people only come to be genuinely altruistic through initially taking pleasure in helping others. Pleasure here might play a *genetic* or *enabling causal* role. But once people have come to be genuinely altruistic, they act for other's benefit, and not their own pleasure. Likewise, perhaps people only learn to hate others through initially taking pleasure in it or because taking away resources from others and competing with them is a means of preserving and gaining resources that are directly linked to personal pleasure. But that does not mean that once they've come to genuinely hate others, they are doing it from any sense of pleasure. It would just be that, in the causal or genetic story about what enables genuinely hateful behaviour, self-love plays a role. Once that behaviour is enabled, it can be genuine hateful behaviour in its own right, and not an expression of self-love. Papish's and Reath's attempts to save the claim that all non-moral action is ultimately self-love driven is, firstly, an issue that needs to be settled empirically (maybe we do only develop commitments or interests due to pleasure, maybe we do not) and, secondly, it is difficult to see how it helps us better understand *all* human action and its many motives. It provides an interesting possibility for how a contentious claim Kant makes could be maintained, but it does not provide convincing reasons that we *should* maintain this claim.

Fifthly, Seiriol Morgan (2005) presents an interpretation of evil and wrongdoing in Kant that notably departs from hedonistic readings. He argues that we should think of the propensity to evil as just this incentive to embrace unrestrained licence. The picture that emerges is of the human being as of her nature inclined to a kind of gratuitous willfulness, in which she simply fetishizes and elevates to a supreme value, trumping all other considerations, the unlimited indulgence of her whims. (Seiriol Morgan 2005, 85)

The main idea of this reading is that evil and by extension wrongdoing is not an attempt to maximize my pleasure, but is due to the will not wanting to accept any restrictions whatsoever. A will that follows through on this boundlessness can infringe on the wills of others. The propensity to evil lies in the 'self-assertive tendency of the will' (Seiriol Morgan 2005, 91–92) and the 'primal lust for an entirely unrestricted outer freedom' (Seiriol Morgan 2005, 111).

Morgan is upfront that this is a rational reconstruction of Kant, but he also maintains that his interpretation can make sense of more of Kant's core ideas than other interpretations. Textually, he mainly draws on Kant's metaphysics of the will in *Groundwork* III and the *Religion* (IV:446–452, VI:25–37). Kant's claims about happiness and his characterization of self-love in the Second Critique and elsewhere do not play a prominent role for Morgan. He rather is concerned with the structure of the will before it becomes embodied and affected with specific needs and inclinations. The lust for unrestricted freedom is an answer to the question of how violations of the moral law are metaphysically possible, not intended as an account of how specific instances of wrongdoing can be explained. Morgan's conception might thus be correct for the will considered in abstraction, but once we have wills under conditions of embodiment these wills will rather be concerned with specific objects that they want, not with their own freedom.

In addition, Morgan's grand metaphysical narrative seems out of place when applied to mundane instances of wrongdoing. It is unlikely that Joe makes excuses for Lizzie's failing to wash her hands because he does not want to accept any restrictions on his willing. Many mundane cases of wrongdoing have much more straightforward explanations than anarchism of the will. Maybe some of the cases we mentioned can be explained on Morgan's framework though. For instance, spite might be a case where someone rebels against restrictions imposed by others and Steve's thoughtlessness signals that he does not consider the claims and needs of others as imposing meaningful restrictions.

In summary, approaches by Kantians to develop his framework into one that is better able to account for the many aspects and subtleties of wrongdoing are promising, in the sense that each interpretation can plausibly account for some of the cases we brought up. However, none of these suggestions can plausibly account for *all* of our cases.

To see this, we will return to those cases and consider what Kant – and Kantians – should say about each of them.

4. Revisiting our cases

4.1. Doing wrong to benefit others

Earlier, we saw that self-love does not have to be exclusively self-directed. For instance, I might help someone else out, but do so (at least in part) from a sense of vanity. However, this isn't the case in our example. In

our example, I want James to get a job, and so lie about his teaching ability. This need not involve any vanity or grandstanding on my part. I might just be a generally helpful person, and want to help James out. This seems wrong, but not straightforwardly a case of self-love. The problem here is that my act of helping James is unhelpful (and misleading) for the search committee that has to evaluate the letters of reference.

It seems that in this case, I suffer from a (very common) form of partiality, where I am willing to be helpful to people in my surrounding or people that I know at the expense of strangers. Maybe there is an element of self-love here, because my actions might be driven by a desire for harmony and to get along with people in my vicinity, as this might facilitate my own pursuit of happiness. However, we should bear in mind that the letter I write for James is confidential, so there is no obvious *quid pro quo* here. It seems that what does the work here is a partial and pathological form of helpfulness, which Kant himself warns of because it is not constraint by the proper moral principles (V:35.37–8), and it is ultimately not grounded in rational and universal principles, as the case of the philanthropist demonstrates (IV:398.8–399.2). We can observe this form of partial helpfulness frequently in social interaction, but it is odd to think of this as a form of self-love, rather than a misdirected attempt to help.

4.2. Thoughtlessness

Kant could maintain that Steve's thoughtlessness is not an action, as there is no corresponding maxim.³⁸ That might be right, but it would generate additional problems for Kant's account, as it then seems that it would be unwarranted to blame Steve at all. After all, moral principles are supposed to evaluate maxims and anything that does not happen on a maxim is difficult to account for on Kant's ethical framework. In addition, Steve might have a habit of leaving his things on various machines such that it is plausible to assume that this is an expression of principled disregard for others and of entitlement and an inflated sense of self. He seems to think that he can take up more space and resources than others.

Paul Formosa (2009, 199) explicitly acknowledges that thoughtlessness is a phenomenon that Kant must be able to account for. For Formosa the

³⁸Frierson (2019), for instance, suggests that evil means not acting on a maxim. See Papish (2018, 44–45) for criticism of Frierson's proposal.

phenomenon of thoughtlessness (understood following Arendt's 1964 influential analysis of Eichmann) shows that not all deliberately evil actions can be explained by self-deception, as, for instance, Allison (1990, 91 and 159) argues. We should rather explain thoughtlessness as a form of compartmentalization. However, it is unclear how Kant can accommodate thoughtlessness as something agents can be responsible for, on a framework that conceives of all actions as either driven by self-love or duty. Formosa (2009, 204–205) maintains that passions might be a motive independently of self-love, since passions can drive agents to actions that are immoral and also not in their considered self-interest. We think that this is a very welcome extension of the resources Kant has to account for actions, but it is hard to believe that this can account even for all cases of thoughtless, since Steve's thoughtlessness can hardly be understood as driven by 'an evil end, such as hatred, malice or an evil ideology, as an all-encompassing passion' (Formosa 2009, 205). Steve does not have a passion for placing his towels on machines that others want to use. Moreover, the particularly bad motives Formosa mentions do not capture the other more mundane cases we put forward (maybe with the exception of spite).

It seems more promising to think of thoughtlessness as a *sui generis* phenomenon that can be a habit which expresses a sense of entitlement and that we can criticize agents for, but it is not a type of action that purposefully strives for satisfaction of desires or pleasure. It is rather something that, for some people, comes with the way they pursue their ends.

4.3. Malpractice and Loyalty

The right approach here might be close to the one to thoughtlessness. Lizzie isn't really paying attention to the ways in which her actions affect others, but it's not thereby clear that this is a case of self-love in any substantial sense. Lizzie doesn't have an elevated sense of self, and she does really care about her patients, but she is negligent³⁹ when it comes to washing her hands. One additional difference between her and Steve is that Lizzie can be expected to know that she ought to wash her hands. She was taught this in dental school and freely committed to the professional standards of dentistry.

³⁹Card (2010, 54–55) points out that '[c]ulpable negligence is a well-known problem for Kant's ethics, as the negligent may have no maxim of negligence (no intention) to subject to the universality test'.

Why doesn't nurse Joe reprimand Lizzie for not washing her hands and report her to her professional body? Well, for one, he thinks 'it's not my place'. He also sees that Lizzie is a good dentist, and nice person, and feels like he shouldn't be too judgmental. These might be failings, but they don't look like failings of self-love. If anything, Joe errs in the other direction, being too deferential. This is in some sense the opposite of Steve, as Joe's sense of self seems to be too deflated. He thinks it is not his place to criticize superiors and he does not want to come across as judgemental, even in cases in which criticism would be appropriate.⁴⁰

4.4. Benevolent Paternalism

When it comes to keeping the letter from Martina, Andrew thinks he knows better than Martina what is good for her, and maybe that is self-love in some substantial sense, in that he has an elevated sense of self. Importantly, though, Andrew is motivated by altruism and there is nothing that he personally stands to gain. He does assert a certain kind of intellectual superiority (knowing better what is good for someone else) though and maybe he gets the warm glow that people sometimes experience when they deem themselves benefactors.

Kant could potentially account for cases of asserting intellectual superiority via his conception of *logical egoism* (VII:128.31–129.17). The logical egoist deems it unnecessary to consult others and take their criticism into account, because he assumes that he knows better than they anyway. However, whilst this is an intellectual failing or vice, it is unclear how logical egoism, on Kant's framework, can *motivate* actions, if the action is not also a matter of self-love (or duty). Kant's dichotomy here is too strict to be able to account for a phenomenon that Kant himself accepts. This, however, does point to a potential solution Kant can avail himself of to address all of our cases: In his wider writings, Kant clearly shows awareness of various different kinds of wrongdoing and he seems to want to accommodate these; but to do so, he would have to give up or relax the claim that all of these forms of wrongdoing reduce to self-love. We will discuss this in the next section.

⁴⁰Papish (2018, 29–33) argues that there are two models of hedonism. The first is actively striving for pleasure. The second is passivity or the path of least resistance where one allows oneself to be determined by environmental factors. This latter model might be able to accommodate cases of deferring to others and excessive loyalty.

4.5. Spite

John's brother is harming his own subjective well-being by living in a terrible spite house, and so this does not seem like acting straightforwardly from self-love. However, that being said, he is also indulging his pettiest revenge fantasies at the expense of his brother, and this does not seem entirely removed from self-love. In addition, we saw that Kant does acknowledge the possibility that agents act wrongfully motivated by a sense of duty. However, once more the question arises how this motivation can be explained on Kant's framework.

One such explanation we can extrapolate from Anna Wehofsits (2020) recent instructive discussion of the relation between passions and self-deception in Kant. Whilst many forms of rationalizing seek to devalue duty by calling into questions its exceptionlessness or purity (IV:405), passions can give rise to a form of rationalizing or self-deception, which seeks to revalue and elevate them over duty, and allows passions to increase their hold on an agent (Wehofsits 2020, 8). This can lead to a state of delusion in which agents deem their passions morally justified, and maybe even think that pursuing them is obligatory (Wehofsits 2020, 21). This might exactly be the situation that John's brother finds himself in, when he chooses a lodging that is by all normal standards inferior, but allows him to spite someone else's happiness. Again, Kant here has resources to accommodate this case of wrongdoing, but the story he can tell does not sit well with his claim that everything non-moral boils down to self-love, since passions are not only an obstacle for morality but also for an agent's long-term self-interest. They are '*pragmatically ruinous*' (VII:267.6), because the satisfaction of everything else is put on hold by them.

5. Pluralism about self-love

We hope to have established that there are a variety of motives for going wrong (e.g. not wanting to accept restrictions, supposed intellectual superiority, inflated and deflated senses of self, a misguided sense of justice, thoughtlessness, etc.), and it is rather forced to reduce these all to one overarching incentive.⁴¹ In doing so, we also lose part of what is

⁴¹The general problem is put very well by Kekes (2005, 4): 'Most of the explanations given in the framework of the religious or the Enlightenment world view assume that evil has a single cause. Evil, however, has many causes: various human propensities; outside influence on their development;

distinctively bad about some of these forms of wrongdoing. We also hope to have established that more subtle interpretations of Kant and developments of his philosophy can accommodate some of the cases we brought up, but that no single interpretation or development of self-love can plausibly account for *all* cases. Moreover, we contend that some of the resources Kant has available to explain different cases of wrongdoing would be more plausible without any appeal to self-love as an overarching principle.

We want to end by suggesting a possible fix for Kant. Kant should acknowledge that the grounds of wrongdoing are not, at bottom, all 'of the same kind' (V:23.16) or commensurate with each other and reducible to a single unifying principle. Acknowledging this would allow Kant to account for the complexity of human motivation, as he in fact does in his discussion of the shortcomings of an ethics founded on happiness. In these discussions, Kant conceives of the phenomenal world as messy and contingent and of striving for anything other than morality as always insecure. There is no need for Kant to maintain that all non-moral actions are structured by one principle. In fact, admitting of a pluralism for non-moral motives would be much more in line with Kant's overall take on the phenomenal world and the human beings in it.

Whilst we think this fix is sensible in the light of our criticism and given the general outlines of Kant's theory, it raises the question of how revisionary this change would be with regards to Kant's conception of moral-psychology, agency and his argument for the special normative status of the moral law. We might particularly wonder whether this means that Kant should give up his statement that all non-moral, principles 'come under the general principle of self-love or one's own happiness' (V.22.6-8). This statement occurs early on in the Second Critique, as the second of four theorems, and its programmatic nature suggests that this is a claim of considerable significance for Kant. There are two options here.

Firstly, a relatively non-intrusive fix is to accept that there are different non-moral principles that cannot be reduced to each other, but to still maintain that all material/non-moral ends/actions etc. come under *a* principle of self-love, where there are various different (potentially incommensurate) principles of self-love: spite, arrogance, desire for pleasure, misguided sense of justice, supposed intellectual superiority, and so on.

and a multiplicity of circumstances in which we live and to which we must respond. Because these causes vary with person, time, and place, an attempt to find the cause of evil is doomed. There is no explanation that fits all or even most cases of evil'.

This would allow Kant to retain a minimally modified second theorem, and to maintain an exclusive dichotomy between duty and self-love.

However, one might worry that this is a largely verbal manoeuvre to make something Kant says come out true even though the underlying picture (pluralism about non-moral incentives) could require a more drastic revision of Kant's position. A related worry is that this looks like an *expansionist* account of self-love(s), and so faces some of the criticisms that we laid out in Section 3. After all, on this version of the theorem Kant would hold that everything can be reduced to *a* form of self-love, except morally worthy actions. This raises the question of why it would be the case that very different incentives, such as the desire for pleasure, the entitlement to take up more space or resources, supposed superiority over others, a misguided sense of justice, i.e. everything that represents possible grounds of motivation, *except duty*, come under the heading of a principle of self-love. We would once more have a framework for everything that an agent desires, except doing the right thing, and it is unclear why doing the right thing would be excluded from this. It would be more consistent to maintain that everything comes under a principle of self-love and one of these principles is respect for the moral law. However, this is something Kant would not accept.

This takes us to our second option: a more revisionary suggestion that avoids the above problem and better accommodates the plurality of incentives, through giving up the claim that all non-moral principles 'come under the general principle of self-love or one's own happiness'. After all, if we accept a pluralism about self-love, where self-love can manifest as spite, arrogance, or a misguided sense of justice, etc. then why not just accept a pluralism about wrongdoing in the first place, where people can be motivated to go wrong in a variety of different ways, without stipulating that these different ways have to come under a principle of self-love. It is sufficient for Kant's overall theory to maintain that the moral incentive, which can endow actions with moral worth, is distinct from all other incentives and rooted in pure practical reason. Or, in other words, Kant could accept that there are many different kinds of sensuous incentives that motivate actions, and that acting on maxims incorporating these incentives is at times permissible, and at other times impermissible⁴², while still maintaining his key thought that there is also one special incentive rooted in pure practical reason. We

⁴²There might also be incentives that we can never incorporate into permissible maxims, such as an incentive to degrade another's humanity (cf. for instance VI:450.3-5).

think this suggestion better accommodates the plurality of grounds that motivate human beings. We also think it chimes with a key belief of Kant's, namely that all incentives other than respect are rooted in the sensuous world, a world riddled with contingency.⁴³

Let us finally elaborate on some of the implications and costs of this proposal, some of which, we think, will make Kant's moral psychology more appealing, where others constitute costs that we think Kantians should be willing to pay.

Firstly, whilst we have focused on *immoral* actions, giving up on the claim that self-love is the only alternative to duty also allows for a more nuanced understanding of *merely permissible* (as opposed to immoral or morally required) actions. For instance, we spend some time in the morning tending to our tomato plants. This is not done from duty, but it's also not clearly done from self-love. It's done from a sense of concern about the plants. Life is full of actions that are neither done from duty, nor from self-love, and our proposal helps us make more sense of these. We take it that potentially many merely permissible actions are motivated by something other than self-love or duty, although elaborating on this is beyond our scope here.

Secondly, so far we have discussed Kant's claims about self-love primarily as a *motive*. But when Kant introduces his claim that all material i.e. non-moral, principles 'come under the general principle of self-love or one's own happiness' at V.22.6-8, he is not only talking about our

⁴³There is a sense in which even on this revisionary conception Kant could maintain a claim to the effect that all non-moral incentives come under self-love, namely, if he made clear that he uses 'self-love' in a highly technical sense as by definition referring to all non-moral incentives whatever they are, and that this does not imply that these incentives have all something in common other than that they are non-moral. This would not be self-love in a substantive sense, and it would be an open task for philosophers to come up with proposals for unifying principles (if such principles can indeed be found). We should not accept that these unifying principles exists as a matter of stipulation or simply because there is a term for the category of all incentives that are non-moral. In fact, since we are here talking about the phenomenal side of human nature, the task of finding unifying principles (if they exist) might be one for empirical psychologists or anthropologists rather than philosophers. These principles might not be *a priori* but simply general (but contingent) features of human volition. Common structures of all non-moral incentives might or might not exist. Availing oneself of a term such as self-love does in no way necessitate or even indicate that they do. Moreover, we might wonder whether even such a technical and non-substantive use of the term can do justice to cases in which agents do wrong because of a misconception of what their duty is. We might rather want to treat this as a third option, neither an admirable acting from duty, nor similar to selfishness, thoughtlessness, etc. Furthermore, the technical use of self-love still lumps together incentives that motivate merely permissible actions (such as tending to our plants, see below) and incentives that motivate immoral actions, even severe wrongdoings such as torturing someone for fun. Yet again, it is not clear that we gain much by availing ourselves of this technical term, other than that a statement Kant makes is in some sense true. Finally, if self-love is merely a technical term, this would still put pressure on Kant's argumentative strategy in the way indicated below, because such a technical term cannot do substantive philosophical work. We are grateful to an anonymous *Inquiry* referee for discussion of this point.

motivations, Kant also wants to establish the *normative* priority of the moral law over all non-moral or material principles.

A worry here is that perhaps, in introducing a plurality of different possible incentives, we are complicating Kant's moral philosophy in an unhelpful way. After all, if it were not for the pluralism we propose, Kant has a straight-forward argument for the specialness and uniqueness of the moral law via a bottom-up strategy that he avails himself of in the Second Critique's Remark I to Theorem II. There, he argues that all non-moral principles, be they concerned with bodily, intellectual or cultural pleasures, have in common that they operate through pleasure (V:23.13-4) and contain 'no determining ground for the will other than such as it is suitable to the lower faculty of desire' (V:24.33-4), and then he contrasts this with the moral principle. Kant here suggests that we can isolate the higher faculty of reason and its principle only if the lower faculty's principles all boil down to a monolithic drive for happiness.

We do think that once hedonism about non-moral principles is abandoned it becomes potentially more difficult for Kant to argue for the singularity and special, elevated status of the categorical imperative, since he cannot simply point out that principles other than morality reduce to self-love. Instead, Kant might have to inspect each candidate principle one at a time and argue for each individually why they are unsuitable as a supreme principle of morality. This potentially opens the door to challenges to the categorical imperative. However, if our argument so far is plausible, then we might just have to accept this. Our psychological lives are more complicated and more nuanced than Kant makes them out to be in his discussion of the Second Theorem, and one upshot of our discussion is that there is no quick and easy way to demonstrate the authority and special status of morality by asserting that everything else comes under the principle of self-love.

Yet, we do not think that this means that Kant's overall argument for the special status of the moral law falls short. After all, he has a number of arguments against principles founded in the lower faculty that potentially extend to principles other than self-love. For instance, that we can never be certain of consequences of actions (IV:418.1-37, VI:215.24-216.6) is a problem not merely for self-love in the narrow sense, but also for other consequence-based principles: We can never be sure that lying in James' letter of reference will have the desired consequences. No one might actually read the letter. John's brother cannot be certain that the spite house will actually annoy John. Maybe John genuinely likes his brother and is glad that he now lives next to him.

Kant avails himself of the nuanced argumentative strategy that we are calling for at the very beginning of *Groundwork* I when he establishes the unconditional goodness of the good will by showing how other goods, such as classical virtues and even happiness, are merely conditionally good.⁴⁴ He does not simply claim that they are all just principles of self-love, but rather presents specific, if very dense, arguments against each of them, showing how these goods could also be bad (IV:393.7–394.12).⁴⁵ Even more importantly, Kant also has a number of top-down strategies to establish the moral law as special, such as the deduction (§⁴⁶) in *Groundwork* III and the Fact of Reason, which, supposedly, shows that the moral law affects us differently from anything else.⁴⁷

Thirdly, part of why Kant seems to think the moral law will enable a feeling of respect sufficient to restrict the claims of self-love is that the pain of respect contrasts so directly with the pleasure sought in wrongdoing (V:73.2–75.19). Respect is effective because it strikes down inclination. This raises the question as to whether our pluralistic account of wrongdoing would necessitate Kant to substantially revise his conception of respect.

The answer to this is ‘no’, if we think that respect is epiphenomenal, merely the way we experience pure practical reason’s determination of the will.⁴⁸ Maybe our account would require some phenomenological variations, depending on what it is that respect strikes down; it might feel different when respect strikes down my inclination to buy all the chocolate in the supermarket (and not leaving any for others), than it does when respect reminds me not to be thoughtless and occupy more than

⁴⁴For recent discussion of Kant’s arguments here, see Watkins (2018).

⁴⁵The fact that, in following our proposal, some alternative candidates for supreme goodness are more difficult to dismiss for Kant, such as ‘moderation in affects and passions, self-control and sober deliberation’ (IV:394.4–5) or ‘strength of soul in overcoming obstacles’ and ‘talents of spirit’ (V:24.4–5) strikes us as an advantage, since these talents and qualities, whilst not unconditionally good, are more conducive to morality on a Kantian framework than the pursuit of pleasure. If Kant simply dismissed self-control, courage and other qualities that are central to many ethical approaches as coming under self-love, we should question whether Kant, at the start of the *Groundwork*, did full justice to their significance. Elsewhere, Kant does recognize the ethical significance of these or similar virtues (e.g., VI:408–409).

⁴⁶Schönecker (2006), thinks that there is only *one* deduction in *Groundwork* III, namely a deduction of the supreme principle of morality; cf. Allison (2011, 274–275) who contends that, while establishing the supreme principle of morality is Kant’s ultimate goal in *Groundwork* III, it is not the only thing that he provides a deduction of. For further discussion of the argumentative strategy of *Groundwork* III, see Allison (2011, 273–363), Bojanowski (2017), Saunders (2021), Timmermann (2007, 120–151).

⁴⁷This phenomenological dimension of the Fact of Reason has recently been stressed by Grenberg (2013), and Ware (2021, 44–70).

⁴⁸This reading is, for instance, supported by the beginning of the Second Critique’s Incentive Chapter (V:71.28–72.232)

one machine in the gym. Yet, the underlying noumenal mechanism would be similar.

Matters could be different, though, if we think of respect as not merely epiphenomenal, but a motive. On this picture, perhaps we do need a more nuanced story. Presumably, respect would always involve the striking down or counteracting of something empirical, but not necessarily of self-love. Respect might counteract and contrast with spite and loyalty, might warn agents that what they are doing might appear morally obligatory to them, but in fact it is not,⁴⁹ that they might be well-intentioned but still act paternalistically, etc.⁵⁰ The moral law would still serve as a corrective, and would help to guide us, but it would not always be a corrective to self-love, but instead a corrective to the variety of different ways in which we can go wrong.

Conclusion

We hope to have shown that not all cases of wrongdoing can be reduced to the principle of self-love. In doing so, we have offered an argument against expansionist readings of self-love in Kant. We have also confronted Kant's moral philosophy with a new variety of everyday cases of wrongdoing, and have discussed how these put pressure on his exclusive dichotomy between morality and self-love. Finally, we put forward a potential solution for Kant, suggesting that he should accept a range of sensuous incentives, and give up the idea that everything non-moral is commensurate and can be reduced to a principle of self-love, and even to self-love altogether.

Disclosure statement

No potential conflict of interest was reported by the author(s).

⁴⁹Respect would here overlap with the function of conscience, the inner judge of oneself (VI:186.17–8, VI:400.27–8).

⁵⁰Interestingly, in the Incentive chapter, Kant talks about *feeling* rather than pleasure as the mediator between non-moral principles and actions (e.g. V:71.30–4, 72.34–73.2, 74.30–3). This contrasts with Kant's claims made elsewhere that pleasure is the only driving force of all actions (see V:222.11–4, VI:399.21–3 and our sec.1). This is presumably because Kant here contrasts non-moral and moral motivation, both of which operate through feeling, the latter through respect for the moral law a 'feeling that is produced by an intellectual ground' (V:73.34–5). Feeling thus seems to be the broadest possible category for that which mediates principles and actions. It is beyond the scope of our paper to discuss whether Kant should also give up the idea that non-moral (and moral) action must always involve feeling. If feeling is just a stand-in for a conative element, nothing in our discussion is at odds with what Kant says. However, we are tempted by the thought that we can be motivated in both moral and non-moral matters by reasons alone, but this would involve a theory of motivation that would take us beyond Kant. For the classic statements of such a theory, see Nagel (1970, 29–30) and McDowell (1998, 79), and for discussion of how Kant himself might adopt such a view, see Saunders (2021).

Literature

Kant's Works (with the exceptions of the First Critique) are quoted volume:page.line of the *Academy Edition*. Translations follow the *Cambridge Edition of the Works of Immanuel Kant*, edited by P. Guyer and A. W. Wood.

ORCID

Joe Saunders  <http://orcid.org/0000-0003-1290-275X>

References

- Allison, H. 1990. *Kant's Theory of Freedom*. Cambridge: Cambridge University Press.
- Allison, H. 2011. *Kant's Groundwork of the Metaphysics of Morals. A Commentary*. Oxford: Oxford University Press.
- Arendt, H. 1964. *Eichmann in Jerusalem*. New York: Viking Press.
- Biss, M. 2014. "Kantian Moral Striving." *Kantian Review* 20 (1): 1–23.
- Blöser, C. 2013. "Grade der Tugend und Rigorismus." In *Akten des XI. Internationalen Kant-Kongresses*, 51–62. Berlin: de Gruyter.
- Bojanowski, J. 2017. "Kant on the Justification of Moral Principles." *Kant-Studien* 108 (1): 55–88.
- Bojanowski, J. 2018. "Thinking About Cases: Applying Kant's Universal Law Formula." *European Journal of Philosophy* 26 (4): 1253–1268.
- Card, C. 2010. *Confronting Evils. Terrorism, Torture, Genocide*. Cambridge: Cambridge University Press.
- Formosa, P. 2009. "Kant on the Limits of Human Evil." *Journal of Philosophical Research* 34: 189–214.
- Formosa, P. 2013. "Evils, Wrongs and Dignity: How to Test a Theory of Evil." *The Journal of Value Inquiry* 47 (3): 235–253.
- Frierson, P. 2019. "Character in Kant's Moral Psychology: Responding to the Situationist Challenge." *Archiv für Geschichte der Philosophie* 101 (4): 508–534.
- Goldberg, Z. 2017. "Can Kant's Theory of Radical Evil be Saved?" *Kantian Review* 22 (3): 395–419.
- Grenberg, J. 2013. *Kant's Defense of Common Moral Experience. A Phenomenological Account*. Cambridge: Cambridge University Press.
- Guyer, P. 2010. "Moral Feelings in the Metaphysics of Morals." In *Kant's Metaphysics of Morals. A Critical Guide*, edited by Lara Denis, 130–152. Cambridge: Cambridge University Press.
- Herman, B. 1981. "On the Value of Acting from the Motive of Duty." *Philosophical Review* 90 (3): 359–382.
- Hills, A. 2006. "Kant on Happiness and Reason." *History of Philosophy Quarterly* 23: 243–261.
- Indregard, J. 2020. "Every Man has his Price: Kant's Argument for Universal Radical Evil." *Inquiry*, doi:[10.1080/0020174X.2020.1724564](https://doi.org/10.1080/0020174X.2020.1724564).

- Kahn, S. 2018. "Kant and the Duty to Promote One's own Happiness." *Inquiry*, doi:10.1080/0020174X.2018.1446047.
- Kekes, J. 2005. *The Roots of Evil*. Ithaca, NY: Cornell University Press.
- Kerstein, S. 2002. *Kant's Search for the Supreme Principle of Morality*. Cambridge: Cambridge University Press.
- Kohl, M. 2017b. "Radical Evil as a Regulative Idea." *Journal of the History of Philosophy* 55 (4): 641–673.
- Kohl, M. 2017. "The Normativity of Prudence." *Kant-Studien* 108 (4): 517–542.
- Korsgaard, C. 1996. *Creating the Kingdom of Ends*. Cambridge: Cambridge University Press.
- Louden, R. 2000. *Kant's Impure Ethics*. Oxford: Oxford University Press.
- McDowell, J. 1998. *Mind, Value, and Reality*. Cambridge: Mass.: Harvard University Press.
- Morgan, S. 2005. "The Missing Formal Proof of Humanity's Radical Evil in Kant's Religion." *The Philosophical Review* 114 (1): 63–114.
- Nagel, T. 1970. *The Possibility of Altruism*. Princeton, NY: Princeton University Press.
- O'Neill, O. 2013. *Acting on Principles. An Essay on Kantian Ethics*. Cambridge: Cambridge University Press.
- Papish, L. 2018. *Kant on Evil, Self-deception, and Moral Reform*. Oxford: Oxford University Press.
- Reath, A. 1993. "Intelligible Character and the Reciprocity Thesis." *Inquiry* 36: 419–430.
- Reath, A. 2006. *Agency and Autonomy in Kant's Moral Theory. Selected Essays*. Oxford: Oxford University Press.
- Ryan Lockhart, J. 2017. "Kant on the Motive of (Imperfect) Duty." *Inquiry* 60 (6): 569–603.
- Saunders, J. 2016. "Kant and the Problem of Recognition: Freedom." *Transcendental Idealism, and the Third-Person. International Journal of Philosophical Studies* 24 (2): 164–182.
- Saunders, J. 2019. "Kant and Degrees of Responsibility." *Journal of Applied Philosophy* 36 (1): 137–154.
- Saunders, J. 2021. "Some Hope for Groundwork III." *Inquiry*, doi:10.1080/0020174X.2021.1997798.
- Schönecker, D. 2006. "How is a Categorical Imperative Possible?" In *Groundwork for the Metaphysics of Morals, Christoph Horn and Diether Schönecker*, 302–323. Berlin: de Gruyter.
- Sensen, O. 2012. *Kant on Moral Autonomy*. Cambridge: Cambridge University Press.
- Sensen, O. 2014. "Universalizing as a Moral Demand." *Estudos Kantianos* 2: 169–184.
- Stark, C. 1997. "The Rationality of Valuing Oneself: A Critique of Kant on Self-Respect." *Journal of the History of Philosophy* 35 (1): 65–82.
- Sticker, M. 2020. "Kant, Eudaimonism, Act-Consequentialism and the Fact of Reason." *Archiv für Geschichte der Philosophie* 102 (2): 209–241.
- Sticker, M. 2021a. *Rationalizing (Vernünfteln)*. Cambridge: Cambridge University Press.
- Sticker, M. 2021b. "Kant, Moral Overdemandingness and Self-Scrutiny." *NOUS* 25 (2): 293–316.
- Timmermann, J. 2007. *Kant's Groundwork of the Metaphysics of Morals. A Commentary*. Cambridge: Cambridge University Press.

- Timmermann, J. 2013. "Kantian Dilemmas? Moral Conflict in Kant's Ethical Theory." *Archiv für Geschichte der Philosophie* 95 (1): 36–64.
- Timmermann, J. 2022. *Kant's Will at the Crossroads. An Essay on the Failings of Practical Rationality*. Oxford: Oxford University Press.
- Varden, H. 2010. "Kant and Lying to the Murderer at the Door ... One More Time. Kant's Legal Philosophy and Lies to Murderers and Nazis." *Journal of Social Philosophy* 41 (4): 403–421.
- Ware, O. 2021. *Kant's Justification of Ethics*. Oxford: Oxford University Press.
- Watkins, E. 2018. "The Unconditioned Goodness of the Good Will." In *Kant on Persons and Agency*, edited by Eric Watkins, 11–28. Cambridge: Cambridge University Press.
- Wehofsits, A. 2020. Passions: Kant's psychology of self-deception. *Inquiry*, <https://www.tandfonline.com/doi/full/10.1080/0020174X.2020.1801498?scroll=top&needAccess=true>.
- Williams, B. 1985. *Ethics and the Limits of Philosophy*. Cambridge: Mass.: Harvard University Press.
- Williams, B. 1993. *Morality. An Introduction to Ethics*. Cambridge: Cambridge University Press.
- Wood, A. 2010. "Kant and the Intelligibility of Evil." In *Kant's Anatomy of Evil*, ed. Pablo Muchnik and Sharon Anderson-Gold, 144–172. Cambridge: Cambridge University Press.
- Wood, A. 2014. "The Evil in Human Nature." In *Kant's Religion Within the Boundaries of Mere Reason: A Critical Guide*, edited by Gordon Michalson, 31–57. Cambridge: Cambridge University Press.
- Woods, J. 2021. "Ordinary Wrongdoing." *Oxford Studies in Normative Ethics* 11: 155–175.