Topoi (2024) 43:503–513
https://doi.org/10.1007/s11245-023-09993-4

Integrity as Incentive-Insensitivity: Moral Incapacity Means One can't be Bought

Etye Steinberg[1]

Accepted: 27 November 2023 / Published online: 12 February 2024

Abstract
This paper develops Bernard Williams's claim that moral incapacity – i.e., one's inability to consider an action as one that could be performed intentionally – 'is proof against reward'. It argues that we should re-construe the notion of moral incapacity in terms of self-identification with a project, commitment, value, etc. in a way that renders this project constitutive of one's self-identity. This consists in one's being insensitive to incentives to reconsider or get oneself to change one's identification with this project. More precisely, self-identification with a project implies that no state-given reason can justify for oneself reconsidering, or getting oneself to revise, or abandon one's identification with that project. This view ties together integrity and self-identification, and avoids problems common to competing views: it avoids regress problems faced by hierarchical theories of identification; it demonstrates that integrationist views of identification overlook the fact that a deep, well-integrated attitude may fail to be incentive-insensitive; and it helps explain what's wrong with 'perverse' cases, where one values acting in a way that one does not all things consider value. It also improves on Williams's own view, by construing moral incapacity not merely in terms of one's incapacity to perform an action (that undermines one's project and thus violates one's integrity), but also in terms of one's incapacity to reconsider one's commitment (to said project).

Keywords Moral incapacity · Integrity · State-given reasons · Self-Identification

1 Introduction

One of, if not the main recurring themes in Bernard Williams's work on morality, is that of personal integrity. It is because of threats to and incompatibility with such integrity that both utilitarianism (1973a) and deontology (1981a) fail as comprehensive moral theories. Our identities as persons rest on commitments and projects that give meaning to our lives (1973b). These projects and commitments are constitutive of our moral lives, of who we are as agents and reasoners. They generate certain 'practical necessities' (1981b), rendering some actions necessary for us, in the sense that not performing them will thereby violate our very identity as persons. The other side of this coin is that these commitments and projects render other actions impossible for us, in the sense that performing them intentionally goes against how we reason and deliberate. We are incapable of performing them intentionally, or of considering them as serious options in our practical deliberation. One's commitments and projects impose on oneself certain moral (or practical, or character) incapacities (1993). They do so through one's own deliberation. Acting against them is a violation of one's very self, according to Williams, in the sense that they are outside the limits of one's moral self. If one is truly committed to a project in a way that generates such moral incapacity, then, according to Williams, one should (or, more strongly, will) resist incentives to perform this action. As Williams puts it,

> moral incapacities are meant to be proof against rewards; and if an agent is not proof against rewards, then we may […] say that, after all, we were wrong to ascribe the incapacity to him. He is one who, after all, could act in that way, because, faced with that reward, he did do so (1993, p. 69).

But what could it mean exactly that moral incapacity is "proof against reward"?

This paper develops an answer to this question, by appealing to the distinction between object-given reasons

✉ Etye Steinberg
etye.steinberg@mail.huji.ac.il

[1] Department of Philosophy, The Hebrew University of Jerusalem, Mount Scopus, 9190501 Jerusalem, Israel

and state-given reasons. Object-given reasons for having an attitude are facts about the *object* of that attitude (say, the object of one's desire, or the content of one's belief). State-given reasons for having an attitude are facts about *having* the attitude, regardless of whether that attitude is warranted, or justified, or correct (say, the benefits of believing something, regardless of whether it is true or not, is a state-given reason for having a belief).[1]

The article construes moral or character incapacity as constituted by self-identification with a project, and argues that the sense in which moral incapacity is "proof against reward" is that self-identification is insensitive to state-given reasons, i.e., reasons that pertain to the fact (and benefits) of having these attitudes of self-identification. Thus, on the view this article develops, while it may be rational to get oneself to revise one's ordinary attitudes based on sufficient state-given reason (e.g., get oneself to believe that someone is kind in order to remain friends with a shared acquaintance), this is hardly ever the case for self-identification. The only state-given reasons that can justify reconsidering or getting oneself to revise one's identification with a project or a commitment (that generates moral incapacity) are reasons of coherence. If two or more projects contradict one another, then that – and only that – constitutes a state-given reason that can justify reconsidering and getting oneself to revise one's identification with (at least one of) these projects. In other words, identifying with a project means that it is irrational for one to reconsider identifying with it for any rewards. On this view, then, moral incapacity is proof against reward because the self-identification underlying it is incentive-insensitive.

In what follows, I flesh out the question of self-identification with a project and its importance for practical reason and agency. I then move on to argue for the incentive-insensitivity view of integrity. This view puts to novel use the distinction between object-given reasons and state-given reasons, based on a re-conception of Williams's notion of moral incapacity. I demonstrate the merits of this view, and how it avoids the pitfalls of competing views. I conclude by discussing some possible objections. Specifically, I discuss the possibility of having a state-given reason for reconsidering an identification attitude that is based on the benefits such a revision will incur on the object of this identification attitude (e.g., God promising one's children will live the best life possible, but only if one gets oneself to stop loving one's children). I argue that such cases are not problem cases: being committed to a project does not imply being higher-order committed to being lower-order committed to said project. Acting in ways that promote the projects to which one is committed does not undermine integrity; and failing to act in such ways merely to sustain one's commitment is straightforwardly fetishistic.

1.1 Williams on Integrity, Identity, and the Moral Self

Before moving on, it is important to emphasize that this is not an article about Williams's work, as such. The article does not presume to provide a detailed exegesis of Bernard Williams's work on integrity, self-identity, and moral incapacity, or the relation between these concepts. The article does not aim to explain Williams's own work, or to rationally reconstruct his views on these topics as such. Instead, it develops a view that is inspired by and based on Williams's work, and can be seen in this sense as Williams-ian. But it is not meant to portray Williams's view accurately.

For example, the article will use the term 'self-identification' as synonymous with Williams's notions of both integrity and personal or self-identity (all of which have different meanings and roles in Williams's own work). The term 'self-identification' is not one that clearly follows from Williams's work. For Williams, personal identity is a term of art appealing to the commitments and projects that give meaning to our lives, and based on which we can show that consequentialist and deontological moral theories fail to leave room for the moral lives of persons (1981a). Integrity, on the other hand, is put forward to argue that utilitarianism is a moral theory in which practical deliberation occurs "outside one's moral self", so to speak (1973a). For Williams, the aim of introducing personal identity (understood in terms of social, moral, emotional or other projects to which an agent is committed) as having weight in practical deliberation, is to relocate the moral point of view: to move it from a universal one that is external to the agent to a particular, even personal point of view that is internal to oneself.

This article also does not engage with the implications of Williams's work on moral theory. Rather, it focuses on the moral-psychological picture that Williams's work can inspire or motivate. In this respect, it is important to note that Williams regards moral incapacity as a case in which performing an action would violate one's integrity, and therefore, one's moral- or self- identity. According to Williams, one's self-identity (i.e., the total set of one's projects and commitments) sets the conditions and limits for practical deliberation that is internal to oneself. Thus, crossing these limits–and reasoning 'from without' oneself–constitutes a violation of one's integrity. Williams's point is that what

[1] Parfit (2001) introduces this distinction. Piller offers a different version of the distinction, focusing on content-related and attitude-related reasons (2006). Hieronymi also draws the distinction differently, based on her understanding of a reason as a consideration that bears on a question (2005, 2011). Thus, commitment-constitutive (i.e., object-given) reasons bear on whether *to have* an attitude, and extrinsic (i.e., state-given) reasons bear on whether *to take the necessary steps* for having an attitude (2005, 2006).

makes an action fail to be integral to an agent's identity (by performing it or deliberating about it) is not the nature of the action, but the nature of the very agent as a person with an identity comprising certain projects and commitments. This is why Williams suggests that only a very malefic type of coercion could justifying the violation of one's moral incapacity (1993, p. 69). Williams stresses that these would be cases in which one would be doing what one rejects—and thus move against one's project or violate one's identity—and not deliberating over one's projects and commitments. Thus, if an agent acts (or deliberates) under coercion, then they are not acting or deliberating within the scope of their moral life. Their action is rather drawn from outside. So, to put it in Williams's words, it is no longer a matter of their character.

The article embraces much of Williams's thought and its spirit, but it does not claim to provide or be based on an accurate description of Williams's work, nor accept it at face value. First, the article argues that we should construe moral incapacity as self-identification with a project. This is neither Williams's view, nor does the article argue that this is how we should interpret Williams's work for Williams scholarship. In other words, the article does not argue that such an interpretation is consistent with what Williams says on this and related topics, or is the best rational reconstruction of Williams's work. Rather, the point of the article is that we should construe moral incapacity this way because then the notion of moral incapacity becomes very useful in other discussions regarding practical reason and agency (discussions which may not necessarily be within the scope of Williams's corpus). Moreover, what follows from the argument in this article is that Williams overlooks an important aspect of how moral incapacity is proof against reward. According to Williams, moral incapacity is proof against acting in ways that undermine one's projects or the things to which one is committed. On the view in this article, moral incapacity is also proof against *reconsidering* those very commitments. In what follows, any explanations of the notion of moral incapacity are meant to portray this re-conceptualization, rather than Williams's view.

2 Integrity and Identifying with Projects

Identification is an important aspect of our lives and agency. It's how we fit together different parts of our mental lives, and how we push away other parts. For example, identifying with a desire has been described as reflectively endorsing it in light of our practical identities (Korsgaard 1996, 2009), as valuing it in motivating one to act (Taylor 1976, 1985; Watson 1975), as caring about acting on it in a way that gives it room in one's will (Frankfurt 1988), as giving it a justificatory role in one's motivationally effective practical deliberation (Bratman 1996, 2007), and as being pleased with how well it causes our actions (Arpaly 2002; Arpaly and Schroeder 1999, 2013; Schroeder & Arpaly 1999). Identification with an attitude is how we constitute ourselves as the particular agents that we are (Korsgaard 2009), how we reconcile conflicting elements in our lives (Frankfurt 1987), and how we render some things more important to us than others (Bratman 2000; Frankfurt 1987; Taylor 1976).

Interestingly, such identification can sometimes apply to mundane desires, that do not play a significant role in one's life. Acting on these mundane desires (or not acting on them) is rather trivial in the overall scheme of things, insofar as one's identity as an agent goes. But sometimes we identify with 'big' things: one can identify with a desire to act honestly (McDowell 1979), or with a pacifist, anti-nuclear-proliferation project (Williams 1973a), or with pursuing studies in grad school, or law school (Chang 2013), or with non-Catholic faith and practices (Frankfurt 1982), or even with a desire to leave one's family behind and pursue a passion for painting (Williams 1981c). These instances of identification are much more important to our constitution of ourselves as ourselves. Going for the mille-feuille rather than the eclair (Taylor 1976) is (usually) hardly as constitutive of one's identity as choosing to become a lawyer over a physician.[2]

Projects and commitments with which one identifies in this deep, self-constitutive way must have deliberational stability: they must be resistant to reconsideration. Part of one's integrity as someone who self-identifies with a project is that one does not reconsider this identification. Of course, this need not be perfect stability. People change. Moreover, sometimes people change for good reasons. In other words, it is *sometimes* rational to change. But then, why is it not *always* rational to change? What kind of reasons could or couldn't justify reconsidering or revising one's identification with projects? How exactly should we understand this deliberational stability?[3]

Bernard Williams's work provides us with resources to construe identification with a project as consisting in one's being unable to act against this project while also maintaining integrity (1973a, b, 1981a, 1993). The total set of the projects and commitments with which we self-identify

[2] Alternatively, having a sugary snack while on a diet is not as detrimental to one's self identity and integrity as having that same snack while observing a religious fast.

[3] Below, I explore several views regarding these questions. I group these into hierarchical views (Bratman 1996, 2000, 2007; Frankfurt 1982, 1987, 1988; Korsgaard 1996, 2009), valuational views (Taylor 1976, 1985; Watson 1975, 1987), and integrationist views (Arpaly 2002; Arpaly and Schroeder 1999; Schroeder & Arpaly 1999). I demonstrate how the view does better than these views.

forms our personal deliberational standpoint. It sets limits to our reasoning, such that acting against them constitutes a form of self-betrayal, a violation of one's integrity. Cases of moral incapacity can thus be understood as cases where one finds oneself incapable of performing a certain action. This incapacity is not due to some extreme psychological aversion (e.g., fear or disgust). The pacifist chemist's incapacity to occupy a temporary position with a plant producing nuclear weapons (Williams 1973a) is not the result of her disgust of war;[4] rather, it is a direct consequence of her deep, self-constitutive commitment to the project of pacifism and nuclear non-proliferation. Her incapacity to intentionally work at a plant providing material for nuclear warfare is of her own doing, *through* her own deliberation and reasoning. Indeed, there may be a sense in which talk of moral incapacity as a result or consequence of reasoning and deliberation is misleading. Instead, we might say that one experiences moral incapacity to perform some action X not *because* of one's deliberation or reasoning, but rather *in* deliberation and reasoning. Integrity *consists* in being morally incapable of reasoning in a way that can justify performing X for oneself.

There is, however, a further aspect of integrity and moral incapacity, that Williams fails to consider. This is the deliberational stability of one's *self-identification* with a project. Being truly committed to a certain project seems to entail not only moral incapacity to perform actions that go against this project. Rather, it also entails a moral incapacity to reconsider one's commitment to this very project. For instance, if partners in a monogamous relationship begin to reconsider their commitment and faithfulness to one another, then there seems to be a sense in which they are thereby no longer really committed to the relationship. This, I argue, is how we should re-construe and use the notion of moral incapacity: if one is really committed to a project (or, in this case, a monogamous relationship), then one should not be preoccupied with whether to be committed to this project. One should not be able to reconsider one's commitment to this project, in the same way that one should not be able to act against this project.

[4] Of course, if one maintains that all moral judgments are grounded in emotional responses, then one may argue that there is no difference between judging that (nuclear) war is a moral atrocity and being extremely disgusted, or saddened, or frightened (or any combination thereof) by (nuclear) war. For the purposes of this article, we can put this option aside. At the very least, we can consider the difference between psychological incapacity (because of extreme disgust) and moral incapacity (because of moral character and self-identification with certain projects) as expanding on the phenomenological difference between being disgusted and being morally outraged.

3 The Incentive-Insensitivity View of Integrity and Self-Identification

So, we have a more comprehensive notion of moral incapacity, which applies not only to acting against those projects with which one self-identifies (i.e., identifies in a way that is constitutive of who one is as an agent); but also to reconsidering one's very self-identification with these projects. Now, Williams maintains that moral incapacity is supposed to be 'proof against reward'. The question is: how do we make sense of such 'reward-proofing', especially regarding this new understanding of moral incapacity? What (if anything) makes it rational for an agent to resist revising her self-identification with a project? Conversely, what kind of reasons might make it rational for the agent, from her own point of view, to reconsider her own self-identification?

Putting the question of self-identification this way puts things squarely in the realm of reasons. So, it might be useful to think of different kinds of reasons. In particular, I think we should explore the distinction between object-given and state-given reasons. Roughly speaking, object-given reasons for having an attitude are facts about the object of that attitude (say, the thing that one desires, or what one believes, or intends). State-given reasons for having an attitude are facts about having the attitude, regardless of whether that attitude is warranted, or justified, or correct (say, the benefits of believing something, regardless of whether it is true or not, is a state-given reason for having a belief). More accurately, state-given reasons for an attitude are reasons for getting oneself to have that attitude, by taking relevant necessary steps towards having this attitude (e.g., conditioning, hypnosis, seeking object-given reasons for this attitude, etc.).

Now, ordinary attitudes are (ideally) sensitive to object-given reasons.[5] For example, if I believe that Italy won the 1994 soccer world cup, and I come across especially strong evidence that Italy in fact lost the final match in penalties, then I now have an object-given reason that justifies revising my belief about the identity of the match's winning team. Or, if I learn that a necessary means to executing my intention to become the world's greatest guitarist is to kill someone, this is an object-given reason to kill that person (or, if I am more committed to certain moral principles, it is a very good object-given reason to abandon my guitar world domination intention).[6]

There is no reason to think that self-identification with a project is any different. In other words, one's commitment to a project is also (ideally) sensitive to object-given reasons.

[5] At least, this is true for "judgment-sensitive" attitudes (Scanlon 1998).

[6] This is all from the point of view of the agent herself. The focus here is on what reasons the agent has (as she takes them to be).

For example, suppose God revealed to Luther that the Catholic way (with its sale of indulgences and what not) is the right way. This would be very strong evidence for Luther that his own judgments about the Catholic Church were incorrect. Luther would have a (quite conclusive) object-given reason to revise said judgments. Or, if one learns (what one takes to be) a terrible fact about a beloved friend, this would be a very good object-given reason to revise one's attitude towards this friend.

The difference in deliberational stability between self-identification and other, ordinary attitudes is to be found, then, in their different sensitivity to state-given reasons – reasons that pertain to the fact (and benefits) of having these attitudes – and how such reasons can justify reconsidering and getting oneself to revise one's judgments.

When it comes to ordinary attitudes, it may be rational to get oneself to revise one's ordinary attitudes (e.g. for beliefs or intentions) based on any sufficient state-given reason. For instance, a sufficient financial incentive may justify – for an agent, from her own point of view – taking steps to getting herself to believe that the Earth is flat. Pascal's wager provides us with a state-given reason for having a belief in God: it provides us with reasons to take steps that can help us come to believe in God's existence and commands (e.g., attend religious ceremonies, read scripture, discuss religion with religious people, etc.). Ordinary attitudes are, in this sense, sensitive to incentives.

This is not the case for self-identification. Or so I argue. To self-identify with a project, to be truly committed to someone, or something, one must be insensitive to incentives. This ties together with the idea of moral incapacity, as implicated in self-identification with a project. Moral incapacity is meant to be 'proof against reward'; and we can understand this claim now as the claim that one's self-identification with a project is supposed to be resistant to state-given reasons: neither for acting (or, getting oneself to intend to act) in ways that go against this project, nor for reconsidering (or, getting oneself to revisit, revise, or abandon) one's very self-identification with that project. Having integrity consists in being insensitive to state-given reasons – i.e., incentives – to change. Experiencing a moral incapacity means that one "cannot be bought".

Consider the case of Martin Luther: there he is, at the Diet of Worms, exclaiming "Here I stand, I can do no other". If God suddenly revealed to Martin Luther that the Catholic church and its practices are exactly right, this would be an object-given reason for Luther to change his mind and say: "Well, actually, after considering new evidence that has come to light, it turns out that I can do other". Changing his religious convictions and his anti-Catholic views and project – with which he deeply self-identifies – would thus be justified for Luther (from his own point of view).

In contrast, suppose that the Holy Roman Emperor offered Luther incredible riches and political power (perhaps even the Papacy itself!) in exchange for Luther's abandoning his anti-Catholic convictions. This would be a state-given reason for Luther to revise and abandon his religious convictions and anti-Catholic project with which he deeply self-identifies. According to the incentive-insensitive view of integrity, this kind of reason can never justify for Luther (from his own point of view) reconsidering and getting himself to revise his convictions and project, nor his identification with these convictions and project.[7]

3.1 Coherence is an Exception

According to the incentive-insensitivity view of integrity and moral incapacity, no state-given reason could ever justify for the agent (from her own point of view) reconsidering, or getting herself to revise or abandon those projects with which she self-identifies, or her very self-identification with those projects.

There is one exception: coherence. According to the incentive-insensitivity view of identification, the only state-given reasons that can justify reconsidering or getting oneself to revise one's self-identification with a project are reasons of coherence. If two or more of the projects with which one self-identifies contradict one another, then that – and only that – constitutes a state-given reason that can justify (for oneself, from one's own point of view) reconsidering or even getting oneself to revise (at least one of) them. Getting oneself to change one's self-identification with a project for non-coherence state-given reasons is never rational, or justified, from the agent's own deliberational, reasoning point of view. The question is: why?

Before answering this question, it is important to explain what 'conflict' or 'incoherence' means, when it comes to projects with which one self-identifies. For the purposes of this argument, we can say that two projects, X and Y, conflict or are incoherent with one another if, and only if, X renders some action *A* a matter of practical necessity (in the sense discussed by Williams 1981b), while Y renders that same action *A* a matter of moral incapacity (in the sense discussed here and by Williams 1993). In such a case, given one's self-identification with projects X and Y, one must do what one must not do.

This is a terrible predicament for an agent, as it threatens her sense of self and the stability of herself. It generates a crisis of self-identity. The way for her to solve this crisis is by reconsidering her identification with projects X and Y,

[7] One may wonder if threats can justify anything like this for Luther, or for anyone who is genuinely committed to- and self-identifies with a project. I discuss this below.

and perhaps get herself to revise this identification, or even abandon one of these projects. So, conflict forces the agent into a situation where she must reconsider and perhaps get herself to revise her projects.

In fact, such conflict does not only *force* the agent to reconsider her self-identification with her projects. It also *justifies* such reconsideration (for her, from her own point of view). This for two reasons. First, projects provide us with reasons to act. They render certain actions unthinkable (i.e., we are morally incapable of performing them) (Williams 1993); they shape our reasoning in ways that render certain actions necessary from our deliberational standpoint (William, 1981b); and they provide us with reasons to prefer performing some actions over others (e.g., pick up child from Karate class rather than take a nap). This means that, if there is conflict between projects with which one identifies, then we must have project-dependent, object-given reasons against at least one of these projects. If projects X and Y conflict in the way described above, then this means that there is at least some object-given reason pertaining to the goodness or value of X (or Y) that also pertains to the badness or disvalue of Y (or X). Discovering that there is a conflict between X and Y gives us a (state-given) reason to change our self-identification with X or Y in this sense, because it gives us a reason to figure out where we went wrong: there is an object-given reason against X (or Y), a fact that pertains to the goodness or badness of X (or Y), that we missed, and that we now have to take into serious consideration (something we failed to do before, apparently).

The second reason why conflict between one's projects (with which one self-identifies) is a state-given reason that can justify reconsidering one's self-identification with one's projects has to do with the role that self-identification with projects is supposed to play in our reasoning and agency.

Self-identification with a project, or having a deep, self-constitutive commitment to some end (or value, or person, or vocation), is supposed to set normative limits to our reasoning and deliberation; to render us morally incapable of performing certain actions (Williams 1993); or make it volitionally necessary that we are unwilling to perform some action, and that we are unwilling to change this unwilliness (Frankfurt 1982); or categorize certain actions as 'base' or 'abhorrent', such that we completely disvalue acting one them (Taylor 1976); or 'silence' courses of action (McDowell 1979); or perhaps make it unjustified or irrational for us to ever perform such actions for specific reasons (Raz 1975). Self-identification with a project or a commitment can function this way – i.e., as setting normative limits to one's reasoning and deliberation – only if it is resistant to all but coherence-related state-given reasons. This is what gives such attitudes more deliberational stability than other, ordinary attitudes. More specifically, if an attitude is not resistant to coherence-independent state-given reasons, it cannot function as a self-identification attitude. That is, it cannot place normative constraints on one's deliberation.

Consider Luther once again: if a self-identification attitude is not resistant to coherence-independent state-given reasons, then Luther might as well have expressed a conditional statement: "If you don't make it worth my effort, I can do no other". If one can be justified in manipulating one's project and self-identity for a state-given reason other than coherence, then this amounts to self-identifying with that attitude only conditionally, i.e., until a good enough incentive comes along. So, without this extra deliberational stability – without incentive-insensitivity – an attitude will lack the deliberational and temporal stability it needs to function the way it does in placing normative limits on one's reasoning and deliberation. It is this extra resistance to reconsideration that self-identification attitudes have over other attitudes, and which renders them self-constitutive.[8]

Here's how this argument works: Self-identification with a project must function as structuring one's reasoning and deliberation. If self-identification is to function as structuring reasoning and deliberation, then it must, among other things, structure one's reasoning in a way that makes it irrational for one to reconsider this very self-identification (if it does not "recommend itself" in this way, then it will not be able to enjoy its special status in one's reasoning, because it will be too easy for one to change one's mind about whether to self-identify with this project).[9] If self-identification self-prescribes in this way, then the only state-given reason to reconsider and get oneself to change it will be a conflicting prescription made by another project with which the agent's self-identifies. Any other state-given reason to reconsider and get oneself to change this attitude will be undermined by the aforementioned self-prescription. The only thing that can outweigh this self-prescription is the self-prescription of another self-identification attitude. It can do so precisely because this other self-identification attitude also structures

[8] Sometimes, *object*-given reasons (reasons that bear on whether one's judgment is correct or warranted) can also be *state*-given reasons. They can justify not just revising one's self-identification attitudes, but also manipulating these attitudes, i.e. *getting oneself to* revise them. Suppose one learns a terrible truth about a close friend. Such a revelation is an object-given reason to revise one's self-identification attitude towards this friend, but it is also a reason to go and find further evidence regarding the friend's true character. This would be a reasonable response: if we learn some shocking or surprising news about something, it is reasonable to find out more information. In this way, one exerts what Pamela Hieronymi calls 'managerial control' (2006, 2008, 2009) over one's self-identification attitude. Indeed, sometimes such news can be so shocking that one might not be able to shake off one's self-identification attitude by simply relying on evidence.

[9] This kind of self-recommendation or self-prescription is similar in its nature to the 'immodesty' of belief-formation methods (Elga 2010; Lewis 1971).

one's reasoning and deliberation. And so, we reach the conclusion that in order to function as structuring one's deliberation, the only state-given reason that can justify getting oneself to change one's self-identification attitude is conflict.

Self-identification with a project is more stable than other, ordinary attitudes (e.g., beliefs, intentions, plans) because, unlike other attitudes, it is irrational for oneself (from one's own point of view) to reconsider or get oneself to revise one's self-identification with one's projects for any state-given reasons other than conflict. This is unlike ordinary attitudes, because sufficiently strong state-given reason (e.g. an incredible reward or cash incentive) could, in principle, justify reconsidering and getting oneself to revise such attitudes. This extra deliberational stability is what distinguishes self-identification with a project from other, ordinary attitudes, and renders it suitable for constituting the self. In this sense, the incentive-insensitivity view of identification is a (modestly) revisionist view: if one can be justified in getting oneself to revise an attitude for coherence-independent state-given reasons, then it is not an instance of self-identification at all.[10]

3.2 Threats are not a Problem for the Incentive-Insensitivity View

We have seen that conflict between projects is a state-given reason that justifies reconsidering one's self-identification with one's projects, in a way that does not undermine one's integrity. One might think that threats, too, should count as state-given reasons that justify reconsidering and getting oneself to revise one's self-identification with one's projects. As Williams points out in his discussion of moral incapacity, there is always the possibility of such terrible threats, that could justify for someone doing what they would otherwise have been morally incapable of doing (1993, p. 69). Now, if terrible threats can justify *doing* something that goes against one's very self, it seems that they can also justify *reconsidering* and *getting oneself to revise* this self. For example, if I tell you that you must stop loving your children to save their lives, then it may very well be justified for you to get yourself to stop loving your children. Such 'getting yourself' to stop loving your children may require immersion and habituation in some bizarre anti-children society, putting yourself through some Pavlovian conditioning to clear out whatever loving feelings you have towards your children, or taking an "anti-loving" pill (if one is available to you). The upshot is that threats seem to pose a problem for the incentive-insensitivity view of integrity and moral incapacity: they seem to provide one with coherence-independent state-given reasons that can justify reconsidering one's self-identification. Doesn't this put into question the incentive-insensitivity view?

There are two important points to note here. First, we have already seen that, for Williams, threats can get us to do what we are morally incapable because coercion forces us outside of our own moral lives. When we are under coercion, we are not acting or reasoning from within anymore. Rather, our actions are drawn from some external sources of deliberation. In this sense, it is no longer a matter of character (and therefore, not a matter of character incapacity).

Second, it is helpful to take a closer look at what makes threats work. Why exactly are threats threatening? How can threats justify reconsidering one's self-identification attitudes? I think that the answer is that they do so by placing someone in a terrible dilemma. Specifically, threats work best precisely when they force an identification conflict on someone. Threats justify reconsidering and even getting oneself to change our commitments precisely when and because they generate a conflict between these commitments.

Suppose the Holy Roman Emperor tells Luther "I'll torture your family if you don't retract your theses!" The threat is successful to the extent that Luther self-identifies with protecting his family from harm. It forces Luther to choose between two terrible options: retracting his theses or failing to protect his family. That's what makes it so terrible. And such threats will be more successful the more they go after deeper instances of self-identification. If Luther self-identifies with his religious views but not with caring for and about his family, then the threat will be less likely to succeed. When structured this way, threats are not really an exception to the rule. Rather, they are simply another class of cases where one may be justified in reconsidering and revising one's self-identification for state-given reasons of coherence. If we assume that the conflict that Luther faces is between projects that Luther does not only care about, but rather cares about them in the special way that entangles his self-identity with them, then Luther's best bet is to reconsider and revise his identification with those projects in the way that is most coherent with other projects with which he self-identifies.

We noted that Luther would be justified in revising his convictions about not doing any other if he learned of reasons that bear on the correctness or warrant of his self-identification attitudes (specifically, his judgments about the things that are wrong with the Catholic Church). This would be the case, for instance, if he came across what he considered as good evidence that his theses were just plain wrong. This is like other judgments, which can be rationally revised for sufficient object-given reasons. Things are

[10] One important note of clarification: The discussion here is restricted to revising and reconsidering *existing* instances of self-identification. It does not discuss reasons to self-identify with *new* projects. I take it that getting oneself to take up such new self-identification can be justified by sufficient state-given reasons. I discuss this in more detail in the concluding section of the article.

similar with the pacifist. If she learned that pacifism was false, this would be an object-given reason that would justify revising her self-identification attitude.

We also noted that, if presented with a threat to the safety of his family, it may be justified for Luther to get himself to revise his attitude about doing no other. Similarly, if the pacifist had to choose between her commitment to non-violence and her children's well-being, it could be justified for her to get herself to revise her principles about violence. This is like other judgments, which can be rationally revised for state-given reasons of coherence.

3.3 Recap: Integrity and Self-Identification as Incentive-Insensitivity

What about other state-given reasons? Could they ever justify getting oneself to revise one's self-identification attitudes? Recall that other, ordinary attitudes are rationally open to manipulation for any sufficient state-given reason. If I am offered a sufficiently attractive reward for believing that two and two are five, I may be perfectly justified in doing whatever I can to manipulate and manage my beliefs accordingly (e.g. go to a hypnotherapist and have them cast away this whole 'two and two are four' rubbish from my head). And if someone offers me a sufficiently strong incentive to *intend* tomorrow to drink some bad medicine next week (so I don't have to actually drink it), it would again be rational of me to go ahead and get myself to have that intention (e.g. condition myself with some Pavlovian method to like the smell of the medicine) (Kavka 1983).

But self-identification must be more resilient than this. Consider the following: Luther is offered a lifetime of riches, provided he changes his mind about whether he can retract and renounce his theses. Or maybe Luther is offered the position of Pope: "Change the system from within, Martin!" the Emperor tells him. Could these or other incentives and reasons ever justify Luther's manipulating his self-identification attitude?

No. It is conceptually impossible for there to be any coherence-independent state-given reason that can justify for Luther to manage and manipulate his self-identification and his projects. If Luther accepted any such offer, this would show us that we were wrong in claiming that he self-identified with his religious convictions in a way that made him unable to do any other. Similarly, if a pacifist started to intentionally profit from violence and war (e.g., started working for an arms dealer) for a high enough salary, then this would show us that we were wrong in claiming that she truly self-identified with pacifism.

The general idea here is that the goodness or badness of self-identification with a project does not justify reconsidering or getting oneself to revise it. Otherwise, one will not experience moral incapacity; one's integrity will fail to be proof against rewards. In other words, one can never have sufficient coherence-independent state-given reasons to manipulate one's projects and self-identification, because self-identification with a project simply must, as a matter of constitutive fact, be resistant to reconsideration for coherence-independent state-given reasons. If a commitment to a project is not resistant to reconsideration in this way, then it just isn't a self-identification kind of commitment or project. It is something else.

4 Conclusion: Benefits, Loose Ends and Objections

4.1 The Benefits of the Incentive-Insensitivity View over the Competition

With the incentive-insensitivity view we get all the benefits of competing views, without the problems. Consider the integrationist view of self-identification. This view maintains that projects and commitments with which one identifies are constitutive of the self in virtue of their being deeply and well integrated: they are causally effective in motivating one to act in certain ways, and one is not displeased with their so functioning (Arpaly and Schroeder 1999). The integrationist view is susceptible to what I will call the 'right price' objection. According to this objection, one may self-identify with a project in an integrationsit way, and still be justified – in one's own lights – to reconsider this self-identification for coherence-independent state-given reasons. This is because of a lack of normative restriction on the notion of being pleased (or not displeased) with the causal efficacy of one's commitment. The problem is that one can be pleased (or not displeased) in many ways. In particular, one may be pleased by way of reward. A sufficient incentive may make reconsidering one's projects and commitments more pleasing to oneself than keeping one's commitment or project. Consider Luther. If we construe his rejection of the Catholic church in terms of integrationism, then we can offer him a reward that would satisfy him so much that it could suffice to suffer the conversion to Catholicism.

The incentive-insensitivity view provides us with a *normative* account of self-identification, unlike the *descriptive* account provided by integrationist views. It construes identification with a project or commitment in terms that are precisely antithetical to the 'right price' problem. On the incentive-insensitivity view, self-identifying with a project or commitment, by definition, precludes the possibility of any coherence-independent state-given reason ever justifying reconsidering or getting oneself to revise or abandon this project.

Moreover, the normative account of self-identification that the incentive-insensitivity view provides does not suffer

from the problems faced by valuational views. According to valuational views, one's self-identity consists in one's values. One's values, on this view, constitute one's true self, and so it is in valuing that one self-constitutes (Taylor 1976; Watson 1975). Valuational views face the problem of 'perverse' cases: sometimes we might value acting based on attitudes that we do not judge as valuable, all things considered (Watson 1987). In other words, our values sometimes contradict our all things considered judgments.

Based on the incentive-insensitivity view, we can rather straightforwardly say that one's agential standpoint can be understood as the overall set of one's non-conflicting projects with which one self-identifies. Consequently, the incentive-insensitivity provides a clean answer to the 'perverse cases' objection: we can say that an action is not expressive of the self yet is expressive of one's valuations – i.e., that one's valuations are not constitutive of the self – precisely when one would be rational to reconsider those valuations for coherence-independent state-given reasons. In one's own view, one would be justified in getting oneself to change how one values a certain action. Therefore, this valuation is not self-constitutive according to the incentive-insensitivity view of self-identification.

Finally, the incentive-insensitivity view avoids the regress problem faced by hierarchical views. According to Hierarchical views of self-identification, identifying with a project consists in having some sort of higher-order pro stance towards said project (Bratman 1996, 2007; Frankfurt 1988; Korsgaard 1996). One common objection raised against hierarchical views is that they lead to regress: true identification with an attitude may require identification with the relevant higher-order pro stance towards that attitude, which would require having a higher-higher-order pro stance, and so on (Bratman 1998; Watson 1975; Wolf 1987).

In contrast, on the incentive-insensitivity view, self-identification with a project only requires that one is not justified in reconsidering and getting oneself to change one's identification with this project for coherence-independent state-given reasons. It does not require that one further self-identifies with this self-identification or that one's incentive-insensitivity itself needs to be justified or incentive-insensitive. So, the incentive-insensitivity view of self-identification gives us all the benefits of other views, without being open to any of their inherent problems.

4.2 Loose Ends: State-Given Reasons *for* Self-Identifying with a *New* Project

The argument in this article focused on how state-given reasons might justify reconsidering and perhaps getting oneself to revise or abandon one's *existing* commitments and projects. I have not discussed here the possibility of self-identifying with a *new* project for state-given reasons. I believe that such cases are actually perfectly fine, on the incentive-insensitive view.

Consider the following case. Suppose I need a kidney transplant, and someone offers me their kidney (which is a match), but only if I truly care *about* (and not merely care *for*) their daughter (Rabinowicz and Rønnow-Rasmussen 2004). The fact that I will get a kidney is a state-given reason to get myself to care about the donor's daughter. The prospect of a new kidney is a reason to (get myself to) have this new commitment; but this reason does not pertain to whether the child herself warrants my caring about her. It is a reason for *getting myself* to care about the donor's daughter. I take it that, all else being equal, the prospect of a new, functioning kidney will probably be a sufficient, state-given reason to get myself to self-identify with this attitude.

Here, caring about the donor's daughter is understood as a project that I take up. This is as opposed to 'caring for', which is understood here as 'taking care of', i.e., as a collection of actions and practices involved in nurturing, protecting, and supporting someone's well-being. Caring *about* someone implies seeing them as worthy of one's caring (e.g., they are precious to oneself, or one loves them). Caring about someone and caring for someone are distinct things. For example, A can care *for* B without caring *about* B (e.g., if A is a nurse and B is someone who killed A's mother). Conversely, A can care *about* B without caring *for* B (e.g. if A is B's father and one of them is in prison). More importantly, A may be justified in caring *for* B without being unjustified in not caring *about* B. The nurse would be justified in caring for his mother's killer while not caring about him. Alternatively, A may be justified in caring *about* B without being unjustified in not caring *for* B. The imprisoned father may care about his child, but he cannot be expected to care for his child while he himself is locked up.

Let's go back to the example. I am offered a kidney. To receive the kidney, I must care *about* the donor's daughter. The prospect of a functioning kidney is a reason to have a caring attitude towards this child. It is a state-given reason to have this attitude. It is a coherence-independent state-given reason to get myself to self-identify with a project (and to get myself to take up that project in the first place).

According to the incentive-insensitivity view, if I do not have a conflicting self-identification commitment or project, there is nothing wrong here in terms of integrity and self-identity. I have a reason to get myself to have a commitment to the donor's daughter and to get myself to self-identify with this commitment. So long as I do not have any conflicting self-identification projects and commitments, then getting myself to form this attitude and to self-identify with it for this state-given reason can be justified (for me, from my own point of view). If there is such contradiction, then I should not take up the new project and abandon an existing one, precisely because this would amount to revising a

commitment for an incentive. If the only way to take up this new incentive-based commitment is by revising an existing commitment, then this is equivalent to revising an existing commitment for an incentive.

The main point here is more broad. Our projects and our self-identification with these projects come from many sources: some are caused independently of our reasoning (e.g., being born into a religious sect), some through education and socialization processes, some through reasoning, and some through strong emotional experiences. Some of these might even be voluntary (one chooses a career in law, or in professional tennis). The kidney case may easily be just another one of these cases. The important question for integrity, however, is not where our projects and commitments come from, but rather what we do with them *once* we have them, and are we justified – in our own lights, at the very least – to hold or revise them. The incentive-insensitivity view is quiet on where our projects and commitments comes from, or how they achieve self-identification status. It only focuses on what it means to self-identify with an attitude once it is already there in our minds. This is a significant enough question in its own rights.

4.3 State-Given Reasons, Commitments, and Instrumental Rationality

Finally, a possible objection. I have argued that the only state-given reason that can justify getting oneself to revise one's self-identification is coherence. One possible counterexample to this would be a case of getting oneself to revise one's self-identification for the benefit of the object of the relevant project.

Suppose God tells you that He can guarantee your children will live a long, happy, and fulfilling life, filled with all and only the things that make one's life go well. This is great news: given your love toward and care about your children, we can assume that this is precisely what you want for them. The only condition is that you must stop loving them and caring about them. It seems, then, that God's promise can justify (for you, in your own view) reconsidering and getting yourself to revise the love and care that you have toward your children. If the well-being of your children is truly important to you, then changing your mind about them should be a small price to pay for securing their future well-being. At least, it is a price you should be willing to pay.

We may formulate the objection like so: if one is committed to some end *E*, then one should be ready to act in ways that promote *E*; at least, assuming that the benefit is significant enough, and one has adequate opportunity to act. On the flip side, risking *E* or acting in an anti-*E* way seems irrational for someone who values or is committed to *E*. If abandoning one's commitment to *E* would significantly promote *E*, then one should abandon one's commitment to E (all else being equal). So, there can be sufficient, coherence-independent state-given reasons to get oneself to revise one's commitments and projects, namely, a significant benefit to these very projects.

I believe that this last conclusion is unwarranted, and so the objection fails. Consider: there is a difference between committing to *E*, and committing to <committing to *E*>. Projects have ends. Committing to a project or self-identifying with a project entails committing to promote those ends. Higher-order committing to so committing objectionably fetishizes self-identification. Self-identification with a project does not require committing to <committing to *E*>. Instead, it requires – or, rather, consists in – reasoning and acting in ways that promote *E*. (Compare: one dies after risking one's life to save a loved one. In dying, one ceases to be committed to the loved one, because someone ceases to be.) Ignoring the requirements of one's projects only to preserve oneself as someone who is committed to these projects is rather vain and self-centered.

The general point here is that projects have ends. These ends place us under the normal requirements of instrumental rationality. In particular, they place us under the requirement to take necessary steps to promote these ends. If we fail to do so, we fail to live up to our own commitments to these projects. Thus, getting oneself to stop loving and caring about one's children – for the divine guarantee of their well-being – is not a problem case for the incentive-insensitivity view of integrity. It is precisely because this is a reason that pertains to the object of one's project and commitment, that it is not an "incentive". It is not external to one's moral life. In getting oneself to change one's commitments for the benefit of the ends towards one is committed, one acts with integrity, because one is not 'bought' at all.[11]

Acknowledgements This work germinated from my dissertation. For countless written comments and discussions on predecessors of this manuscript, I am wholeheartedly grateful to Philip Clark, Arthur Ripstein, and Sergio Tenenbaum. I am also indebted to two anonymous reviewers for this issue of Topoi for their comments and suggestions, as well as to anonymous reviewers from other journals.

Funding This research was supported by the ISRAEL SCIENCE FOUNDATION within the MAPATZ program (grand No. 399/23).

Conflict of interest There are no conflicts of interests to disclose.

[11] These comments are very brief, and open up further questions. For instance: perhaps there is room for distinguishing between object-directed state-given reasons (state-given reasons for revising an attitude to benefit the object of the attitude) and state-directed state-given reasons (any other state-given reasons). Based on this distinction, God's promise for the well-being of one's children would be an object-directed state-given reason. I hope to address this and related questions in further work.

References

Arpaly N (2002) Unprincipled Virtue. Oxford University Press, New York

Arpaly N, Schroeder T (1999) Praise, blame and the whole self. Philos Stud 93(2):161–188

Arpaly N, Schroeder T (2013) In praise of desire. Oxford University Press, New York

Bratman ME (1996) Identification, decision, and treating as a reason. Philos Top 24(2):1–18

Bratman ME (1998) The sources of Normativity. Philos Phenomenol Res 58(3):699–709

Bratman ME (2000) Reflection, planning, and temporally extended Agency. Philos Rev 109(1):35–61

Bratman ME (2007) Structures of agency. Oxford University Press, New York

Chang R (2013) Grounding practical normativity: going hybrid. Philos Stud 164(1):163–187

Elga A (2010) How to disagree about how to disagree. In: Feldman R, Warfield TA (eds) Disagreement. Oxford University Press, New York, pp 175–186

Frankfurt HG (1982) The importance of what we care about. Synthese 53(2):257–272

Frankfurt HG (1987) Identification and wholeheartedness. In: Schoeman F (ed) Responsibility, Character, and the emotions. Cambridge University Press, Cambridge, pp 27–45

Frankfurt HG (1988) The importance of what we care about: philosophical essays. Cambridge University Press, New York

Hieronymi P (2005) The wrong Kind of reason. J Philos 102(9):437–457

Hieronymi P (2006) Controlling attitudes. Pac Philos Q 87(1):45–74

Hieronymi P (2008) Responsibility for believing. Synthese 161(3):357–373

Hieronymi P (2009) Two kinds of Agency. In: O'Brien L, Soteriou M (eds) Mental actions. Oxford University Press, Oxford, pp 138–162

Hieronymi P (2011) Reasons for Action. Proc Aristot Soc 111(3):407–427

Kavka GS (1983) The toxin puzzle. Analysis 43(1):33–36

Korsgaard CM, O'Neil O (1996) The sources of normativity. Cambridge University Press, Cambridge

Korsgaard CM (2009) Self-constitution: Agency, identity, and integrity. Oxford University Press, Oxford

Lewis D (1971) Immodest inductive methods. Philos Sci 38(1):54–63

McDowell J (1979) Virtue and reason. The Monist 62(3):331–350

Parfit D (2001) Rationality and reasons. In: Edgonsson D, Josefsson J, Petersson B, Ronnow-Rasmussen T (eds) Exploring practical philosophy: from action to values. Ashgate, London, pp 17–39

Piller C (2006) Content-related and attitude-related reasons for preferences. Royal Inst Philos Supplement 59:155–182

Rabinowicz W, Rønnow-Rasmussen T (2004) The strike of the demon: on fitting pro-attitudes and value. Ethics 114(3):391–423

Raz J (1975) Practical reason and norms. Oxford University Press, Oxford

Scanlon TM (1998) What we owe to each other. Harvard University Press, Cambridge, MA

Schroeder T, Arpaly N (1999) Alienation and externality. Can J Philos 29(3):371–387

Taylor C (1976) Responsibility for self. In: Rorty AO (ed) The identities of persons. University of California Press, Berkeley, pp 281–299

Taylor C (1985) Human Agency and Language: Philosophical Papers 1. Cambridge University Press, Cambridge

Watson G (1975) Free Agency. J Philos 72(8):205–220

Watson G (1987) Free action and free will. Mind 96(382):145–172

Williams BAO (1973a) A critique of utilitarianism. Utilitarianism: For and Against. Cambridge University Press, Cambridge, pp 77–150

Williams BAO (1973b) The Makropolous Case. Problems of the self. Cambridge University Press, Cambridge

Williams BAO (1981a) Persons, Character and Morality. In: Moral luck: philosophical papers, 1973–1980. Cambridge University Press, Cambridge, pp 1–19

Williams BAO (1981b) Practical necessity. In: Moral luck: philosophical papers, 1973–1980. Cambridge University Press, Cambridge, pp 124–131

Williams BAO (1981c) Moral luck. In: Moral luck: philosophical papers, 1973–1980. Cambridge University Press, Cambridge, pp 20–39

Williams BAO (1993) Moral Incapacity. Proc Aristot Soc 93:59–70

Wolf S (1987) Sanity and the metaphysics of responsibility. In: Schoeman F (ed) Responsibility, Character, and the emotions. Cambridge University Press, Canbridge, pp 46–62