Topoi
https://doi.org/10.1007/s11245-018-9608-7

Confabulation, Rationalisation and Morality

Anneli Jefferson[1]

© The Author(s) 2018

Abstract
In everyday confabulation and rationalisation of behaviour, agents provide sincerely believed explanations of behaviour which are ill-grounded and normally inaccurate. In this paper, I look at the commonalities and differences between confabulations and rationalisations and investigate their moral costs and benefits. Following Summers and Velleman, I argue that both can be beneficial because they constrain future behaviour through self-consistency motivations. However, I then show that the same features that make confabulations and rationalisations beneficial in some cases can also make them morally costly, when behaviour is explained and justified through the endorsement of bad moral principles. I show that these effects are most likely to occur where the central element of confabulation, self-explanation, and the central element of rationalisation, self-justification, coincide.

Keywords Confabulation · Rationalisation · Sel-justification · Self-explanation · Morality

1 Introduction

When we confabulate, we sincerely report something that we take to be true, but our claim is ill grounded and, in most cases of confabulation, not in fact true. Some definitions (Fotopoulou 2009) limit confabulations to memory distortions. Broader definitions "encompass any type of false or unjustified belief" (Bortolotti and Cox 2009). Classic cases of confabulation are inaccurate memory reports of past events, or, more relevantly for the paper at hand, inaccurate explanations for what we have done, where our explanations do not link to the facts adequately. Whether these kinds of confabulations are best described as false memories or as inferences about past motivations need not concern us here.

Confabulation is by definition an epistemically problematic activity: individuals are unable to recall their past actions and motivations but confidently make assertions about these.[1] When we confabulate, we lack knowledge about our own past, but also self-knowledge, because we mistakenly attribute certain past emotions, beliefs etc. to ourselves (Strijbos and de Bruin 2015). The concept of confabulation has its original home in clinical cases, for example in amnesic patients. There is some debate on how broad the notion of confabulation should be and what the relevant differences between confabulation in the clinicial and non-clinical population are (Bortolotti and Cox 2009).

In this paper, I will only be concerned with a small subset of non-clinical confabulations and rationalisations. I will be looking at what Lisa Bortolotti calls 'everyday confabulations' (Bortolotti 2018), where agents give justifications for behaviour which are normally false and insufficiently or not at all supported by the evidence available to the agent. More specifically still, I will be interested in moral rationalisations and confabulations, instances where we give an explanation for past behaviour by citing moral reasons for said behaviour, when the claim that these moral reasons were what in fact motivated our action is not justified by the evidence. In other words, the agent confabulates certain moral motives when explaining their past behaviour. I will show both benefits and costs of confabulation and rationalisation, and argue that rationalisations which are *also* confabulations are most likely to have either a strong positive or negative effect on future moral conduct. Whether the effect is positive or negative does not depend on the inaccuracy of the rationalisation, but on the moral principles endorsed in giving it.

The paper is structured as follows—I first introduce the notions of rationalisation and confabulation and ask whether

✉ Anneli Jefferson
a.jefferson@bham.ac.uk

[1] University of Birmingham, Birmingham B15 2TT, UK

[1] While confabulations are generally false, this need not be the case, there is in principle the possibility that individuals accidentally come up with a memory that is not false, but nevertheless only accidentally represents the past accurately.

Published online: 08 November 2018

we can draw a principled distinction between these two. I will argue that while there are cases of pure confabulation, where the main goal is self-explanation, and cases of pure rationalisation which are only geared at self-justification, the two phenomena often coincide.

I then turn to the benefits of confabulation. Summers (2017) has recently argued that there are some benefits to rationalisation and confabulation in the moral realm. I will show that Summers is correct and—drawing on his work and that of David Velleman—I will delineate the kinds of cases where we should expect rationalisation and confabulation to have beneficial effects on agents' moral conduct.

However, I then show the flipside of these mechanisms. While Summers focuses on the positive effects,[2] I argue that the *same* kinds of mechanisms can also lead to negative consequences. Rationalisation and confabulation can have considerable moral costs when they reinforce bad moral arguments and justifications. In some cases, when we explain moral behaviour to ourselves and others in such a way as to make ourselves look good, a central characteristic of motivated reasoning and cognition, this can lead to skewed moral judgment: rather than adjusting our behaviour to our standards, we adjust our standards to our behaviour. The same mechanisms of self-consistency and self-enhancement that lead to positive results can have detrimental results. And, importantly, it is not the fact that confabulations and rationalisations are false at the time of adoption or that they are sometimes enlisted in order to retain a positive self-image that makes them morally costly. Both of these features can drive positive effects. It is only when they coincide with bad moral justifications that things go seriously wrong. Finally, I turn to the question whether confabulation is less susceptible to this kind of problem, because it does not rely on the same kinds of motivations as rationalisation. I conclude that it is when confabulation and rationalisation coincide that we will have the strongest effects, be they positive or negative.

2 Confabulation and Rationalisation

2.1 Introducing the Phenomena

In this section, I introduce the notions of confabulation and rationalisation and explain important similarities and differences between them. The cases of confabulation I am interested in for the purposes of this paper are ones where we give explanations for our beliefs, decisions, or actions which are not well supported by the available evidence and, at least normally, untrue. Probably the most famous example comes from Nisbett and Wilson's studies reported in a 1977 paper. In two studies, participants were asked to evaluate the quality of different nightgowns or identical nylon tights, and their choices were best explained by a placement effect, they always evaluated the items furthest to the right highest. However, when asked to explain their choices, they said they had chosen the items of clothing because of their colour, the superior knitting or similar (Nisbett and Wilson 1977). Because the real reasons for action were not introspectively accessible to participants, they confabulated an explanation.[3] This does not mean that the reasons they confabulated were random. They can be seen as a kind of inference to the best explanation, where participants came up with explanations which were plausible in the light of criteria one might well apply when choosing items of clothing.

Similarly, there are cases of confabulation in moral judgment. In fact, Jonathan Haidt claims that moral reason giving generally is confabulation, because the real drivers of moral judgments are not the ones we appeal to when giving reasons for our moral judgments. According to Haidt (2001), we make moral judgments based on emotions, and the moral theories we use to explain our judgments are confabulations. He illustrates this with an example, where people are given a vignette of a couple of siblings who engage in sex which, according to the vignette, is a one off event which does no psychological harm and does not lead to pregnancy because they use contraceptives. Haidt points out that even though general principles that would make incest wrong (for example risk of birth defects, risk of psychological damage) are excluded by the way the example is constructed, people still believe that the action is wrong and put forward these kinds of reasons when justifying their judgment. When faced with the objection that the general principles do not apply, they are unable to come up with an alternative explanation for the wrongness of incest in this particular case. According to Haidt, they are confabulating an explanation for their moral judgment, even though the real underlying reason is a gut emotional reaction.[4]

A core feature of confabulation is that the correct explanation for our behaviour is not available to us. Things are slightly different in cases of rationalisation, where alternative explanations may be available to the agent, but are not endorsed. Haidt's case is at the boundary between

[2] It is important to note here that Summers does not deny that there may well be negative effects, they just aren't the focus of his paper.

[3] Note that we can sometimes tell that an explanation is likely to be a confabulation even if we do not know what the real underlying reasons for actions were, because the explanation is supported by obviously self-serving or biased reasoning. I thank an anonymous reviewer for pointing this out.

[4] Much more could be said about Haidt's interpretation of the experimental results and whether they establish his general claim. For current purposes, it is sufficient that his interpretation gives us a model of what confabulation in the moral realm might look like.

rationalisation and confabulation, as 'it just feels wrong' may be an explanation introspectively available to the individual, but it is certainly not considered satisfactory.

Rationalisation is not just employed to explain a moral judgment, it is also a common way of justifying *past decisions* or *intended actions*. This can be illustrated by the following example. I am set to meet a friend at the tube station. On leaving the station, I walk past a beggar, who is sitting out in the icy rain. I do not give him any money. I then see my friend emerging from the tube; he stops and gives the beggar money before he joins me in the doorway where I am waiting for him. The contrast between our behaviour leads me to consider why *I* did not give money. It may also lead to cognitive dissonance, because it threatens my self-view as generous and caring. In this situation, I might come up with the following rationalisation: "I should not give money to the beggar because he would just go and buy drugs with it anyway. It's got nothing to do with being stingy or uncaring." What we have here is a rationalisation of my behaviour such as to make it seem morally acceptable.

Rationalisations need not be *moral* rationalisations, they can also be used to avoid the threat of seeming to be irrational or violating prudential norms. So, for example, I might rationalise the purchase of expensive shoes by saying that I was not really being extravagant. as the shoes will last longer and be worn more often than cheaper shoes, and are therefore a good investment. Here, I am just trying to show that I am meeting a standard for spending money in such a way that it benefits me most, and thereby meeting a rationality standard in my purchase.

2.2 The Difference Between Confabulation and Rationalisation

Let us now focus on the differences between confabulation and rationalisation. At first glance, it would seem that rationalisation and confabulation are different in ways that are relevant to their epistemic and moral evaluation. Confabulation is a response to ignorance, where individuals confidently fill the gaps by either confabulating past events (memory confabulations) or confabulating explanations for past or present behaviour. While confabulation is generally understood as epistemically problematic because the confabulated beliefs are unwarranted and normally untrue, confabulations are attempts to avoid admitting ignorance, rather than being attempts to avoid a certain truth that is in principle available to us (cf Bortolotti 2018).

By contrast, rationalisation looks like an instance of motivated cognition. In motivated cognition, we reason in such a way as to achieve certain beliefs that are subjectively desirable, or in such a way as to avoid subjectively undesirable beliefs, or both (Hughes and Zaki 2015). For example, positive illusions, where individuals exhibit an unrealistically positive self-image—also known as the better than average effect—(Taylor 1989), or have unrealistically positive expectations for the future, are classic cases of motivated cognition. In these cases of motivated cognition, evidence is selected in a biased way, so as to reach certain subjectively desirable conclusions. In the example above, we can see the rationalisation of not giving money to the beggar working in the following way: I want to avoid the conclusion that my actions are uncaring and morally questionable, or even immoral. I therefore recruit the thought that beggars often use money to buy drugs as a reason to justify my action.

2.2.1 Self-Explanation and Self-Justification

One important thing to note is that rationalisation can serve two purposes. In the case presented above, I can use the claim 'beggars use money they are given to finance their drug consumption' as a moral justification, without claiming that this was *in fact* the reason why I didn't give money. I could, for example, say, "The reason I didn't give money was because I was cold and it was inconvenient to dig out my wallet in the rain. However, as it happens, this was also the correct thing to do, because the beggar would have only spent the money on drugs anyway, and I would not really have been benefiting him." Alternatively, and more commonly, I could claim that the justification of my action was also my reason for acting the way I did at the time.

Summers characterizes rationalisation as providing the agent both with a justification and an explanation for their behaviour. "The rationalizer offers a justification of her own action as if it were an explanation of her action." (Summers 2017, p. S21) Self-explanation and self-justification often coincide, when I explain my actions in such a way that my explanation of my motivating reasons for action also justifies that action. I may well not have thought about whether or not to give money to the beggar, and only feel the need to access and question my motivations when I see my friend giving money to the beggar. But once I feel the need to explain my behaviour, I explain it in such a way that makes sense of my behaviour (self-explanation) while at the same time justifying it (self-justification). It takes a special kind of detachment to say "I didn't give money because it was inconvenient, but as it happens, that is also the right thing to do." It seems to make the rightness of my action accidental, which is something most people will feel uncomfortable with.

Nevertheless, in rationalisation, justification and explanation can, and in some cases do, come apart, as the above example shows. Only when a rationalisation also serves the purpose of self-explanation will it be an instance of confabulation. Focusing on the two goals of self-justification and self-explanation, we can draw a preliminary distinction. Confabulation primarily serves the purpose of self-explanation, whereas rationalisation is primarily concerned with

self-justification. (Moral) rationalisation can be successful even if it only serves the purpose of self-justification, self-explanation is, as it were, optional.

By contrast, the kinds of everyday confabulation cases we have been considering are best characterised as attempts at self-explanation. In a recent paper, Coltheart (2017) even argues that all confabulation, whether clinical or non-clinical, derives from a feature of human cognition that he calls 'the drive for causal understanding'. I take no stance on whether this is true for all instances of confabulation, including false memory reports, but it is almost trivially a feature of confabulations which explain past decisions and behaviour.

However, there is frequently also an element of self-justification in confabulation. Successful self-explanation does not just require there to be any kind of story, it requires a story that makes sense of one's behaviour and makes it appear rational or moral. One wants to meet certain epistemic or moral standards when explaining one's actions. 'I don't know' is not a satisfactory explanation for one's behaviour, but neither is 'I felt like it' or 'I was in a good mood', because these explanations (while they may actually be true) do not really provide reasons, they are at most causes. This is why the participants in Haidt's study do not settle for 'it just feels wrong' when explaining their judgment that incest is wrong even in the carefully designed vignette they are presented with.

Rationalisations can therefore at the same time be confabulations in cases where individuals are rationalising to fill a gap and to explain their past behaviour to themselves. For those who are unhappy with merging the concepts of confabulation and rationalisation in this way, an alternative way of making this point is to say that some instances of sincere but mistaken explanations of behaviour can be construed either as rationalisations or as confabulations.

While self-justification plays a role in confabulation as well, it seems at first glance that pure confabulation is, at least from a moral point of view, less suspect than confabulation which involves rationalisation, because it is more open to different kinds of explanations for behaviour. While the individual does not want to admit ignorance with respect to the reasons and causes for their decisions and actions, they are not set on a certain explanation that makes them appear in a favourable light, anything that makes sense of their decisions would in principle do. However, what makes sense is plausibly constrained by certain rationality standards, which then shape confabulations.

3 Moral Benefits of Confabulation and Rationalisation

Summers argues that confabulation and rationalisation can have moral benefits because they put pressure on individuals to act consistently with self-ascribed motives.[5] In what follows, I will focus on those rationalisations which also count as confabulations because they arise from the need to fill a gap in the person's understanding of their own behaviour. To illustrate the benefits of rationalisation, Summers also employs the example of a beggar, but in the scenario he describes, a person gives money to the beggar because they are actually afraid of him. However, they do not access that reason and rationalise that it is good to give to people in need. This rationalisation, Summers claims, can have positive knock-on effects by putting pressure on subjects to act in accordance with the self-avowed motive for action in future.

> I now put practical pressure on myself—insofar as I care about or am committed to being a reasonable, consistent, and moral person who treats likes alike—to defend this as a good reason and act according to it in future cases, or at least to distinguish apparently similar cases so that I can still claim to have consistent motives. If I turn the corner and walk past yet another person who appears similarly in need, I will feel some pressure to do one of the following: give to him as well, distinguish the cases ("This person doesn't actually seem to be suffering"), or add some nuance to the reason I previously endorsed ("... when I have extra cash in my wallet.") (Summers 2017, p. S29).

The idea that self-consistency motivations have a positive influence on future action can be found in other writers as well. Velleman (2000) stresses the importance of the self-consistency motivation for making us behave in accordance with certain traits or motivations we self-ascribe. In summarizing psychological literature on cognitive dissonance, self enhancement, self verification and self consistency, he says: "The research appears to show that we tend to act in accordance with the motives and traits of character that we perceive ourselves as having" (Velleman 2000, p. 362, 363). Leon Festinger, the founder of cognitive dissonance theory, also saw the need for cognitive consistency as a core human motive (Gawronski 2012).

Self-consistency motivations and the avoidance of cognitive dissonance are known to drive rationalisations and reinterpretations of past behaviour. In a famous experiment by Festinger and Carlsmith (1959), students were asked to do a boring task (sorting spools) and were then split into three

[5] Summers also lists meaningfulness as a further benefit of rationalisation, I will not discuss that issue here.

groups. The control group were debriefed straight away. One experimental group was asked to introduce the task to the next participant and claim that it was a lot of fun. They were given one dollar for doing this. The second experimental group was also asked to introduce the task and claim that it was fun, but received 20 dollars as compensation. In a debrief, it turned out that when participants were later asked how enjoyable they found the task, the control group and the 20 dollar group reported not finding it enjoyable, with the control group reporting least enjoyment. The one dollar group, by contrast, reported finding it enjoyable. Festinger and Carlsmith explain this by saying that there is more cognitive dissonance pressure for the one dollar than the 20 dollar group, because the higher payment provides the 20 dollar group with a plausible motivation and explanation for claiming the task was interesting. They therefore do not have to reduce cognitive dissonance in the way the one dollar group does. Characterising the task as enjoyable can be interpreted as an implicit rationalisation for doing something that they have no obvious motivation to do.

But, importantly in the current context, consistency motivations also constrain behaviour looking forward. According to Velleman, we act as we do in order to retain a consistent self-image. This consistency is crucial for us to make sense of ourselves. The agent acts in accordance with an existing self-image in order to know what they are going to do (Velleman 2000, p. 366). In some cases of confabulation and rationalisation, this can have positive effects on our behaviour, because we will try to act in ways that fit with our image of ourselves as people who help those in need, or people who are honest etc. These kinds of considerations make it seem that rationalisations that construe our behaviour in a morally positive light are desirable, because they will constrain our behaviour by making us want to act in accordance with our professed motives and values. In the quote by Summers above, we can see how professed reasons for action and moral commitments produce pressure to act in accordance with these. This is also what cognitive dissonance theory teaches us. In the study by Festinger and Carlsmith, cognitive dissonance was reduced post hoc, but it is also possible to preempt cognitive dissonance by acting in accordance with the values we profess to have.

The effect of labelling on future behaviour extends to labels others apply to us as well. Studies with children have shown that labelling them as tidy has a more positive effect on their tidiness than impressing on them the importance of tidiness (Miller et al. 1975), and that telling children that they are generous makes them more likely to behave generously in future (Holte et al. 1984). These kinds of findings suggest that attributions of positive traits and motives will influence behaviour in such a way that people will act in accordance with these attributions, whether they are self-generated or given to us by others.

Self-consistency motivations are however not the only forces at work when we try to live up to our professed moral principles. Self-enhancement motives also play a positive role in motivating individuals to live up to positive attributions. Thus, in the cases where others ascribe positive attributes to individuals as in the studies cited above, one motivation for acting in accordance with the description of oneself as tidy or generous will be the desire to retain a positive self-image. The children who are labelled as tidy and then behave accordingly want to be able to apply positively valued traits to themselves and are therefore driven by self-enhancement motives. A further motivation here may be that we want others to retain their positive image of us, so the motivation is one of wanting external approval rather than internal consistency or validation. But these further motivations are compatible with the hypothesis that others' belief in us as, e.g. 'generous', reinforces that specific positive self-conception[6] and thereby supports behaviour that is in line with that self-conception. An important difference between positive self-attributions and positive attributions made by others is that the standards of what actually constitutes good behaviour are public in the latter case, whereas the individual has far more leeway in the way they define desirable behaviour when attributing positive characteristics, motives or actions to themselves. I will return to this issue in Sect. 4. However, in as far as agents do endorse moral self-descriptions which commit them to specific values and corresponding types of behaviour, such as giving to charity regularly, visiting their friends when they are ill, not eating meat because they care about animals, being a reliable colleagues who answers e-mails on time, these self-ascriptions will put pressure on them to behave in accordance with their self-conception from a self-consistency motivation. The pressure to live up to our own rationalisations will be even higher if these are made public, if we endorse them in front of others.

These considerations suggest that rationalisations in which we endorse certain moral values or principles as driving our behaviour will facilitate behaving in accordance with moral values, as self-consistency motives and self-enhancement motivations put pressure on the individual to live up to their professed values. Self-consistency will lead us to act in a way that fits our existing self-image in order to help us make sense of ourselves and the world. Our need for a morally positive self image constrains our space of available actions and motivates us to act in accordance with what we believe to be right, so that we can maintain a positive moral

[6] For example, Murray et al. (1996) found that when one partner had an unrealistically positive view of their romantic partner, this tended to affect the other partner's self-image over time, which also became increasingly positive.

self-image. Clearly, self-consistency and self-enhancement are only two of the factors that will be at work, so we should not expect a perfect fit between self-professed values and actual behaviour, but we should expect them to have some constraining power.[7]

Furthermore, and more speculatively, our moral self-image may have an indirect effect on what we feel morally obligated to do. If we assume that ought implies can, then being morally incompetent releases us from moral obligations. People frequently use their personal limitations, such as for example illnesses, as excuses for not meeting moral expectations others may have of them (Pickard 2011). This need not be disingenuous. On the plausible assumption that we cannot have moral obligations that we are unable to meet, a self-conception as a person who does not care about the well-being of others, or someone whose behaviour is driven by addiction, makes an individual feel unable to meet moral demands. This, in turn, will make them unlikely to try to meet them. We therefore need to be able to tell stories about our behaviour which emphasize our agency, rather than portraying us as compelled by forces such as addiction or disorder. Sometimes, a positive rationalisation or confabulation will open up space for the adoption of certain motives for action in a way that a more realistic self-assessment would not. If we are concerned with the effects of rationalisation, it does not matter so much that these rationalisations are untrue at the time of adoption. What matters is that they allow us to shape ourselves in a better way going forward.

Self-narratives which emphasize our moral failings risk being self-defeating, by making it seem impossible to act from moral motivations. This can range from extremely determinist narratives which deny agency and moral capacity, to more mildly self-defeating views on one's own behaviour. For example, when we accurately explain our own behaviour to ourselves by saying 'I did not give to the beggar because I don't much care about the plight of others' or 'I did not stand up to the bully because I am a bit of a coward' this may well be accurate, and we are not claiming that we *could* not act from different, more moral motivations. But it would probably be more beneficial for our future behaviour if we are less honest with ourselves. In some cases, inaccurate self-reports which are instances of confabulation or self-justification may actually be preferable in terms of bolstering our moral agency.

[7] Strijbos and DeBruin (2015) point out that we may have to manage our behaviour in a number of ways by setting external incentives and constraints to live up to our self-professed values. Avowal of these values may not be enough, but it guides further steps we take to shape our behaviour.

4 Costs of Rationalisation and Confabulation—The Distortion of Moral Judgment

It is tempting to conclude from the above discussion that we can confabulate and rationalise all we like in the moral domain, and that as long as we end up with explanations which justify what we do as moral, it will all be to the good, because this allows us to become better persons. Unfortunately, this conclusion is unjustified. Put very simply, the effects will only be as good as the moral principles the rationalisation appeals to. What matters is not how accurately the justification provides the agent's actual motives, but how good the justification is in terms of picking out good moral principles, i.e. whether it is an objectively good moral justification and the individual would be morally justified if this had indeed been their reason for action. This characterisation of course raises a further issue, which is whether there are correct and incorrect moral justifications. Resolving this question is beyond the scope of this paper. Instead, I will just assume that we can get it wrong morally. As a working hypothesis, moral principles which do not generalise or ignore the well-being of others should be rejected.

Not all instances of confabulation and rationalisation that show us in a morally positive light are actually morally defensible. Rather, rationalisations can attempt to justify behaviour that should not be justified. Take, once again, the example from the beginning, where I don't give to the beggar and justify this by saying that beggars spend all their money on drugs anyway. This rationalisation of my behaviour is also a rationalisation of not helping a certain subset of those in need, beggars, in a certain way, by giving them money. Leaving aside the question whether this is a justified moral stance to take, it is clear that these kinds of rationalisations can have a severely negative impact on our moral judgment and decision making, because they provide a justification for not helping others, in this case based on some empirical assumptions about the likely effects.

The claim that not all moral self-justifications and rationalisations are defensible is hardly a novel one: Shakespeare caricatures this kind of moral self-justification, in *Julius Caesar* (Shakespeare 1988). In the play, Caesar's killers rationalise their murder, claiming that they are doing Caesar a favour by freeing him from the fear of death:

BRUTUS
Fates, we will know your pleasures:
That we shall die, we know; 'tis but the time
And drawing days out, that men stand upon.
CASSIUS
Why, he that cuts off twenty years of life
Cuts off so many years of fearing death.
BRUTUS

> Grant that, and then is death a benefit:
> So are we Caesar's friends, that have abridged
> His time of fearing death. (Act 3, Scene 1)

While this is not meant to be a psychologically realistic instance of self-justification,[8] it does illustrate something important about the ubiquity and importance of the self-justification motive. The urge to put a positive spin on our actions that makes them appear acceptable is a strong one, and it can influence our moral judgments. In a discussion of self-enhancement and self-protection, Alicke and Sedikides point to ways in which we alter our perception and interpretation of events when we cannot reach a positive self-image by making changes in the external world: "When people cannot promote themselves objectively, they have recourse to construal mechanisms such as reinterpreting the meaning of social or task feedback, misremembering or reconstructing events in a self-serving way, and making excuses for poor behaviour or performance" (Alicke and Sedikides 2009).

Our moral compass is susceptible to being skewed by mechanisms of self-justification, especially when we adjust our judgments of what we deem acceptable gradually through post hoc rationalisations. Existing standards can be eroded in the attempt to find a post hoc or advance justification for unethical behaviour in a process of cognitive dissonance reduction. Changes in our perception of what constitutes acceptable behaviour can often take the form of a slow slide down a slippery slope. We begin by justifying a small departure from ethical conduct: "everyone lies in a situation like that". And once this type of behaviour has come to be perceived as acceptable, more serious lies get justified as not being so different from what one has already declared acceptable, and so on.[9] LaFollette remarks: "[t]hinking that bad behavior will probably cause worse behavior is not surprising: it is precisely what one would expect given our habitual natures" (LaFollette 2005).

It is interesting to note that these real-life slippery slope events are relevantly different from empirical slippery slope arguments put forward in ethical and policy debates, because the first step onto the slope need not be perceived as acceptable. Rather, there are pressures which make this course of action seem attractive (or less unattractive than the alternatives), and the moral evaluation is made to follow suit.[10] We do what we may initially believe to be wrong and rationalise it post hoc. This leaves us with a more permissive approach to a specific issue with the potential to become even more permissive if there are further motivations to do what we had initially believed wrong and we become habituated to this kind of behaviour.

What I have just described gives us a mechanism how self-enhancement motivations can lead to bad rationalisations. These in turn can reinforce bad behaviour by taking moral pressure off the individual, who feels morally justified in their action. If the behaviour is embraced or defended as acceptable, it is likely to be reinforced by self-consistency motives.

In his book on moral disengagement, the psychologist Albert Bandura claims that 'People do not usually engage in harmful conduct until they have justified to themselves the morality of their actions' (Bandura 2016). However, people do sometimes spontaneously act in ways that violate their moral standards. They also sometimes act in situations that they do not yet have a moral standard for, because they haven't considered them from a moral perspective. I take it that Bandura would not disagree, but rather, that his point is that if we do consider the moral character of our action before we act, we will find a way to justify it to ourselves before we perform the action. But we do not always consider the morality or the harmfulness of our actions prior to acting. Bandura's 2016 book on moral disengagement is replete with examples of planned action where agents rationalise their behaviour as morally acceptable. Thus, for example, he looks at the cover-up of child abuse within the catholic church, citing church officials who justified covering up individuals misdeeds by appeal to the damage it would do to the authority of the church if these matters came to light. These types of justification can be, and have been, recruited either before or after embarking on morally dubious behaviour. In either case, they illustrate the downside of rationalisation. This can be rationalisation which is also confabulation, where we try to explain our own behaviour to ourselves because we are genuinely trying to figure out our reason for action. But it may also be pure rationalisations, where we are trying to explain away unwelcome motives.

[8] Note also that this is one of the cases where motive and (the caricature of) self-justification come apart, the motive for killing Caesar was political.

[9] Tenbrunsel et al. (2010) describe a similar scenario, where a consultant becomes habituated to overbilling their clients. At the start, the amount is so small that it does not register on their moral compass and, incrementally, the amount of time the consultant charges for become larger and larger.

[10] In contrast, slippery slope arguments which are put forward in policy debates often assume, at least for the sake of argument, that the action is considered is not problematic in and of itself, but will lead to problematic consequences. For an overview on factors that influence both empirical predictions and the likelihood of slippery slope events occurring, see Jefferson (2014).

5 Conclusion

We find ourselves in rather an odd place. Following arguments by Velleman and Summers, I have shown that there are possible benefits to rationalisation and confabulation, and that they can have a positive effect on our moral conduct. I have argued that what makes rationalisation and confabulation problematic—when they are problematic—is not their falsity or lack of warrant, nor that they stem from the desire to make ourselves appear in a positive light. In fact, the self-enhancement motive can be a positive thing because it makes us try to live up to a desirable moral self-image. If the explanation and justification we give for our behaviour is one that embraces good moral principles, such as being honest, helping those in need or similar, they are likely to lead to good results. This is true even if the self-ascribed moral motive, e.g. benevolence, was not actually driving our behaviour at the time of action and our explanation is therefore strictly speaking false as an explanation of past behaviour.

What makes rationalisation problematic is the fact that there can be rationalisations for an action which satisfy the agent that their action is morally justified when it is not, and allow him to adopt self-serving moral principles, for example 'We do not need to help the poor as they have brought their plight on themselves.' This has further negative effects by leaving the agent with bad moral principles and making future immoral behaviour more likely via self-consistency pressures.

Is this a danger which affects confabulation and rationalisation equally? I have argued earlier that where we have a justificatory gap and the agent aims *both* at self-justification and at self-explanation, confabulation and rationalisation coincide. It is in these cases where there is a motivational element of self-justification in a confabulation that the effects, be they positive or negative, will be strongest, because the justificatory element is at the forefront. This means that these cases of rationalisation involving confabulation are more likely to lead to positive *or* negative effects. If we adopt a rationalisation that is incorrect but reinforces desirable moral values, this is good for our long term moral development and conduct, even if it shows lack of self-knowledge at the time. If we come up with a rationalisation that justifies immoral behaviour as moral, then this is bad for us as moral agents, because it will lead us to continue behaving in that way.

By contrast, there can be a confabulation which explains our behaviour to us without at the same time justifying it. In fact, we may think that our behaviour was unjustified. For example, I may refuse to loan a friend who needs money for a business venture money and, in a self-accusatory mood, think this is because I have always been slightly envious of them and their entrepreneurial spirit. In reality, however, I am just worried about my own finances and the possibility of suffering financial loss. Because I have not tried to find a moral justification for my behaviour, be it an objectively good one or a misguided one, there is no consistency-pressure for me to act up to a specific moral principle I endorse.

Frequently, the need for self-explanation and self-justification coincide, especially in cases of moral confabulation and rationalisation. But to the extent that confabulation is primarily geared towards self-explanation, it is less likely to affect moral agency to quite the same degree as rationalisation.

Acknowledgements I would like to acknowledge the support of Hope and Optimism: Conceptual and Empirical Investigations, a funding initiative by the Templeton Foundation for a project entitled 'Costs and Benefits of Optimism' [Grant ID 46501]; and the support of the Leverhulme Trust for the project *Mental Disorders, Brain Disorders and Moral Responsibility* [ECF-2015-493]. I have presented some of the ideas in this paper at the Society of Applied Philosophy Annual Conference and at the LSE Rational Choice Group and received helpful feedback. My thanks also go to Lisa Bortolotti, Jan-Hendrik Heinrichs, Kathy Puddifoot and two anonymous reviewers at *Topoi* for their comments on earlier drafts.

Compliance with Ethical Standards

Conflict of interest The author declares that she has no conflict of interest.

Research Involving Human and Animal Participants This article does not contain any studies with human participants or animals performed by any of the authors.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

Alicke M, Sedikides C (2009) Self-enhancement and self-protection: what they are and what they do. Eur Rev Soc Psychol 20:1–48

Bandura A (2016) Moral disengagement—how people do harm and live with themselves. Worth Publishers, New York

Bortolotti L (2018) Stranger than fiction: costs and benefits of everyday confabulation. Rev Philos Psychol 9:227–249

Bortolotti L, Cox RE (2009) 'Faultless' ignorance: strengths and limitations of epistemic definitions of confabulation. Conscious Cogn 18:952–965

Coltheart M (2017) Confabulation and conversation. Cortex 87:62–68

Festinger L, Carlsmith J (1959) Cognitive consequences of forced compliance. J Abnorm Soc Psychol 58:203–210

Fotopoulou A (2009) Disentangling the motivational theories of confabulation. In: Hirstein W (ed) Confabulation—views from

neuroscience, psychiatry, psychology and philosophy. Oxford University Press, Oxford
Gawronski B (2012) Back to the future of dissonance theory: cognitive consistency as a core motive. Soc Cogn 30:652–668. https://doi.org/10.1521/soco.2012.30.6.652
Haidt J (2001) The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychol Rev 108:814–834
Holte CS, Jamruszka V, Gustafson J, Beaman AL, Camp GC (1984) Influence of children's positive self-perceptions on donating behavior in a naturalistic setting. J Sch Psychol 22:145–153
Hughes BL, Zaki J (2015) The neuroscience of motivated cognition. Trends Cognit Sci 19:62–64
Jefferson A (2014) Slippery slope arguments. Philos Compass 9:672–680
LaFollette H (2005) Living on a slippery slope. J Ethics 9:475–499
Miller RL, Brickman P, Bolen D (1975) Attribution versus persuasion as a means for modifying behavior. J Pers Soc Psychol 31:430–441
Murray SL, Holmes JG, Griffin DW (1996) The self-fulfilling nature of positive illusions in romantic relationships: love is not blind, but prescient. J Pers Soc Psychol 71:1155–1180
Nisbett RE, Wilson DS (1977) Telling more than we can know: verbal reports on mental processes. Psychol Rev 84:231–259
Pickard H (2011) Responsibility without blame: empathy and effective treatment of personality disorder. Philos Psychiatry Psychol 18:209–224
Shakespeare W (1988) Julius Caesar. In: Wells S, Taylor G (eds) William Shakespeare—the complete works. Oxford University Press, Oxford
Strijbos D, de Bruin L (2015) Self-interpretation as first-person mind-shaping: implications for confabulation research. Eth Theory Moral Pract 18:297–307
Summers JS (2017) Post hoc ergo propter hoc: some benefits of rationalization. Philos Explor 20:21–36
Taylor SE (1989) Positive illusions: creative self-deception and the healthy mind. Basic Books, New York
Tenbrunsel AE, Diekmann KA, Wade-Benzoni KA, Bazerman MH (2010) The ethical mirage: a temporal explanation as to why we are not as ethical as we think we are. Res Org Behav 30:153–173
Velleman JD (2000) From self psychology to moral philosophy. Philos Perspect 14:349–377