Ethic Theory Moral Prac (2015) 18:237–250
DOI 10.1007/s10677-015-9575-7

# How Does Reasoning (Fail to) Contribute to Moral Judgment? Dumbfounding and Disengagement

**Frank Hindriks**

Accepted: 24 February 2015 / Published online: 17 March 2015

**Abstract** Recent experiments in moral psychology have been taken to imply that moral reasoning only serves to reaffirm prior moral intuitions. More specifically, Jonathan Haidt concludes from his moral dumbfounding experiments, in which people condemn other people's behavior, that moral reasoning is biased and ineffective, as it rarely makes people change their mind. I present complementary evidence pertaining to self-directed reasoning about what to do. More specifically, Albert Bandura's experiments concerning moral disengagement reveal that moral reasoning often does contribute effectively to the formation of moral judgments. And such reasoning need not be biased. Once this evidence is taken into account, it becomes clear that both cognition and affect can play a destructive as well as a constructive role in the formation of moral judgments.

**Keywords** Moral reasoning · Moral dumbfounding · Moral disengagement · Cognitive dissonance · Reason · Emotion

Folk wisdom has it that you should think before you act. The underlying idea is that people often come to regret acting on impulse. The folk wisdom seems to be inspired at least in part by moral concerns. People might end up doing the wrong thing when they act in a rash manner. On a pre-reflective or common sense picture of morality, then, it is a good idea to consider moral matters explicitly and think about them carefully. This picture, however, has come under substantial pressure. A lot of research in psychology suggests that conscious reasoning as such is of limited use and merely serves to confirm beliefs we already have (Nisbett and Wilson 1977; Nickerson 1998; Wilson 2002; Johansson et al. 2005).

Jonathan Haidt (2001, 2012) has developed this argument for the case of moral thought. He maintains that, typically, moral reasoning merely serves to confirm prior intuitions and is in this sense *biased*. As people are not interested in changing their opinions, they hardly ever change their moral views. This means that individual moral reasoning is also *ineffective*. Haidt's experiments suggest that people simply reaffirm their pre-reflective moral intuitions when they run out of arguments – that is, in Haidt's terms, when they are 'morally dumbfounded'. This holds for almost all of us for almost all of the time (Haidt mentions philosophers as an exception, as they form 'one of the few groups that has been found to

F. Hindriks (✉)
Department of Ethics, Social and Political Philosophy, University of Groningen,
Oude Boteringestraat 52, 9712GL Groningen, The Netherlands
e-mail: f.a.hindriks@rug.nl

reason well'; 2001: 819). In light of this, I argue in section 2 that the thrust of Haidt's dumbfounding research is this: when it comes to moral matters, *reason is powerless, and intuition carries the day.*

I argue, against Haidt, that objective and effective moral reasoning lies within reach of lots of people and need not be a rare exception. As Monin, Pizarro, and Beer (2007) point out, Haidt's claim to the contrary is based on a rather limited range of empirical research. In order to show that other research supports a more optimistic view, I contrast Haidt's research concerning moral dumbfounding with Bandura (1999; Bandura et al. 1996) research on what he calls 'moral disengagement'. Moral disengagement occurs when someone suspects a conflict between an envisaged action and prior intuitions. Such a conflict gives rise to cognitive dissonance, which is resolved by means of rationalization that facilitates preferred behavior. As such, moral reasoning is *self-serving*. Bandura's research suggests that moral rationalizations often are *effective* in altering the beliefs people form. As cognitive dissonance involves cognition as well as affect, Bandura's disengagement research can be used to illustrate what Monin, Pizarro, and Beer call 'the interplay between emotion and reason' (2007: 102). Whereas Haidt is exclusively concerned with *post hoc* rationalization, Bandura's disengagement research concerns rationalization that occurs prior to an action, or 'anticipatory rationalization'. Moral disengagement typically causes people to flout their moral principles. As I argue in section 3, however, moral reasoning sometimes serves the role of a *disinterested judge*. Before discussing disengagement (section 3) and dumbfounding (section 2) in more detail, I explain what I mean by moral reasoning (section 1).

## 1 Moral Reasoning

Imagine you are driving on the highway and you notice that you are speeding. You realize that you violate the traffic rules and perhaps you recognize that you thereby increase the risk of getting into an accident. At the same time, however, you notice that other people are also speeding. On the basis of this second consideration you conclude that it is ok to drive as fast as you do. In this scenario you engage in moral reasoning. You consider an issue that is morally relevant – whether it is ok to speed given that you thereby increase the risk of harm to which you expose other people. And you bring to bear an observation on this matter – that other people do the same. You use this observation to formulate an argument – it is ok to speed because other people do so as well – and you form a moral judgment – that it is ok to speed.

Without having the ambition to define exactly what it is, let me characterize the notion of moral reasoning in a way that fixes the phenomenon in sufficient detail for the purposes of this paper (cf. Richardson 2013). *An agent engages in moral reasoning when she assesses moral considerations in order to arrive at a conclusion as to what some agent is permitted, prohibited, or obliged to do.*[1] As the agent that is evaluated need not be the agent that is doing the evaluating, moral reasoning *can be other-directed or self-directed*. Arriving at a conclusion is a matter of forming or reaffirming a judgment. Whereas this characterization fixes the *proximate goal* of moral reasoning – to evaluate the propriety of actions – it leaves open whether the *ultimate aim* is, for instance, to persuade people of a possibly prior conclusion (Mercier and Sperber 2011 argue that this is the point of reasoning as such). Persuasion can also be other-directed or self-directed. In the speeding example, you engage in

[1] As this is only a sufficient condition, it is, for instance, consistent with reasoning concerning values being moral reasoning as well.

self-directed reasoning in order to arrive at a conclusion as to whether it is ok for you to speed there and then. And presumably your aim is to persuade yourself that it is.

People sometimes engage in *motivated* reasoning, which is driven by a preference that the agent has for a particular conclusion.[2] If the prior judgment you seek to affirm is in your own interest, as it probably is in the speeding example, you engage in *self-serving* reasoning. In this paper, I follow common practice within psychology and equate rationalization with motivated reasoning.[3] On this usage of the term, your speeding self engages in rationalization, whereas the self that slowed down did not. In terms of the legal metaphor that Haidt uses: the self that slows down reasons as a judge, whereas the speeding self reasons as a lawyer.[4]

Having commented on the goals that moral reasoning can have, I now turn to its content, and ask what the *moral considerations* are that are assessed in moral reasoning. Moral considerations can be *moral principles* – such as the principles of non-maleficence and beneficence – or *values* – such as fairness, autonomy, or friendship. They can also be *moral reasons* more narrowly conceived, such as particular harm or benefit, or particular disrespectful behaviors. As the speeding example reveals, the content of moral reasoning can also feature apparent reasons, such as the observation that everybody else is speeding. For all I know, and as I will suppose, this is not a legitimate consideration. Instead, it is something that you merely take to be a reason. Your reasoning contributes to your judgment in the sense that your judgment depends on it: you could have arrived at another moral judgment had you taken different arguments into account or if you weighted the ones you did consider differently.

In addition to content, there are external influences that bear on the conclusions someone draws.[5] You might, for instance, suddenly remember that you will be home alone tonight and that there is little reason to be home early today. Or you might notice that someone put some flowers on the side of the road, which makes the risk of harming others more vivid in your mind. Perhaps you start feeling like a hypocrite, as you always tell your children to drive safely. The fact that something pops up in your memory, that something is particularly vivid in your mind, or that you feel guilty because of your hypocrisy are factors that do not as such feature in your reasoning. They do, however, influence the conclusion you reach. They do so by influencing the cognitive salience of a consideration, and by influencing the weight that you attach to it.

External factors such as these play an important role in Jesse Prinz's conception of moral reasoning. He takes it to boil down to an emotional struggle: '[W]e deliberate about moral dilemmas by pitting emotions against emotions. Conflicting rules have different emotional strength, and the stronger emotions win out.' (Prinz 2007: 25) Even though moral rules or principles feature in its content, the outcome of the reasoning process is determined by the strength of the emotions that support them. The same holds for reasoning concerning values, which are 'rock bottom' or 'thin' reasons that hardly leave any room for reasoning (2007: 31 and 125). Emotions determine which conclusion the agent draws.

Emotions play a central role in cognitive dissonance. Cognitive dissonance occurs when some cognitive discrepancy, such as a perceived inconsistency between two beliefs, leads to psychological discomfort, for instance in the guise of guilt feelings. This discomfort motivates

[2] Ditto, Pizarro and Tannenbaum define motivated moral reasoning as reasoning 'in which an individual has an affective stake in perceiving a given act or person as either moral or immoral, and this preference alters reasoning processes in a way that adjusts moral assessments in line with the desired conclusion.' (2009: 312)

[3] Philosophers typically say that an agent rationalizes a belief, intention, or action when she provides reasons for it that play a causal role in the formation of the attitude or the performance of the action (Davidson 1963). What motivates the agent in doing so is irrelevant to whether providing reasons counts as rationalizing.

[4] See Haidt (2001, 2012) and Ditto et al. 2009: 309–12) for similar uses of the lawyer-judge metaphor.

[5] I thank an anonymous referee for pressing me to distinguish clearly between content and goals of moral reasoning, as well as factors that influence it.

the agent to reason and find a way to resolve the cognitive discrepancy.[6] In section 3.1, I discuss the role that cognitive dissonance plays in moral disengagement. More specifically, there I explain how people use reasoning to get themselves to believe that an envisaged action that apparently conflict with their moral principles does not do so after all.

How does the concept of moral reasoning relate to the dominant theoretical framework with which psychologists work nowadays: Dual System Theory? Dual Systems Theory postulates two systems of mental processing (Evans 2008; Kahneman 2011). System 1 is automatic, unconscious, and quick, process 2 is controlled, conscious, and slow. Although System 1 usually processes affect and System 2 is dedicated primarily to reasoning processes, the distinction between these two systems does not match perfectly on to that between emotion and reasoning. Emotions can be conscious, and not all cognitive processes require consciousness. The argument I go on to present can be developed in terms of both distinctions. Here I prefer to develop the argument in terms of emotion versus reasoning.[7] This facilitates connecting my conclusions to the debate about sentimentalism versus rationalism. Note also that some moral psychologists use this terminology as well. As mentioned in the introduction, Monin et al. (2007) talk about the interplay of emotion and reasoning. In sections 3 I set out to make precise how affect and cognition can interact, and what role cognitive dissonance plays in the process.

## 2 Moral Dumbfounding

Haidt's (2001; Haidt et al. 2000) experiments about moral reasoning focus on harmless taboo violations performed by unknown others. Examples include consensual sibling sex, cleaning a toilet with the national flag, and eating the pet dog that just died in a car accident. Haidt discovered that, when he asks people to defend the negative verdicts they tend to form, people quickly run out of arguments. A number of those who are morally dumbfounded in this manner refer to emotions as their point of last resort. They simply say: 'It's just disgusting.' And this is supposed to be the end of the matter.

The conclusion Haidt draws from these experiments is that reasoning does not play a significant role in the formation of moral judgments. He claims that many of the arguments people give are post-hoc rationalizations – biased arguments provided after the judgment has been formed – or confabulations – invented arguments that did not play a role in causing the judgment and hence cannot have justified it. Such arguments serve to confirm prior moral intuitions that are due to unconscious, automatic and often affective processes. This implies that moral reasoning does not function as an objective and disinterested judge, but as a prejudiced *lawyer* whose goal it is to make prior intuitions look good. The bottom line of this research is supposed to be that, for the most part, *reason is powerless, and intuition carries the day.*[8]

My main criticism of Haidt's view will be that the dumbfounding findings provide an incomplete picture of moral reasoning. The dumbfounding experiments have, however, also been criticized for suffering from internal problems.[9] Consider, for instance, Haidt's

[6] Churchland (2011) provides useful ideas for developing this account of moral reasoning further. She argues, for instance, that understanding a moral rule is best understood in terms of skills such as cue-based reasoning that involve the use of 'case-based analogies, emotions, memory, and imagination' (2011: 171, 183–84).

[7] In Hindriks (2014) I focus on the role intuitions play in the formation of moral judgment and present what I call 'Sentimental Rationalism' as an alternative to the views of Haidt, Nichols, and Prinz.

[8] Haidt (2012) compares our 'affective receptors' to taste buds. The affective responses they give rise to can be full-blown emotions, which Haidt takes to be cognitive appraisals (2001, 2012).

[9] For more internal criticisms, see Churchland (2012), Fine (2006), Paxton, Ungar, and Greene (2010), Pizarro and Bloom (2003), Saltzstein and Kasachkoff (2004), Sauer (2011), and Sneddon (2007).

assumption that the fact that the taboo violations are harmless implies that the scenarios that feature them form an exception to a moral prohibition. This can reasonably be denied, as many deontologists and rule consequentialists will do. They take morality to consist in rules that do not have exceptions. In light of this, they could, for instance, reject Haidt's presumption that sibling sex is permitted when it involves no harm. This criticism is particularly apt in the case of participants who, rather than appealing to an emotion, invoke a norm when they run out of (other) arguments. For all we know, this is exactly the thing to do, which means that the reasoning that dumbfounded people engage in need not be biased.

These considerations derive support from research conducted by Kohlberg and by Neo-Kohlbergians, the tradition that Haidt vehemently opposes. Kohlberg distinguishes three levels of moral development: preconventional, conventional, and postconventional. In the (advanced stage of the) conventional level, people focus on maintaining social order. Neo-Kohlbergians argue that people rely in this stage on the Maintaining Norms Schema. In light of this schema, the commonly provided response "because you're not supposed to do that" may well be a perfectly legitimate response (Rest et al. 2000; cf. Prinz 2007: 34–35). The response ceases to be valid for those who reach the postconventional level. Perhaps the minority of the participants in the dumbfounding experiments that does change its mind have progressed to this level. They respond rationally in a way that fits their more advanced level of moral development.[10] Incidentally, Neo-Kohlbergians take cognitive dissonance to play an important role in explaining how people move from one stage of moral development to another (Rest et al. 2000). As I discuss in section 3.1, cognitive dissonance tends to generate self-justifying reasoning. This suggests that the Neo-Kohlbergians can also explain why people feel pressed to provide arguments in order to defend the moral judgments they form. The main conclusion to draw at this point is that the pessimistic conclusions Haidt draws from the dumbfounding findings do not follow. Rather than moral reasoning being biased and ineffective, it may just as well be that the different responses fit different stages of development.

In the remainder of this paper, I develop a line of criticism that is external to the dumbfounding experiments as such. I argue that the evidence that Haidt considers is too one-sided to warrant the general conclusions that he draws about moral reasoning. The argument below is inspired by Monin, Pizarro, and Beer who observe that the dumbfounding experiments concern only reasoning that is directed at others and maintain that in such situations 'emotions are primary when judging the shocking infractions of others' (ibid.: 104). They also argue, however, that '[r]easoning is primary when [people are] confronted with first-person dilemmas' (ibid.). Furthermore, they suggest that research into moral temptation and moral self-image will 'reflect a greater interplay between cognition and emotion' (ibid.: 105). I follow their lead by considering empirical studies about moral disengagement that concern the agent's own actions reasoning. A striking feature of these studies is that, rather than *post hoc* rationalization they concern *anticipatory* rationalization. These studies enable us to see that both affect and cognition can play destructive as well as constructive roles in moral reasoning.

Before turning to moral disengagement, I should mention a methodological problem that this kind of research faces. Haidt needs some criterion for characterizing certain forms of reasoning as bad. The problem is that none of us has special access either to which moral judgments are correct, or to the epistemic standing of particular lines of moral reasoning. Haidt

[10] In Hindriks (2014) I develop the more basic idea that the minority might be right. In terms of what I call 'the dominant minority argument', I argue that their reasoning reveals what role justification is supposed to play in relation to moral judgment. This remains true if the majority reasons in a way that is appropriate to their (lower) level of moral development.

solves this problem by using an indicator of bad moral reasoning or irrationality. Reasoning does not function well or is ineffective when an agent who has no arguments left to support her intuition does not change her judgment. As is implied by the preceding discussion, however, this is at best an imperfect indicator. Sometimes it is perfectly rational not to revise your views even if you are not aware of a good argument in favor of it (other than that the action violates a particular taboo). In what follows I rely on a closely related indicator of ineffective reasoning: not changing your judgment in the light of arguments against it. Just as the one Haidt uses, this indicator is imperfect. There is, however, no reason to think that it functions worse than Haidt's indicator. After all, it does not bode well for the quality of the agent's reasoning when she ignores counterarguments. Using this indicator, I argue that moral disengagement research reveals that moral reasoning is often causally *effective* and sometimes even *disinterested*.

## 3 Moral Disengagement

Moral dumbfounding occurs when people are questioned about the judgments they form about other people. Rather than with other-directed judgments, moral disengagement is concerned with self-directed judgments. Moral disengagement is triggered when someone suspects that the action she wants to perform is inconsistent with her moral standards. The negative affect that this thought causes motivates her to reason her way to the conclusion that the conflict was only apparent, as this will resolve the negative affect. Moral Disengagement Theory (MDT) is intimately related to cognitive dissonance theory. Appreciating the relation between them will help to understand MDT. It will also serve to appreciate a role that rationalization can play other than reaffirming prior judgments, that of facilitating immoral behavior. Once I have discussed cognitive dissonance and its relation to moral disengagement (sections 3.1 and 3.2), I go on to argue that moral disengagement research reveals that moral reasoning can in fact be *effective* – by resolving cognitive conflicts (section 3.3) – and *unbiased* – when the desire to perform the incongruent action is outweighed by other factors (section 3.4).

### 3.1 Cognitive Dissonance Theory

The core claim of Leon Festinger's (1957) Cognitive Dissonance Theory (CDT) is that a discrepancy between two cognitions causes the agent to experience psychological discomfort, which he seeks to reduce by changing one or more of his cognitions. Two cognitions are dissonant with each other when one of them entails the negation of the other. Such a cognitive discrepancy gives rise to psychological discomfort in the form of negative affect. Whereas Festinger used the term 'cognitive dissonance' both for the discrepancy and the psychological discomfort it causes, I follow Eddie Harmon-Jones (2000: 121, 136–37) and use it only for the discomfort. Festinger distinguishes three ways in which dissonance can be reduced: forming new consonant beliefs, reducing the number of dissonant beliefs, and increasing the importance of consonant beliefs. The agent can achieve this, for instance, by misperceiving, misinterpreting, or rejecting information she receives (Harmon-Jones and Mills 1999a).

Festinger and Carlsmith (1959) found evidence for CDT by considering the consequences of paying someone for telling others that a boring task was interesting. If belief were a matter of incentives and reinforcement, paying someone more would increase the degree to which the agent would in fact come to believe that the task he promotes is in fact interesting. CDT predicts the opposite. A substantial amount of money provides sufficient justification for someone to promote the task. When he receives a low amount, the agent experiences substantial cognitive dissonance and has to justify to himself the fact that he pretends that the task was interesting. He does this by changing his cognitions and coming to believe that the task was more interesting than he believed previously. Doing so involves reasoning that is motivated by the desire to relieve the psychological discomfort, which implies that reasoning triggered by cognitive dissonance is motivated reasoning.

Although (Mills 1999: 32) observes that he mentions anticipatory rationalization at some point, Festinger proceeded on the assumption that motivated reasoning always takes plays after the agent performed the act that creates the discrepancy. The evidence reveals, however, that biased reasoning also occurs prior to acting. Furthermore, such anticipatory rationalizations can facilitate behavior such as continuing to play a card game against professional gamblers while loosing money (Mills 1999: 36). Anticipatory rationalizations that facilitate behavior are central to moral disengagement.

Now why do cognitive discrepancies lead to cognitive dissonance? Elliot Aronson (1999) explains this in terms of the fact that people want to have a consistent and positive self-concept. Dissonance reduction enables the agent to maintain her self-concept by means of self-justification. In the example just discussed, an agent who lies about how interesting a task was will feel guilty. In order for him to restore the self-concept and continue to regard himself as a moral person, this guilt has to be resolved by means of rationalization (Aronson 1999: 111–12). Morality tends to be a significant dimension of the self-concept. The idea that people desire to maintain self-consistency also plays a central role in Bandura's MDT.

### 3.2 Moral Disengagement Theory

Moral disengagement theory is embedded in Bandura's social cognitive theory of self-regulation (1986, 1989, 2001). Social cognitive theory concerns the role that beliefs about the self play in generating actions. Self-beliefs such as self-efficacy beliefs feature in self-regulatory processes that link thought to action. Self-regulation involves a process of proactive guidance as well as a process of reactive adjustment, two processes that people need to balance. The standards by which people guide themselves include not only personal standards but also moral standards.

According to social cognitive theory, moral standards can be more or less central to someone's self-conception. And whether a standard impacts on someone's action depends on whether it is accessed at the relevant moment. This in turn depends both on the agent and on her environment: both on how central the standard is to her self-understanding and on whether environmental features make the standard accessible. When it is accessible the desire to maintain self-consistency may be frustrated due to an action that the agent wants to perform, as the action might conflict with the agent's moral standards. This results in psychological discomfort in the form of anticipatory guilt feelings. Although Bandura rarely if ever uses the term in writing, these feelings constitute cognitive dissonance.[11]

Moral disengagement occurs in situations in which someone is tempted to flout his own moral standards, and thereby to frustrate his desire to maintain self-consistency. Due to a real or imagined cognitive conflict between the envisaged action and the standard, this person experiences guilt feelings before even performing the action. The guilt feelings form a kind of anticipatory self-sanctioning, which makes the agent more aware of the perceived conflict. As discussed towards the end of section 3.1, such anticipatory guilt feelings constitute cognitive

[11] Aquino et al. (2005: 386) recognize that moral disengagement is a matter of reducing cognitive dissonance by means of rationalizations. Moore (2008) does not and maintains instead that disengagement pre-empts cognitive dissonance. This may well be due to a failure to acknowledge the possibility of anticipatory guilt feelings and anticipatory rationalizations.

dissonance. Cognitive dissonance, in turn, tends to trigger processes of moral reasoning that are aimed at resolving the conflict. Moral disengagement occurs at this point when, rather than enlisting them, people selectively disengage moral standards (Bandura 2001: 9). Moral disengagement, then, consists of anticipatory rationalization that serves to facilitate action that is apparently in conflict with the agent's moral standards. Rather than post hoc rationalization, such anticipatory reasoning is a matter of what I call 'ante hoc rationalization'. Note that moral disengagement can also take place when the agent has already started performing the relevant action, as in the speeding example. As it is aimed at getting rid of unpleasant guilt feelings, the reasoning or rationalization that the agent engages in is not only *motivated reasoning*, but also *self-serving*.

Having explained what it is, I now turn to the argumentative strategies that people rely on when morally disengaging, as well as to the empirical evidence that support people employ those strategies. Recall the argument presented in favor of speeding in section 1: everybody does it. This is an instance of what Bandura calls 'responsibility diffusion'. This strategy enables you arrive at a cognitive re-construal of the situation on which speeding does not appear very wrong and seems hardly blameworthy if at all. Bandura distinguishes seven other 'disengagement mechanisms', which are reasoning strategies aimed at undermining the thought that there is a genuine conflict between what you want to do and what your moral standards permit you to do. These reasoning strategies include well-known strategies such as blaming the victim ('she should not have worn such a revealing dress') and dehumanization ('they are filthy like dogs'). These mechanisms provide people with bad arguments, which give rise to the belief that some ostensible conflict with the moral principles to which they subscribe is illusory. In this way, her rationalizations provide apparent justification for performing the action, which serves to avoid self-condemnation and disengage the process of self-sanctioning.[12]

In one of the most elaborate empirical studies performed on this topic, Bandura et al. (1996) find that moral disengagement decreases helpfulness and cooperativeness and increases aggression and delinquency in children.[13] Bandura et al. (1996) studied 124 children most of whom were in their final year of elementary school or in the first few years of junior high school. Parents, peers, and teachers, as well as the children themselves contributed to the collection of data. At the beginning of the study, the children rated their acceptance of moral exonerations concerning, for instance, physical and verbal abuse, deception and theft. These ratings were used to measure moral disengagement. As it turned out, the children employed a wide range of disengagement mechanisms including in particular distorting the consequences (they construe injurious behavior as serving righteous purposes), displacement of responsibility (they disown responsibility for harmful effects), and dehumanization (they devalue those who are maltreated).

Questionnaires filled in by parents and teachers in combination with peer ratings were subsequently used to obtain data concerning prosocial and antisocial behavior. One of the findings is that moral disengagement fosters proneness to aggression, which in turn promotes delinquent behavior. What makes this study of particular interest for the purposes of this paper, however, is that it includes measures of affect and cognition: a hostile rumination measure and an irascibility measure, as well as a guilt and restitution measure. These measures provide for a

[12] Four of the eight mechanisms of moral disengagement that Bandura 1999, Bandura et al. 1996) distinguishes pertain to conduct and its consequences: moral justification, advantageous comparison, euphemistic labeling, and disregarding or distorting the consequences. Both displacement and diffusion of responsibility pertain to the agent. Dehumanization, and attribution of blame pertain to the victim.

[13] See Gini et al. (2014) for a meta-analysis that finds that moral disengagement significantly correlates with aggressive behavior among children and adolescents.

unique perspective on moral disengagement, as they facilitated testing the relation between moral disengagement and guilt feelings. In a structural equation model that Bandura et al. (1996) estimated, they found that moral disengagement decreases anticipatory guilt over transgressions. This confirms the role of cognitive dissonance in moral disengagement.

A study concerning prison personnel from maximum-security penitentiaries provides more evidence concerning the role of cognitive dissonance (Osofsky, Bandura, and Zimbardo 2005). In all likelihood executioners agree that prisoners on death row deserve the death penalty. One might think that they do not need to employ disengagement mechanisms in order to do their job. Nevertheless, they do employ a number of disengagement mechanisms including euphemistic labeling, advantageous comparison, and displacement of responsibility (ibid.: 379). As it turns out, actually killing someone requires disengagement even in the absence of a strict inconsistency with agential moral standards. Apparently, it is not easy to eradicate ethical qualms about killing even when the agent recognizes the case as an exception. This study includes experiential reports, which reveal that the executioners experience affective resistance to killing and rely heavily on emotion regulation to do their job (ibid.: 389). In the semi-structured interviews that were conducted, the executioners report managing their thought processes in order to enable themselves to go through with the killing.[14]

Empirical studies of moral disengagement that include measures of behavior are in fact rare. Most studies relate moral disengagement to attitudes such as support for war and terrorism (McAlister 2001). The studies that do concern behavior provide support for the idea that moral disengagement facilitates behaviors such as cheating, lying, and stealing (Detert et al. 2008; Moore et al. 2012). Detert et al. (2008) find that moral disengagement has predictive power over individual differences such as empathy and trait cynicism. Furthermore, the effect of individual differences is mediated by moral disengagement. Moore et al. go as far as claiming that the adult measure for moral disengagement they construct 'is the strongest individual difference predictor of unethical behavior to date', this in comparison to rivals such as Macchiavellianism, cognitive moral development, and empathy (2012: 40).

So how does disengagement relate to dumbfounding? There are some striking differences between the two. Whereas dumbfounding concerns other-directed post hoc reasoning, disengagement is a matter of self-directed ante hoc reasoning. Furthermore, the former is aimed at reaffirming pre-existing intuitions, whereas the latter is aimed at facilitating behavior that is at least apparently inconsistent with the agent's intuitions. Such inconsistencies are brought to light by anticipatory guilt feelings.

In spite of these differences, however, the post hoc rationalizations involved in dumbfounding may well be motivated by the same desire as disengagement – the desire to maintain a consistent self-concept as a rational and moral agent. After all, participants in the dumbfounding experiments are highly motivated to try and justify their point of view, as is suggested by the fact that they even invented victims in scenarios in which no one was actually harmed. In light of this, it may well be that at least some of the participants who are left without an argument experience cognitive dissonance. What remains to be done is see whether, in light of these similarities and differences, Haidt's claims about moral reasoning being ineffective and biased generalize to moral disengagement.

[14] Osofsky et al. describe this part of the interviews as follows: 'Also discussed was their emotional reactions in preparation, during, and after an execution, the extent to which they discuss their experiences with others, their perception of the stressfulness of the process, and the ways they tried to manage their stress in this situation. The interviews ranged from 30 min to 2 h with an average interview time of 1 h.' (2005: 381)

### 3.3 The Effectiveness of Moral Disengagement

Haidt claims that (private) moral rationalizations typically reaffirm prior judgments and is as such ineffective. Is moral disengagement effective in changing people's attitudes? The evidence suggests that cognitive dissonance in general and moral disengagement in particular does indeed cause genuine changes in beliefs. Moral disengagement generates beliefs by mechanisms that facilitate judgments and behaviors that conflict with the agent's moral standards. To be sure, it is not effective in the sense of overturning existing beliefs. Disengagement is triggered by the suspicion that there might be a conflict. This suspicion is replaced by the belief that there is no such conflict. Thus, disengagement prevents the agent from forming the belief that there is a genuine conflict between the action that she wants to perform and the moral standards to which she subscribes. In this sense, moral disengagement is *effective* in generating a belief.

Haidt argues that moral reasoning does little else than confirm prior intuitions. How do moral intuitions feature in moral disengagement? For one thing, moral standards can be seen as general moral intuitions when conceptualized as affect-backed rules (Nichols 2004). And the affect infused in those standards might give rise to the specific intuitions Haidt is concerned with. Those moral standards contribute to the thought that a particular action that the agent wants to perform is in conflict with them. This perceived cognitive discrepancy gives rise to cognitive dissonance or an anticipatory guilt feeling. This can be seen as an intuition that it might be wrong to perform the action under consideration. Moral disengagement overrules this intuition. In this sense, and in contrast to what Haidt claims, moral reasoning is *effective*. The upshot is that, *pace* Haidt, *intuition does not always carry the day. Reasoning can be powerful enough to overturn it*.

The extent to which this is reason for celebration is limited, however. Just as reasoning prior to dumbfounding, reasoning leading up to disengagement is motivated reasoning. Moreover, and in contrast to dumbfounding, disengagement is tailored to facilitate immoral behavior (supposing the agent's moral standards are adequate to begin with). And even if the standards are questionable, disengagement mechanisms such as blaming the victim and dehumanization are likely to result in immoral behavior.[15] Hence, even though *moral reasoning is often effective*, it tends to make the quality of moral judgments *worse*. As it will often lead to immoral actions, moral disengagement can in this sense be destructive. In the legal terms used by Haidt, the empirical findings discussed suggest that *reason functions as a lawyer that seeks to overturn an intuition, even when intuition should prevail*.

### 3.4 The Moral High Road: Unbiased Reasoning

Thus far, I have argued that moral reasoning can be effective in overturning an intuition and contributing to a judgment that the agent would not have formed in the absence of moral disengagement. What remains for me to argue is that moral reasoning leading up to action need not be biased. The underlying idea is that some people resist at least some of the time the temptation to employ one or more moral disengagement mechanisms. Rather than taking the low road of disengagement, they take the high road in that they take their guilt feelings

[15] The effects of the rationalization can be temporary. The pressure to rationalize can, for instance, recede when one has performed the act. In that case, the agent may come to see that she was blinded to the discrepancy. As Till Vierkant pointed out to me, however, it may also be that someone's rationalizations reveal his real moral commitments. It might be, for instance, that someone thinks he opposes speeding, but discovers during a process of rationalization that he in fact regards speeding as unproblematic. Perhaps he tells himself that the increase in risk is negligible, and comes to the conclusion that this in fact what he genuinely believes.

seriously. If they conclude that those feelings point them in the right direction, they will refrain from performing the action they initially wanted to perform. Thus, their reasoning is not susceptible to the bias involved in reasoning involved in moral disengagement.

In order to determine whether there is evidence for such unbiased reasoning, we need to know which factors influence the extent to which people disengage. Celia Moore (2008) discusses this in terms of what she calls ‘the propensity to morally disengage’. As it turns out, women have a lower propensity to disengage than men (Bandura et al. 1996; McAlister 2001; Detert et al. 2008). The same holds for younger people in comparison to older ones (Osofsky, Bandura, and Zimbardo 2005).[16] Furthermore, an individual trait such as cynicism correlates with a high propensity, whereas empathy and moral identity correlate with a low propensity (Moore 2008). In addition to this, awareness and discussion tend to decrease the extent to which people morally disengage (ibid.). Uhlmann et al. (2009) note in this connection that the extent to which people engage in motivated reasoning is limited: people only do so when it does not pose a threat to their self-image ‘as fair and objective judges’ (ibid.: 314).[17]

Recall that, according to Bandura, how central the moral standards are to the agent’s self-conception bears on moral disengagement. This idea is developed further in Moral Identity Theory (MIT). Karl Aquino defines moral identity in terms of someone’s moral commitments, including commitments to values, goals, traits and behavioral scripts (Aquino et al. 2009: 124, see also Aquino and Reed 2002). Someone with a strong moral identity regards it as more important than other self-concepts, including for instance her private, public and collective self-concept. People are, the thought is, more inclined to stick to their moral standards when their moral identity plays a central role in their thinking. In this way, someone’s moral identity can strengthen her motivation to stick to her moral standards and act accordingly. Moral identity facilitates action in accordance with the agent’s moral standards in particular when it is activated, conscious, or accessible.

How exactly does someone’s moral identity bear on moral reasoning? It can have an effect on someone’s moral judgment without her experiencing cognitive dissonance. Someone with a strong identity will usually be conscious of her moral commitments and will hardly be tempted to act in a way that conflict with them. In such cases, those commitments can feature in the content of her reasoning. Furthermore, they will carry a lot of weight. Someone with a weaker identity who has less access to her moral commitments might be influenced by it unconsciously. As discussed in section 1, they will then function as external factors and influence reasoning as such. Such a person will be more prone to temptation and thereby more likely to experience cognitive dissonance. Note that this leaves open the possibility that he ends up honoring her commitments. After all, someone’s moral identity might be weak enough to open the door for temptation, but strong enough to resist it. In such a case, moral identity helps to curb the temptation to violate internalized moral standards. The upshot is that, at least in theory, conscious moral thought strengthens moral motivation and is good for moral or at least self-consistent action.

So what does moral identity do when it influences thinking and decision-making? Aquino and Reed (2003) find that, in comparison to people with a weak moral identity, people with a strong moral identity are nicer to outsiders, i.e., to people who do not belong to their group. Such people are inclined to define the in-group in more expansive terms and report stronger

[16] Osofsky et al. (2005) also mention education and ethnicity as factors that correlate with moral disengagement. According to them, the less educated and Caucasians morally engage more than their contrast classes.

[17] Uhlmann et al. (2009) present evidence that people sometimes use moral principles selectively in order to support and rationalize their preferred conclusions. They also find, however, that people do not switch between principles across situations when a within-subject design is used. They conclude that the affective stakes must be high for motivated reasoning to occur when its low quality is particularly salient.

moral obligations towards out-group members. Furthermore, they are more favorable regarding relief efforts to assist them, and less accepting towards harming innocent out-group members in the course of military retaliation. Finally, they tend to donate to an out-group rather than an in-group member when presented with this choice. Thus, moral identity mitigates in-group favoritism and out-group hostility.

Aquino et al. (2005), Devert et al. (2008), and Moore et al. (2012) explore the connection between moral identity and moral disengagement. They find a negative correlation between them, which means that those with a strong moral identity are less prone to moral disengagement. The upshot is that, when an agent is aware of her moral identity, this strengthens moral motivation and is conducive to action that is self-consistent.[18]

These empirical findings strongly suggest that moral identity contributes to the quality of the moral judgment someone forms. Having an accessible moral identity helps people to stick to their moral standards. It serves to avoid rationalization aimed at obscuring conflicts between envisaged actions on the one hand and moral commitments on the other. Furthermore, they treat such conflicts as a sufficient reason for not performing the action. Moral identity research thereby reveals that the way in which someone arrives at a decision to perform an action need *not be biased*. Given adequate moral standards, moral identity can function in a constructive manner so as to support morally appropriate action. Someone with a relatively strong moral identity is likely to be conscious of her moral standards when they bear on an action that the agent wants to perform. To the extent that such a person engages in reasoning it will be unbiased in the sense that he objectively checks whether an action conflicts with self-imposed moral standards. The agent's reasoning functions as a disinterested *judge*. Thus, MIT supports the conclusion that *reason can be powerful even if intuition carries the day.*

### 3.5 The Cognitive Dissonance Model of Moral Reasoning

These findings concerning moral disengagement and moral identity can be used to add further detail to the Cognitive Dissonance Model of moral reasoning that I have introduced elsewhere (Hindriks 2014). The point of departure of this model is the assumption that people have a moral self-concept that encompasses the moral standards to which they subscribe (stage 0). Against this background, an agent might contemplate performing an action that might be prohibited by one of these standards. The thought that there might be such a discrepancy gives rise to cognitive dissonance or anticipatory guilt feelings (stage 1). This dissonance gives rise to moral reasoning aimed at resolving the cognitive discrepancy (stage 2). This reasoning can but need not be biased and self-serving. How likely it is that an agent's reasoning in support of a resolution will be biased depends on how conscious she is of her moral standards. This in turn depends on a number of factors including the strength of her commitment to the moral standards, i.e., her moral identity. Finally, she will come to a decision in which she does or does not remain true to her moral standards (stage 3).

## 4 Conclusion

Moral Disengagement Theory supports a picture of moral reasoning that differs from the picture that Haidt paints in three respects discussed respectively in sections 3.2-3.4. (1) Moral reasoning need not be post hoc reasoning aimed at reaffirming prior affect-based judgments. It

[18] The moral disengagement mechanisms that Aquino et al. (2005) explore are advantageous comparison, and moral justification. Detert et al. (2008) investigate minimizing or misconstruing harm to others.

can also be ante hoc reasoning aimed at resolving apparent inconsistencies brought to light by guilt feelings. (2) Such reasoning frequently leads to a change in someone's opinion, which means that moral reasoning often is effective. (3) Such reasoning can be unbiased, and sometimes fulfills the role of a disinterested judge.

Given the role that cognitive dissonance turns out to play in moral reasoning, it is not credible that affect and cognition are independent forces in moral reasoning, as Haidt suggests. Instead, affect and cognition often contribute to the formation of moral judgments together. Hence, any plausible theory has to be hybrid and combine features from sentimentalism and rationalism, as it does in Sentimental Rationalism, the position that I propose and defend in Hindriks 2014. According to Sentimental Rationalism, both cognition and affect can play a constructive role in the formation of moral judgments.

**Acknowledgments** I presented this paper at the workshop on Everyday Reasons held in Rotterdam, November 2013. In addition to the participants of that workshop, I thank Katrien Schaubroeck and two anonymous referees for very helpful comments.

## References

Aquino K, Reed A (2002) The self-importance of moral identity. J Personal Soc Psychol Rev 83:1423–40

Aquino K, Reed A (2003) Moral identity and the expanding circle of moral regard toward Out-groups. J Personal Soc Psychol Rev 84:1270–86

Aquino K, Reed A, Thau S, Freeman D (2005) A grotesque and dark beauty: How moral identity and mechanisms of moral disengagement influence cognitive and emotional reactions to War. J Exp Soc Psychol 43:385–92

Aquino K, Reed A, Freeman D, Lim VKG, Felps W (2009) Testing a social-cognitive model of moral behavior: the interactive influence of situations and moral identity centrality. J Pers Soc Psychol 97:123–41

Aronson E (1999) Dissonance, Hypocrisy, and the Self-Concept. In: Harmon Jones and Mills 103–26

Bandura A (1986) Social foundations of thought and action: a social cognitive theory. Prentice-Hall, Englewood Cliffs

Bandura A (1989) Human agency in social cognitive theory. Am Psychol 44:1175–84

Bandura A (1999) Moral disengagement in the perpetration of inhumanities. Personal Soc Psychol Rev 3:193–209

Bandura A (2001) Social cognitive theory: an agentic perspective. Annu Rev Psychol 52:1–26

Bandura A, Barbaranelli C, Caprara GV, Pastorelli C (1996) Mechanisms of moral disengagement in the exercise of moral agency. J Pers Soc Psychol 71:364–74

Detert JR, Trevino LK, Sweitzer VL (2008) Moral disengagement in ethical decision making: a study of antecedents and outcomes. J Appl Psychol 93:374–91

Ditto PH, Pizarro DA, Tannenbaum D (2009) Motivated Moral Reasoning. In: Bartels DM, Baliman CW, Skitka LJ, Medin DL (eds.), Psychology of Learning and Motivation 50: 307–38

Evans JSBT (2008) Dual-processing accounts of reasoning, judgment, and social cognition. Annu Rev Psychol 59:255–78

Festinger L (1957) A theory of cognitive dissonance. Stanford University Press, Stanford

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

Fine C (2006) Is the emotional Dog wagging its rational tail, or chasing It? reason in moral judgment. Philos Explor 9:83–98

Gini G, Pozzoli T, Hymel S (2014) Moral disengagement among children and youth: a meta-analytic review of links to aggressive behavior. Aggress Behav 40:56–68

Haidt J (2001) The emotional Dog and its rational tail: a social intuitionist approach to moral Judgment'. Psychol Rev 108:814–34

Haidt J (2012) The righteous mind: Why good people Are divided by politics and religion. Books, Pantheon

Haidt J, Björklund F, Murphy S (2000) Moral Dumbfounding: When Intuition Finds No Reason. Unpublished manuscript

Harmon-Jones E (2000) An update of cognitive dissonance theory, with a focus on the Self'. In: Tesser A, Felser R, Suls J (eds) Psychological perspectives on self and identity. American Psychological Association, Washington, pp 119–44

Johansson P, Hall L, Sikström S, Olsson A (2005) Failure to detect mismatch between intention and outcome in a simple decision task. Science 310:116–19

Kahneman D (2011) Thinking fast and thinking slow. Allen Lane, London

McAlister AL (2001) Moral disengagement: measurement and modification. J Peace Res 38:87–99

Mercier H, Sperber D (2011) 'Why Do Humans Reasons? Arguments from an Argumentative Theory'. Behav Brain Sci 34:57–111

Mills J (1999) Improving the 1957 Version of Dissonance Theory. In: Harmon Jones and Mills 25–42

Monin B, Pizarro DA, Beer DS (2007) Deciding versus reacting: conceptions of moral judgment and the reason-affect debate. Rev Gen Psychol 11:99–111

Moore C (2008) Moral disengagement in processes of organizational corruption. J Bus Ethics 80:129–39

Moore C, Detert JR, Trevino LK, Baker VL, Mayer DM (2012) Why employees Do Bad things: moral disengagement and unethical organizational behavior. Pers Psychol 65:1–48

Nickerson RS (1998) Confirmation bias: a ubiquitous phenomenon in many guises. Rev Gen Psychol 2:175–220

Nisbett RE, Wilson TD (1977) Telling more than We Can know: verbal reports on mental processes. Psychol Rev 84:231–59

Osofsky MJ, Bandura A, Zimbardo PG (2005) The role of moral disengagement in the execution process. Law Hum Behav 29:371–93

Paxton JM, Ungar L, Greene JD (2010) Reflection and reasoning in moral judgment. Cogn Sci 36:163–77

Pizarro DA, Bloom P (2003) The intelligence of the moral intuitions: comment on haidt (2001). Psychol Rev 110:193–196

Prinz J (2007) The Emotional Construction of Morals. Oxford University Press, Oxford

Rest J, Narvaez D, Thoma SJ, Bebeau MJ (2000) A Neo-kohlbergian approach to morality research. J Moral Educ 29:381–95

Richardson HS (2013) Moral Reasoning. The Stanford Encyclopedia of Philosophy (Spring 2013 Edition), Edward N. Zalta (ed.), URL=<http://plato.stanford.edu/archives/spr2013/entries/reasoning-moral/. Accessed 9 Mar 2013

Saltzstein HD, Kasachkoff T (2004) Haidt's moral intuitionist theory: a psychological and philosophical critique. Rev Gen Psychol 8:273–282

Sauer H (2011) Social intuitionism and the psychology of moral reasoning. Philosophy Compass 10:708–21

Sneddon A (2007) A social model of moral dumbfounding: implications for studying moral reasoning and moral judgment. Philos Psychol 20:731–48

Uhlmann EL, Pizarro DA, Tannenbaum D, Ditto PH (2009) The motivated Use of moral principles. Jud Decision Making 4:476–91

Wilson TD (2002) Strangers to ourselves: discovering the adaptive unconscious. The Belknap Press of Harvard University Press, Cambridge