Journal of the American Philosophical Association (2016) 420–436 © American Philosophical Association
DOI: 10.1017/apa.2016.20

Is Punishment Morally Justified? Developing Nietzsche's Critique of Compatibilism in *The Wanderer and His Shadow*, Section 23

ABSTRACT: *Nietzsche is mostly known for denying moral responsibility on account of lack of libertarian free will, thus betraying an incompatibilist approach to moral responsibility. In this paper, however, I focus on a different, less familiar argument by Nietzsche, one that I interpret as a critique of a compatibilist conception of moral responsibility. The critique shows why punishment and our moral sanctions in general are morally unjustified by the compatibilist's own lights. In addition, I articulate what I call Nietzsche's explanatory challenge, which challenges the compatibilist to explain the performance of an immoral action without appealing to conditions that would exempt or excuse the wrongdoer or otherwise relieve the wrongdoer from responsibility and would thus make punishing the wrongdoer morally unjustified. By drawing on the work of R. Jay Wallace, I reconstruct Nietzsche's anticompatibilist argument and defend it against four possible objections.*

KEYWORDS: Nietzsche, compatibilism, responsibility, punishment, desert

Introduction

Nietzsche is well known for his critique of moral responsibility and of the moral justification for the attendant practices of blame and punishment.[1] Starting especially in his so-called middle period (*Human, All Too Human*, which includes *The Wanderer and His Shadow, Daybreak,* and the first four books of *The Gay Science*),[2] Nietzsche consistently attacked the notion of moral responsibility and the related moral practices, arguing that they are philosophically indefensible. Nietzsche's criticism of these notions took the form of arguing, first, that the libertarian belief in the idea of a metaphysically free will is an error, and second, that

1 Throughout this paper I will use the term 'punishment' broadly to refer not merely to various forms of corporeal punishment but to any kind of behavior meant to cause some kind of displeasure in the person who is perceived to have violated some moral norm—a displeasure conceived as deserved by the wrongdoer precisely insofar as he or she violated the moral norm.

2 I use the standard abbreviations for Nietzsche's works, a glossary of which can be found in the reference list. The German edition of Nietzsche's works used is the *Kritische Studienausgabe*.

holding people accountable on the basis of such an erroneous belief is unjustified.[3] This line of thought, focused as it is on the denial of a metaphysically free will and the consequent denial of moral responsibility, evinces clear incompatibilist leanings on Nietzsche's part at this stage of his career.[4]

While Nietzsche's arguments from this period take several different forms (see, for example, HH 18, WS 11), one central argument presupposes the truth of determinism, where determinism is the view that a description of the total state of the world at time *t* and the laws of nature entail a description of the total state of the world at time *t*+1. Thus, in response to the impression that a waterfall might make on us as something that enjoys 'freedom of will and capriciousness [*Freiheit des Willens und Belieben*]' (HH 106), Nietzsche claims that

> everything here is necessary [*nothwendig*], every motion mathematically calculable. So it is too in the case of human actions; if one were all-knowing, one would be able to calculate every individual action . . . for if one moment the wheel of the world were to stand still, and there were an all-knowing, calculating intelligence to make use of this pause, it could narrate the future of every creature to the remotest ages and describe every track along which this wheel had yet to roll. (HH 106, cf. HH 39)

Though Nietzsche does not employ the term 'determinism' explicitly, he could be seen here as invoking the famous image of the Laplacean demon, who can calculate the entire course of the world using the laws of nature and the current state of affairs. On the basis of this deterministic view, Nietzsche constructs a familiar argument. He holds that given the truth of determinism, every element of our psychological makeup, every decision we make and every action we perform is necessary. From this necessity, it follows that no person could have acted differently than he actually did and that consequently 'man can be made accountable for nothing, not for his nature, not for his motives, not for his actions, nor for the effects he produces' (HH 39). Nietzsche writes: 'He who has fully grasped the theory of total unaccountability can no longer accommodate the so-called justice that punishes and rewards . . . . For he who is punished does not deserve the punishment . . . for he could not have acted otherwise' (HH 105).

These Nietzschean ideas are familiar. Recently, however, scholars have suggested that despite Nietzsche's critique of 'desert free will', one can nevertheless attribute to Nietzsche a conception of 'agency free will', according to which an agent's freedom is an ideal to aspire to and involves self-mastery, unity among one's drives, and

3 For detailed discussion of Nietzsche's middle period see Abbey (2000) and Franco (2011). While these two texts are invaluable resources for the study of Nietzsche's middle period, they do not provide detailed analysis of the argument I will focus on from WS 23. Peter Sedgwick (2013: 112–13) and François Raffoul (2010: 113–14) discuss this passage but do not analyze it closely or interpret it as a criticism of compatibilism, as I do here.

4 It is important to emphasize that in the later stage of his thought (BGE 21) Nietzsche argued that his denial of free will in the metaphysical sense does not commit him to the opposite view according to which the will is *un*-free. I will not examine this claim here.

https://doi.org/10.1017/apa.2016.20 Published online by Cambridge University Press

self-discipline (Gemes 2009). First, agency free will is understood as *compatible* with the lack of metaphysical free will Nietzsche adheres to in his middle-period, and second, it does not ground our moral practices of blame and punishment (for a related view and an account of the feeling of freedom as based on a sense of self-efficacy in overcoming resistance, see Dries [2015]).

This raises the question of whether it is possible to defend against Nietzsche's critique of moral responsibility and desert precisely by invoking the *compatibilist* perspective, according to which determinism and lack of free will in the metaphysical sense are compatible with moral responsibility and the related practices of blame and punishment. If such a strategy can succeed, then Nietzsche's critiques could perhaps give cause for alarm among the libertarians, but they do not pose a threat for the compatibilist.

In this paper I argue that this is not the case by interpreting a relatively neglected section in *The Wanderer and His Shadow*—section 23 (WS 23)—as an anticompatibilist argument. On the reading I propose, the argument clearly manifests Nietzsche's hitherto unappreciated sensitivities to considerations that compatibilists typically appeal to and argues for the claim that punishment could not be morally defended by compatibilism's own light. The argument thus shows that Nietzsche's thought poses a challenge to compatibilism as well.[5]

Since compatibilism comes in different forms, it is necessary to narrow down the construal of Nietzsche's target by focusing on one specific version of compatibilism. I will therefore concentrate on and succinctly summarize one prominent view—that of R. Jay Wallace as expounded in his 1994 *Responsibility and the Moral Sentiments*. Wallace's view elegantly analyzes compatibilism using distinctions and terms that nicely map on to those that Nietzsche himself could be seen to be appealing to in his argument. Wallace's work is thus especially fruitful in the attempt to understand Nietzsche's critique of compatibilism.

In section 1 I summarize Wallace's compatibilist position. This will provide the requisite background for the analysis of the argument in WS, which I examine in detail in section 2. In section 3 I consider four possible objections to Nietzsche's argument and offer replies on Nietzsche's behalf.

1. R. Jay Wallace's Compatibilism

One of compatibilism's attractions and advantages lies in that it strives to maintain proximity to, and to elucidate conceptually, our everyday moral practices without having recourse to any substantial metaphysical assumptions. Accordingly, Wallace

5 Ken Gemes (2009) draws attention to this passage from WS, but does not analyze the argument Nietzsche propounds there. Leiter (2007) attributes to Nietzsche an argument that seems to have anticompatibilist implications. The argument is that since for Nietzsche conscious willing is epiphenomenal and since 'on *any* account of free will and moral responsibility, the will must be causal' (Leiter 2007: 11, emphasis added), it follows that compatibilism cannot defend the idea of moral responsibility and (consequently) morally justified punishment. I don't think, however, that compatibilism is committed to the view that the 'will' must necessarily be conscious for ascriptions of moral responsibility to be justified. Compatibilism is thus immune to this specific criticism.

https://doi.org/10.1017/apa.2016.20 Published online by Cambridge University Press

holds that moral responsibility is not a metaphysical fact to be ascertained from the theoretical point of view, a fact that obtains independently of our moral practices (Wallace 1994: 88). Rather, the concept of moral responsibility should be analyzed precisely *from within* the perspective of our moral practices, for it is these, guided as they are by our moral norms, that determine when and on what grounds one could be justifiably taken as morally responsible for one's actions. More specifically, a person is justifiably deemed morally responsible when we correctly judge that it would be justified to hold him or her so. In other words, moral responsibility is an essentially normative concept.

When would it be justified, then, to hold a person morally accountable? Wallace claims that for a person to be considered a morally responsible agent in general that person has to possess what he calls *reflective self-control* (this encompasses what Wallace calls 'accountability conditions' or A-conditions). When a person possesses the powers of reflective self-control it would be 'fair to demand that [he] comply with moral obligations we accept' (1994: 157). And when such a person fails to live up to our demands, it would be appropriate to resent that person, feel indignant, or impose various kinds of moral sanctions, such as moral blame, 'avoidance, reproach, scolding, denunciation, remonstration and (at the limit) punishment' (1994: 54)—sanctions that express our reactive attitudes. Of course, whether any one of these sanctions would be appropriate and what severity would be deemed proportional is to be determined with reference to the specificities of the case. The point is that when a morally responsible person performs a morally impermissible action, he meets—in virtue of being justifiably considered morally responsible—the necessary conditions for being rendered susceptible to such sanctions.

Reflective self-control involves capacities of the following two kinds: (1) 'the power to grasp and apply moral reasons, and (2) the power to control or regulate [one's] behavior by the light of such reasons' (1994: 157). In effect, this implies the following interconnected set of powers (1994: 157–59). Moral responsibility involves the understanding of general moral principles. This involves not only a cognitive grasp of the concepts these principles are concerned with but also the ability to be moved by them, that is, the affective capacity to care about morality and its demands: one has to be motivationally responsive to such principles and the reasons they engender. Further, one has to be able to apply the general principle to the concrete case in which one acts—that is, one has to possess the power of judgment in virtue of which the particular could be subsumed under the general. Next, one must be able to deliberate on the basis of the relevant moral principles and the specific reasons they provide in the specific circumstance as well as in light of other salient features. This involves the ability of critical reflection: 'one must be able to stand back from one's immediate desires and assess the actions they incline one to perform' (1994: 158). In addition, moral responsibility requires the capacity to make choices in light of such critical moral reflection. A person who lacks the capacity to choose on the basis of moral reflection will 'lack the ability to control what she does *by* grasping moral reasons' (ibid.)—and would thus not fully count as a person who enjoys reflective self-control. Lastly, reflective self-control requires the ability to translate one's choices into action—persons who lack this

https://doi.org/10.1017/apa.2016.20 Published online by Cambridge University Press

capacity would not be able to express their choices in action and would thus not be considered to possess full reflective self-control.

Wallace clarifies (1994: 160) that these capacities come in degrees but that nevertheless a certain degree of reflective self-control is necessary for moral responsibility: when a person lacks such self-control 'it would be unfair to hold [that person] to moral obligations one accepts' (1994: 161). Why? Because, Wallace explains, it would be '*unreasonable* to hold the person to moral obligations under these conditions. To make this proposal is, in effect, to postulate a moral principle of reasonableness, namely, that it is unreasonable to demand that people do something . . . if they lack the general power to grasp and comply with the reasons that support the demand' (ibid.). By analogy, 'it would seem unreasonable to demand that a child should be a star athlete if the child lacks the basic physical talents that are necessary to excel in school sports' (ibid.) In other words, the important point here is that it would be unreasonable to demand of people that they do something they just cannot do; thus, it would be unreasonable to blame people (or morally condemn or punish, etc.) for an immoral action they performed if they lacked the capacities that would enable them to do what is right.

When would it then *not* be appropriate to regard a person as morally responsible for his or her actions? Wallace explains that this would be the case when certain conditions obtain, conditions he labels *exemptions* or *excuses*. Let's start with exemptions. When these obtain, Wallace explains (1994: ch. 6), then it would be unjustified to regard the person as morally accountable in general, and, consequently, it would be unjustified to resent that person for specific wrongdoings and to blame or expose him or her to any moral sanction. Examples of exemptions include: (1) immaturity (the case of children); (2) mental illnesses (e.g., psychopathy); (3) action under hypnosis; (4) extreme mental stress; (5) addiction (which undermines at least to some extent one's moral responsibility). In all these different cases, the agent lacks, to some degree or other, one or more of the components that constitute reflective self-control. Notice that such exemptions fall roughly into two camps: chronic or long-standing conditions (cases 1 and 2, for example) and short-term conditions (cases 3 and 4).

Let's turn now to excuses (Wallace 1994: ch. 5). When these obtain, then even though a person might be justifiably held to be a morally accountable agent in general, it would be wrong to hold that person responsible for a particular action performed under these conditions. Put differently, when excuses obtain, blameworthiness conditions (B-conditions) do not apply, and the person would not be appropriately held responsible for a specific moral transgression. Examples of excuses include: (1) not knowing what one is doing, thus not intending to do what one did; (2) acting inadvertently or by mistake; (3) acting under coercion.

When neither excuses nor exemptions obtain, we would be justified in holding persons morally responsible for their actions and exposing them to the various moral sanctions mentioned. As we can see, according to this compatibilist view, holding others responsible and subjecting them to our moral sanctions does not in any way presuppose the falsity (or truth) of determinism: our moral practices can be rendered intelligible and reasonable without settling such metaphysical issues one way or the other.

https://doi.org/10.1017/apa.2016.20 Published online by Cambridge University Press

2. Nietzsche's Critique of Compatibilism and the Explanatory Challenge

I now turn to examine WS 23—Nietzsche's criticism of compatibilism; in this discussion I will use the term 'compatibilism' without any qualification, but it is important to keep in mind that the version of compatibilism presupposed here is Wallace's. The argument, on my reading, aims to show that given the compatibilist's view of moral responsibility and blameworthiness, no person ever deserves to be punished. Let me first frame my interpretation of the passage with the following qualifications. First, Nietzsche of course does not use the label 'compatibilism'—nor does he employ any of the other familiar technical terms. Nevertheless, as I show below, his argument could be reconstructed as addressing precisely the features of accountable action appealed to by a compatibilist of the kind presented above. Second, though Nietzsche is discussing here 'people who judge and punish as a profession' (WS 23), given that the position attributed to them is manifestly similar to the compatibilist view outlined above, I will interpret Nietzsche as presenting an argument directed at our general practices of holding others responsible, practices compatibilist thought appeals to. This should not be a surprising move given that 'the criminal law reflects central assumptions about moral responsibility, and the two concepts of responsibility [i.e., the legal and compatibilist] have very similar structure' (Brink and Nelkin 2013: 284). Accordingly, references to the law in the argument will be read—and they indeed can be easily so read—as references to the *moral* law, that is, to moral principles. As will become apparent, these modifications do nothing to alter Nietzsche's argument and actually render it more philosophically interesting than when read as targeting merely the legal practices of judges. Further, though Nietzsche is discussing the right to punish, it should be clear that this discussion immediately bears also on the appropriateness of the other kinds of moral responses mentioned above as well, those of blame, condemnation, denunciation, etc. Finally, it is important to emphasize that Nietzsche is *not* presenting us with his own theory of action or his own conception of moral responsibility in this passage: he presents us with what he takes to be a prevalent conception among those who 'judge and punish'—a conception that, as I will argue, is essentially compatibilistic.

The section is entitled 'Have the adherents of the theory of free will the right to punish?'. This can give the impression that the section is addressed to all believers in free will and is not specifically concerned with compatibilism's understanding of free will. I suggest, however, that a most fruitful way to read WS 23 is as specifically targeting compatibilism (without mentioning it by name, of course), for it addresses the distinctions and criteria employed by compatibilism. Specifically, Nietzsche argues in this section that no one deserves to be punished *not* because the libertarian idea of free will is an error of some sort (in contrast to the more familiar Nietzschean line of argument I discussed above in the introduction), but because no wrongdoing agent—metaphysically free or not—could be intelligibly seen as meeting the *compatibilist's* criteria of punishability. The section can be divided into two main parts. In the first part, Nietzsche could be read as arguing against compatibilistic justifications for punishment, and in the second, Nietzsche shows how a last-ditch attempt by the compatibilist to justify punishment appeals

https://doi.org/10.1017/apa.2016.20 Published online by Cambridge University Press

to a notion of metaphysical free will. This attempt fails as well, but not because of difficulties that beset the notion of metaphysical free will as such; rather, the problem is that the assumption of metaphysical free will cannot help to meet the *compatibilist's* own criteria of punishability. I will now quote the first part of the section:

> **Have the adherents of the theory of free-will the right to punish** [*strafen dürfen*]?—People who judge and punish as a profession try to establish in each case whether an ill-doer is at all accountable [*verantwortlich*] for his deed, whether he was [1] *able* to employ his reason [*Vernunft*], [2] whether he acted for *reasons* [*aus Gründen handelte*] and [3] not unknowingly [*nicht unbewußt*] or [4] under coercion [*Zwange*]. If he is punished, he is punished for having preferred the worse reasons for the better [*die schlechteren Gründe den besseren vorzog*]: which he must therefore have [5] *known*. Where this knowledge is lacking, a man is, according to the prevailing view, unfree and not responsible [*unfrei und nicht verantwortlich*]: except if his lack of knowledge, his *ignoratia legis* [*ignorance of the law*], for example, is a result of an intentional neglect to learn; in which case, when he failed to learn what he should have learned he had already preferred the worse reasons to the better and must now suffer the consequences of his bad choice. If, on the other hand, he did not see the better reasons, perhaps [6] from dull-wittedness or [7] weakness of mind [*Stumpf- und Blödsinn*], it is not usual to punish him: he lacked, one says, the capacity [8] to choose [*Wahl*], he [9] acted as an animal would. For an offence to be punishable presupposes that its perpetrator intentionally denied the better dictates of his reason [*die absichtliche Verleugnung der besseren Vernunft*]. **But how can anyone intentionally be less reasonable/rational than he has to be?** [*Wie kann aber Jemand absichtlich unvernünftiger sein, als er sein muss?*] Whence comes the decision when the scales are weighted with good and bad motives? Not from error, from blindness, not from an external nor from an internal compulsion? (Consider, moreover, that every so-called 'external compulsion' is nothing more than the internal compulsion of fear and pain). Whence? One asks again and again. *Reason* [*die Vernunft*] should not be the cause, because it could not decide against the better reasons? (WS 23, translation slightly modified, numbering and emphasis in bold added)

Given the outline of the compatibilist view offered above, I believe it is not hard to see how Nietzsche's argument directly addresses the main concepts and distinctions this view employs and could thus be read as targeting it. Let us look closer.

First, Nietzsche can be seen here as positing, in the name of the view he is addressing, conditions for moral accountability strikingly similar to the ones encountered above under the concept of reflective self-control. I will now track this similarity with reference to the passage just cited, with the numbers in the following

https://doi.org/10.1017/apa.2016.20 Published online by Cambridge University Press

corresponding to those added to the passage. According to the view Nietzsche is presenting, one is morally responsible when (1) one has the power to employ his or her reason (*Vernunft*) and thus (5) knows what moral principles there are and which ones are relevant. More precisely, the responsible individual can tell which (2) specific reasons (*Gründe*) are applicable to the case at hand (which involves the power of judgment.) Further, the individual must be able (8) to choose on the basis of these reasons, that is, to deliberate morally based on recognition of the relevant reasons and decide which course of action to pursue. If this ability to choose upon reflection is lacking, the agent's choices would not be choices at all; they would not be expressive of the agent's reflective self-control (to use Wallace's terminology), but would be merely the 'choices' [9] of an animal, not of a morally responsible person. Finally, one must possess the ability (2) to translate one's choices into action, and again the action must be carried out not merely in accordance with one's reason but, as Nietzsche puts it, *for* those reasons.

Next, Nietzsche could be seen here as referencing what Wallace calls exemptions and excuses. When exemptions obtain, recall, we are justified in not seeing the individual as a morally responsible agent in general. Nietzsche's first example of this is what he calls (6) dull-wittedness (*Stumpfsinn*), which could be read as involving a kind of apathetic mindset that, in this context, might involve *affective* indifference to the force of moral reasons. The second case Nietzsche gives is that of (7) weakness of mind (not to be confused with weakness of will) or imbecility (*Blödsinn*), which seems to suggest some *cognitive* lack in one's ability to grasp moral reasons at all or in one's ability to 'see the better reasons' as they apply to the particular case. Next, excuses, when they obtain, remove an individual's responsibility for a specific action. The examples Nietzsche gives here are similar to the cases Wallace considers: we have the excuse of (3) not knowing what one is doing, thus not acting intentionally, as well as that of (4) acting under coercion.

Now, Nietzsche claims, if it turns out that any of these excuses or exemptions obtain, then, according to the compatibilist view under examination, it would be inappropriate to hold the person accountable, and so one should refrain from punishment (or from other moral sanctions, we can add). Nietzsche's presentation of the compatibilist view of those who 'judge and punish' thus corresponds closely to Wallace's account of the conditions for moral accountability and to his account of exemptions and excuses. Crucially, the position Nietzsche presents here ascribes freedom and responsibility to an agent not in virtue of a capacity to will freely in some metaphysical sense or in virtue of an agent's access to alternative courses of action; rather, freedom and responsibility require meeting conditions that obtain, or not, independent of the truth of any metaphysical thesis about the will.

So when *would* the ill-doer be justifiability held responsible and seen as punishable according to Nietzsche's compatibilist? Nietzsche explains: 'If he is punished, he is punished for having preferred the worse reasons for the better' (WS 23). The other formulation Nietzsche gives is this: 'For an offence to be punishable presupposes that its perpetrator intentionally renounced the better dictates of his reason' (ibid.). These formulations complement each other, for the agent described here is one who in intentionally preferring the worse reason for the better is

https://doi.org/10.1017/apa.2016.20 Published online by Cambridge University Press

intentionally denying or renouncing those better reasons. Here, in articulating what he takes to be the compatibilist's criterion of punishability, Nietzsche is expressing one of the central premises of his argument on which the criticism of compatibilism I ascribe to him turns. I will now turn to reconstructing this argument.

Nietzsche's reasoning can be reconstructed as follows. For punishment to be appropriate the ill-doer must justifiably be held morally responsible. When excuses or exemptions obtain (to use Wallace's terms), we should not hold the wrongdoer responsible for the particular action performed. But this means that we should hold offenders accountable only when they were in their right mind, not coerced in any way, when they acted intentionally and out of full knowledge of what morality requires in the particular case. But then, Nietzsche asks, how can we explain the agent's wrong action? *Why* did the agent act immorally? The only possible answer left, Nietzsche holds, is that reason 'is the cause', that is, that the agent rationally and intentionally preferred to be irrational and act on his or her worse reasons. I believe the string of rhetorical questions at the end of the passage as well as the bewildered expression '*Reason* [*die Vernunft*] should not be the cause because it could not decide against the better reasons?' is indicative of Nietzsche's view that such a predicament is indeed radically unintelligible. But if this predicament is one that is absolutely unintelligible, it therefore should be considered impossible. Consequently, there is no reason to think there exists a situation where an agent is both morally responsible and does the morally impermissible thing. Hence, there is no situation where punishment is morally justified.

As we saw in the quotation from WS 23 above, Nietzsche ends his argument with a string of rhetorical questions: 'But how can anyone intentionally be less reasonable than he has to be? Whence comes the decision when the scales are weighted with good and bad motives? . . . Whence? One asks again and again' (WS 23). By asking these questions, Nietzsche is raising what I will call the *explanatory challenge*: the challenge is to explain the performance of an immoral action without appealing to conditions that would exempt or excuse the wrongdoer or otherwise relieve the wrongdoer from responsibility and would thus make punishing the wrongdoer morally unjustified. Nietzsche's implicit view here is, of course, that the challenge cannot be met and that consequently, as the argument reconstructed above aims to show, there is no moral justification for punishment (and by extension, for all other moral sanctions).

Before examining a number of objections to this argument and attempts to answer the explanatory challenge (in the next section of the paper), we have to understand the argument's nature better. Specifically, we should ask: (1) why should the compatibilist rise to the explanatory challenge at all; and (2) why is it unintelligible that an agent rationally and intentionally prefers the worse reasons to the better ones? I will address the questions in this order.

The reason the compatibilist should worry about the explanatory challenge is that if Nietzsche's argument is right, then the only scenario in which agents could be justifiably blamed and punished *by the compatibilist's own lights* is when agents rationally and intentionally act against what they themselves took

https://doi.org/10.1017/apa.2016.20 Published online by Cambridge University Press

to be their better reasons. But if it turns out that this scenario cannot be made intelligible and should therefore be seen as impossible, it would follow for the compatibilist that the only situation in which blame and punishment are claimed to be deserved should be seen as impossible. This, however, would entail that we have reason to think that there is no situation in which punishment is morally justified. Consequently, if compatibilism is to make room in its theory for the appropriateness of moral sanctions, it must confront Nietzsche's challenge and explain how an agent can perform an immoral action and be held morally responsible for it.

Let me now turn to the second question. The reason why an agent rationally and intentionally preferring the worse reasons and acting on them is unintelligible is that this involves an extreme case of deliberate practical self-contradiction and self-denial. The agent Nietzsche has in mind here, remember, is an agent who possesses reflective self-control and is thus guided by her morally informed reason. This agent, by supposition, knows that it would be wrong to Φ and understands the reasons why, has the power to choose and perform not-Φ (or avoid Φ), and yet intentionally chooses to Φ. If we leave out exempting and excusing considerations (as well as other possibilities considered below, such as weakness of will), then we are left with the option that it is the agent's reason itself that generates reasons to Φ and ultimately leads the agent to prefer those worse reasons and do the wrong thing. And yet it is the very same reason that proclaims that this act should *not* be performed—that there are better reasons, i.e., moral reasons, not to do it. Thus, such an agent's reason can be construed as pronouncing *both* that there is more reason not to Φ (insofar as the agent possesses reflective self-control) *and* that there is more reason to Φ (insofar as the agent prefers the worse reasons to the better). This is already mysterious, for how can a properly functioning reason generate contradictory pronouncements such as these? Would this not be like a properly functioning computer giving two answers—one correct and one incorrect—to the same arithmetical problem? Hence, it seems that it can't be reason that is responsible for the agent's view that the worse reasons are better. (Of course, it can be a *failure* of reason that is responsible for such preference, but the question is whether such failure should not relieve the agent of moral responsibility. Nietzsche argues that it should.) To make things even more baffling, the agent, through the use of reason, is supposed then to proceed and intentionally deny reason and its better reasons and prefer the worse ones. Moreover, finally, for this agent, it is the worse reasons that prevail, for the agent ends up Φ-ing after all. Clearly, this is an extremely bewildering predicament, and insofar as compatibilism cannot make sense of it, it cannot render intelligible the one case in which, by its own light (according to Nietzsche's interpretation), the ill-doer can be rightly held responsible, blamed, and punished. It turns out that the conditions for moral responsibility cannot serve as an explanation for the performance of an immoral action. But if that is so, and if the explanation has to appeal to other factors, then responsibility for the misdeed cannot be pinned to the agent in his or her capacity as morally responsible.

It is at this point in the argument that Nietzsche introduces a surprising twist: faced with the utter unintelligibility of an agent rationally and intentionally

https://doi.org/10.1017/apa.2016.20 Published online by Cambridge University Press

preferring the worst reasons and acting against the better reasons, compatibilism 'calls "free-will" to [its] aid' and proclaims:

> It is *pure wilfulness* [*vollendete Belieben*] that is supposed to decide, an impulse is supposed to enter within which motive plays no part, in which the deed, arising out of nothing, occurs as a miracle. It is this supposed *wilfulness* [*Beliebigkeit*], in a case in which wilfulness ought not to reign, which is punished: reason, which knows law, prohibition, and command, ought to have permitted no choice and to have had the effect of compulsion and a higher power. Thus the offender is punished because he employs 'free will', that is to say, because he acted without a reason where he ought to have acted in accordance with reasons. Why did he do this? But it is precisely this question that can no longer even be *asked*: it was a deed without a 'for that reason', without motive, without origin, something purposeless and non-rational [*Vernunftloses*].—*But such a deed too ought*, in accordance with the first condition of all punishability laid down above, *not to be punished*! It is not as if something had *not* been done here, something omitted, reason had *not* been employed: for the omission is under all circumstances *unintentional*! And only the intentional omission to perform what the law commands counts as punishable. The offender certainly preferred the worse reasons for the better, but *without* reason or intention . . . . The presupposition that for an offence to be punishable its perpetrator must have intentionally denied his reason—it is precisely this presupposition that is annulled by the assumption of 'free-will'. (WS 23, translation slightly modified)

Nietzsche's claim here, I take it, is that once the compatibilist realizes that no sense can be made of the moral practices of blame and punishment, the temptation will arise, as a last-ditch attempt to find the agent responsible, to explain the wrongdoer's actions as arising out of free will in the *metaphysical sense*, the sense we encountered in the introduction. Specifically, the free will invoked here is a pure willing, constrained and directed by nothing, a pure *causa sui*.[6] It is a willing absolutely free from the guidance of reason (for if it were guided by reason this would take us back to the argument above) and yet one that does not *intentionally* aim at sidestepping reason, for if the agent intentionally and knowingly acted against reason, then either the agent suffered from lack of reflective self-control (and so should not be held responsible), or the agent's reason acted against itself (which would again lead us to the argument above). But the problem, as Nietzsche notes, is that such nonrational capriciousness should not be punished by the compatibilist's own criterion of punishability, for an agent acting out of such pure free will is

6 For some discussion of the idea of *causa sui* in BGE 21 see Leiter (2002: 88–91) and Leiter (2007, especially footnote 11). Ken Gemes (2009) recently noticed that the concept of free will invoked in WS 23 is criticized under the idea of *causa sui* in BGE 21. He rightly observes that this free will by itself is not sufficient for moral accountability precisely because of its arbitrariness (Gemes 2009: 40–41).

https://doi.org/10.1017/apa.2016.20 Published online by Cambridge University Press

not an agent who acts intentionally at all. Thus, even this desperate attempt by the compatibilist to appeal to the notion of metaphysical free will fails. But this is no surprise, as Nietzsche concludes the section, for the whole idea of free will and the compatibilist principles of punishment 'are at bottom nothing but a very peculiar conceptual mythology; and the hen that hatched it sat on her eggs in a place far removed from reality' (WS 23). Put differently, upon reflection, the whole conceptual nexus of accountability, punishment, blame, freedom, etc. does not stand up to critical scrutiny. I will now turn to considering four possible objections to the argument and offer replies on Nietzsche's behalf.

3. Objections and Replies on Nietzsche's Behalf

(1) It might be objected that it is quite easy to conceive of and explain agents acting against their better reasons, for such cases are familiar enough: they involve *weakness of the will* or *akrasia*. Akratic agents can be (noncontroversially, it is to be hoped) described as possessing moral knowledge and discernment and yet failing occasionally to act on their better reasons because of a sudden influence of some unruly passion that causes them to do the wrong thing. And yet we typically hold akratic agents morally responsible for their actions. (Thus, Fischer [2012: 129–31] criticizes Harry Frankfurt's approach to weakness of the will precisely on the grounds that it generates the counterintuitive conclusion that akratic agents are not morally responsible.)

Interestingly, Nietzsche, to the best of my knowledge, never discusses at length the phenomenon of weakness of the will in the above sense, and his usage of the term (e.g., BGE 212) typically refers to an agent's incapacity to commit to long-term and challenging tasks. But here in section 23 from WS Nietzsche seems to address the issue directly: he seems to be saying that every internal compulsion, such as the kind that figures in accounts of akrasia, should be excusable just as much as external ones, for the latter are 'nothing more than the internal compulsion of fear and pain' (WS 23). In other words, external compulsion, in its relevant aspects, could be reduced to internal compulsion, and since the former can serve as an excuse (as we saw in section 1 above), so should the latter. Can this idea hold water?

In his account of threats and their excusing force Wallace explains that acting under threats serves as an excusing consideration since it shows that the coerced agent did not really choose to perform action of kind *k* and thus violate our moral expectations, but rather chose to perform action-of-kind-*k*-*rather-than*-*y*, where *y* is some unwelcome consequence, such as one's own death or the death of one's loved ones (Wallace 1994: 143–45). This kind of choice, Wallace explains, is excusable for it does not violate our moral principles and obligations (at least in some cases). Can Nietzsche appeal to this kind of excuse in his remark on internal compulsion? This would be implausible, for the akratic agent typically does not express a choice of this kind: the akratic does not express the choice of eating-another-piece-of-cake-rather-than-suffering-some-horrible-fate; neither does the akratic express some other kind of choice we would find morally excusable. So this attempt to defend Nietzsche would not be successful.

https://doi.org/10.1017/apa.2016.20 Published online by Cambridge University Press

A different way to defend Nietzsche's argument is to recall the principle of reasonableness Wallace invokes when explaining accountability: according to this principle it would be unreasonable of us to expect of an agent to perform action *X* when the agent does not possess the general ability to do *X*. Now I wish to argue that in the case before us, where *X* stands for doing the right thing (or refraining from doing the wrong thing), the agent performs the wrong action. The agent does so not because she rationally chooses not to do *X*, but rather because she is unable to do *X* (if she can, why doesn't she?), and she is unable to do *X* because she is unable to translate her rational choice into action, and she cannot do that because she is deprived of her reflective self-control by her overpowering passions. Hence, it would be unreasonable to hold her responsible for her wrong action and unreasonable to blame or punish her.

It will be objected at this point that this completely elides the important distinction between chronic or general incapacity (like that of the psychopath) and local, episodic failure—a distinction that bears a lot of normative weight in our assessment of moral responsibility given that it is our practice to exempt only agents suffering from the former from the burdens of accountability. But this means that the akratic agent should be considered responsible and that he or she consequently could justifiably be seen as liable to moral sanction.

In response we should be aware, first, that even relatively temporary conditions, such as stress or distress (Wallace gives the examples of 'the loss of one's job or a sudden death of a family member' [1994: 179]), could be regarded as exempting a person from responsibility, and Wallace acknowledges this explicitly (Wallace 1994: 179–80). Second—and this is the more crucial point—recall the reason it is unreasonable to hold a person responsible according to Wallace: it is unreasonable to demand of someone to act if he or she does not possess the requisite *general* capacity to act. In contrast, I wish to argue here that the reason why it is unreasonable to demand of people to do *X* when they can't do *X* is *not* that they suffer from a general or chronic debilitating condition but that they suffer from a debilitating condition, period. It is the condition that prevents people from acting morally, not its chronicity. The chronicity only strengthens our assurance that someone indeed lacks the capacity requisite to meet our moral expectations, but it in itself is not the morally relevant feature of the agent. It is rather the capacity that is morally relevant, a capacity the akratic *lacks* (at the time of action).

We might still insist that even though the akratic is not wholly responsible for the wrong actions, the akratic is more responsible than the person who suffers from some mental disorder or acts under threats. We might feel that although the akratic was, in a way, compelled to act in this way, an akratic still deserves to be rebuked or punished *at least to some extent*. After all, the typical akratic agent seems to indulge in the act and to approve of it (or at least seems to at the moment of action). But this seems to be all that the compatibilist needs, for wouldn't it then be justifiable to punish the agent at least to some degree? The problem, however, with the view that a partially responsible agent deserves to be partially punished is that it presupposes that an agent who is *fully* responsible deserves to be punished without any amelioration. But it is precisely this presupposition that Nietzsche's argument calls into question, for he thinks it is precisely this the compatibilist

https://doi.org/10.1017/apa.2016.20 Published online by Cambridge University Press

fails to show. After all, the upshot of Nietzsche's argument is that punishment cannot be morally justified even in the best of cases where we are dealing with a paradigmatically responsible agent, and, a fortiori, it can be justified much less when an incapacitated agent is at issue.

The topic of akrasia is of course extremely complex, and the sketch of an argument given above is not meant as a detailed defense of akratic moral irresponsibility, but it is consonant with Nietzsche's general position and the specific views expressed in the passage and enjoys sufficient plausibility to be considered seriously. Note, finally, that even if the Nietzschean defense of akratic action given above remains ultimately unsatisfying, it is open for Nietzsche to reply by saying that surely it is not *only* the akratic individual whom we tend to hold accountable: the scope of our ascriptions of moral accountability, with its attendant practices of blame and punishment, is far wider and requires justification. Clearly, not every wrongdoer who is held responsible is an akratic (the cool-headed embezzler, for example), so how could we explain the remaining cases that form (perhaps) the large majority? Nietzsche's argument thus raises a considerable challenge to a large portion of our moral practices.

(2) It could be objected that Nietzsche's explanatory challenge is a spurious one: he asks us to explain the morally wrong actions of a perfectly rational agent, but this is clearly impossible by definition. In contrast, the moral actions of an *imperfect* agent are clearly intelligible, precisely in virtue of the agent's imperfection. Indeed, most of us are somewhere on the continuum—neither angels nor monsters—and yet we hold ourselves and others morally accountable. (The individual I have in mind here in this objection is someone we can think of as possessing only a moderate reason-responsive mechanism, of the kind Fischer and Ravizza discuss [1998]).

Let us consider this objection. For example, if an agent steals a book, then we can explain this in terms of the agent's failing to see the reasons (general or particular) against stealing the book or failing to be properly affected by them at the moment of action or failing to deliberate correctly on their basis or failing to translate the choice made into action. But then, regardless of which of these obtains, it would follow that the agent at the moment of action did not possess sufficient reflective self-control to perform the right action (and avoid doing what is wrong). But then it would be *unreasonable* of us to expect the agent to do what is morally right in that instance (not steal) since, as I argued above, it is unreasonable to expect people to act in a way that requires the possession of capacities they lack at the time of action. But then it would be unreasonable to hold the agent morally responsible for that action.

(3) We might object that Nietzsche's compatibilist is implicitly assuming that moral accountability requires *actual exercise* of the general capacities for reflective self-control, but that this assumption is false. Differently put, we might say that wrongdoers should be held morally responsible even if they do not actually choose as a result of their capacities springing into action—what matters is that the agent *possesses* these general reflective capacities, *not* that these are exercised in fact. As Wallace puts it: 'To require actual exercise of these abilities as a condition of responsibility would therefore rule out responsibility in cases in which people do things that are morally wrong' (Wallace 1994: 190). In other words, were

https://doi.org/10.1017/apa.2016.20 Published online by Cambridge University Press

the actual exercise of reflective self-control—which issues in morally permissible action—a necessary condition for moral responsibility, then we would have no reason to hold moral wrongdoers accountable for their actions, for their acting immorally would attest to their *not* exercising these abilities, and this would undermine responsibility. Consequently, actual exercise is not required for justified ascriptions of moral responsibility—*possession* of reflective self-control is, however, sufficient (and necessary). This could also allow us to justify our ascription of moral responsibility to the book thief above: the thief *did* in fact possess full reflective self-control; it just was not exercised. This approach thus meets Nietzsche's challenge, for here we have an intelligible case of a person who is morally accountable (the person possesses reflective self-control) but nevertheless does the wrong thing (the person does not exercise reflective self-control at the moment of action).

Nietzsche, I think, would continue to press the compatibilist with his explanatory challenge: if the agent really possessed these general capacities of reflective self-control, how can it be explained that they failed to spring into action in this particular instance and exactly at the time when they were most needed? Arguably, the whole point of possessing such capacities is that they *do* spring into action and get exercised when the proper moment arrives. Either there was some momentary collapse in their proper function or not. If the former, then it would seem unreasonable to hold the person responsible given that at the time of action this person lacked the powers to do what is expected of him or her. If the latter, then, again, why didn't the capacities function properly?

(4) Finally, one might object that the situation Nietzsche has in mind in his challenge is quite intelligible: an agent can know morality's demands, recognize their force and relevance to the particular occasion, and yet intentionally choose to act for a different reason. And the agent may so choose because he or she does not think that moral principles always issue in stronger reasons for action and because he or she believes that the case at issue is precisely of such a kind where amoral or immoral considerations outweigh the demands of morality. In other words, the objection maintains that Nietzsche's criticism (as I construed it here) presupposes the Kantian view that to act against morality's dictates is irrational or, more weakly, that the reasons provided by morality are always stronger than any other reasons, that there is always more reason to act in accordance with morality. This presupposition, however, can be challenged precisely by someone with a Nietzschean view who holds that great individuals or burgeoning higher spirits should not be swayed by the temptations of morality and lose their way but should rather hold fast to their individual calling that provides them with stronger reasons than those of morality. For them, the danger at issue is that rather 'than tolerate (even welcome) suffering, the [higher spirit] will seek relief from hardship... rather than practice "severe self-love"... [the higher spirit] will embrace the ideology of altruism and reject "self-love" as improper' (Leiter 2002: 299, see GS 338). Thus, we can meet Nietzsche's challenge: a person may possess and exercise the capacities for reflective self-control (which makes the person responsible) and yet rationally decide to act immorally (given that he or she rationally thinks there are better reasons for action than the moral ones).

https://doi.org/10.1017/apa.2016.20 Published online by Cambridge University Press

In response it should be realized that this way out of Nietzsche's criticism is in itself quite radical and involves the abandonment of a central view of moral thought according to which morality's reasons are always overriding and enjoy unconditional authority (for a critique of morality in general and of this particular feature of this 'strange institution' see Williams [1984]). To act immorally *is* to act irrationally (at least to some extent), and so to knowingly and intentionally go against morality's dictates would indeed amount to intentional irrationality, something that, Nietzsche argues, it is hard to make sense of.

Of course, we can bite the bullet and, like Nietzsche, hold that this view of morality should itself be subject to criticism and be exposed for what it is—a prejudice. In other words, we might hold that in *some cases* agents have *more reason* to act on immoral or amoral considerations and that such a case is the one before us. But then on what grounds would we *morally* blame (punish, etc.) a wrongdoer? If we think that that person did the rational thing, that in this situation it was indeed more rational for that person, say, to look after himself or herself rather than bother with morality, then blaming this wrongdoer would not make sense. This person will rightly be held by us as accountable—meeting all the criteria of accountability—but would not be justifiably held as deserving blame or sanction of whatever sort. Indeed, it would not even be appropriate for us to resent that person, given that insofar as we think that he had good reason to act as he did, we ourselves would have done the same had we been in his shoes. Nietzsche can thus conclude that although in this case we would be justified in holding the other person responsible for his actions, we would *not* be justified in blaming or punishing him.

4. Conclusion

In this paper I have reconstructed and examined Nietzsche's argument in WS 23. According to my interpretation, this argument could be read as an anticompatibilist argument that seeks to show that by the compatibilist's own lights there is no moral justification for punishment (and for our moral sanctions in general). In addition, I have claimed Nietzsche can be seen as raising in WS 23 what I called the *explanatory challenge*, which challenges the compatibilist to explain the performance of an immoral action without appealing to conditions that would exempt or excuse the wrongdoer or otherwise relieve the wrongdoer from responsibility and would thus make punishment morally unjustified. The argument, as reconstructed here, by bringing to light Nietzsche's hitherto unappreciated sensitivities to compatibilist considerations regarding moral responsibility, reveals the breadth and depth—as well as the relevance—of Nietzsche's thinking about these issues.

GUY ELGAT
SCHOOL OF THE ART INSTITUTE OF CHICAGO
Guyelgat2011@u.northwestern.edu

https://doi.org/10.1017/apa.2016.20 Published online by Cambridge University Press

References

LIST OF ABBREVIATIONS OF NIETZSCHE'S WORKS

BGE: *Beyond Good and Evil*. Translated by Walter Kaufmann (New York: Vintage), 1966.
GS: *Gay Science*. Translated by Walter Kaufmann (New York: Vintage), 1974.
HH: *Human, All Too Human*. Translated by John R. Hollingdale (Cambridge, UK: Cambridge University Press), 1986.
WS: *The Wanderer and His Shadow*. Translated by John R. Hollingdale. In *Human all too Human* (Cambridge, UK: Cambridge University Press), 1986.
KSA: *Kritische Studienausgabe*. Compiled under the general editorship of G. Colli and M. Montinari (Berlin and New York: Walter de Gruyter), 1999.

SECONDARY LITERATURE

Abbey, R. (2000) *Nietzsche's Middle Period*. Oxford: Oxford University Press.
Brink, D. O., and Nelkin, D. K. (2013) 'Fairness and the Architecture of Responsibility'. In D. Shoemaker (ed.), *Oxford Studies in Agency and Responsibility*, *Volume 1* (Oxford: Oxford University Press), 284–314.
Dries, M. (2015) 'Freedom, Resistance, Agency'. In M. Dries and P. J. E. Kail (eds.), *Nietzsche on Mind and Nature* (Oxford: Oxford University Press), 142–62.
Fischer, J. H. (2012) 'Semicompatibilism and Its Rivals'. *Journal of Ethics*, 16, 117–43.
Fischer, J. H., and S. J. Ravizza. (1998) *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge, UK: Cambridge University Press.
Franco, P. (2011) *Nietzsche's Enlightenment*. Chicago: University of Chicago Press.
Gemes, K. (2009) 'Nietzsche on Free Will, Autonomy and the Sovereign Individual'. In K. Gemes and S. May (eds.), *Nietzsche on Freedom and Autonomy* (New York: Oxford University Press), 33–50.
Leiter, B. (2002) *Nietzsche on Morality*. London: Routledge.
Leiter, B. (2007) 'Nietzsche's Theory of the Will'. *Philosophers' Imprint*, 7, 1–14.
Raffoul, F. (2010) *The Origins of Responsibility*. Bloomington, IN: Indiana University Press.
Sedgwick, P. (2013) *Nietzsche's Justice: Naturalism in Search of an Ethics*. Montreal & Kingston: McGill-Queen's University Press.
Wallace, R. J. (1994) *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.
Williams, B. (1984) *Ethics and the Limits of Philosophy*. Cambridge, MA: Harvard University Press.