Philosophical Explorations
Vol. 15, No. 3, September 2012, 255–275

WINNER OF THE PHILOSOPHICAL EXPLORATIONS ESSAY PRIZE 2012

Educated intuitions. Automaticity and rationality in moral judgement

Hanno Sauer*

Institute for Philosophy, University of Leiden, Matthias de Vrieshof 4/Witte Singel 25, 2311 BZ Leiden, The Netherlands

Moral judgements are based on automatic processes. Moral judgements are based on reason. In this paper, I argue that both of these claims are true, and show how they can be reconciled. Neither the automaticity of moral judgement nor the *post hoc* nature of conscious moral reasoning pose a threat to rationalist models of moral cognition. The relation moral reasoning bears to our moral judgements is not primarily mediated by episodes of conscious reasoning, but by the acquisition, formation and maintenance – in short: education – of our moral intuitions.

Keywords: moral judgement; moral reasoning; moral intuition; moral rationalism; Jonathan Haidt

Thought is, indeed, essential to humanity. It is this that distinguishes us from the brutes. In sensation, cognition, and intellection; in our instincts and volitions, as far as they are truly human, Thought is an invariable element. (Hegel, *The Philosophy of History*)

1. Introduction

Moral judgements are based on automatic processes. Moral judgements are based on reason. In this paper, I argue that both of these claims are true, and show how they can be reconciled.

To some, this might seem like a trivial endeavour, and indeed a superfluous one. But this is not so. Here is why: recent years have witnessed a revolution in the empirical psychology of moral judgement and cognition. After decades of rationalist dominance under the auspices of a cognitivist paradigm (Kohlberg 1969), moral psychology has undergone an emotionist turn (Haidt 2007; Haidt and Kesebir 2010; Sauer 2011a). Contrary to the emphasis on moral reasoning typical for rationalist models of moral education and development, studies on mental disorder and brain lesion suggest that emotions are critically necessary for moral judgement (Damasio 1994; Blair 1995; Blair, Mitchell, and Blair 2005; Koenigs et al. 2007). Research on emotion manipulation and mood induction provides evidence for the sufficiency of emotions for moral attitudes (Wheatley and Haidt 2005;

*Email: h.c.sauer@hum.leidenuniv.nl

ISSN 1386-9795 print/ISSN 1741-5918 online

http://dx.doi.org/10.1080/13869795.2012.706822
http://www.tandfonline.com

ValdeSolo and DeSteno 2006; Schnall et al. 2008). Evidence from neuroimaging suggests that an important part of moral cognition is shaped by automatic emotional reactions (Greene et al. 2001, 2004; Singer 2005; Sauer, forthcoming). On top of that, recent studies seem to have shown that we arrive at moral verdicts on the basis of quick, often emotionally charged intuitions, rather than episodes of controlled reasoning and conscious deliberation (Haidt 2001; Uhlmann et al. 2009).

These developments are part of a more general trend in the psychology of cognition and action to study the pervasive automaticity of human judgement and behaviour. It seems that to a surprising extent, judgement formation and action are based on processes that remain largely unconscious (Bargh 1994; Bargh and Chartrand 1999; Wilson 2002; Dijksterhuis 2004, 2006). People often do not have access to what really drives their behaviour (Wegner 2002; Sie 2009), and are oblivious to what triggers a certain judgemental or behavioural response (Nisbett and Wilson 1977, 1978; Langer, Blank, and Chanowitz 1978). Emotional and intuitive processes of the kind moral psychologists are interested in are subsets of automatic processes.[1] Empirically minded philosophers have thus taken the aforementioned findings to support a broadly sentimentalist account of moral judgement and cognition, and to provide the building blocks of an empirical refutation of rationalist models of moral judgement (Nichols 2004; Prinz 2007). Moral judgements are not, these philosophers argue, based on critical reflection and proper weighing of reasons, but on uncontrolled, emotionally charged states of intuitive (dis)approval. Call this the *automaticity-challenge*.[2]

In this paper, I argue that this challenge to rationalist accounts of moral judgement can be met. More precisely, I aim to show that the automaticity challenge rests on problematic, and ultimately mistaken, assumptions concerning the nature of automatic mental processes. In a nutshell, I maintain that emotionists about moral judgement, broadly construed, make a plausible case for the automaticity of moral judgement, but fail to show that there is anything about the automaticity of a mental process that excludes it as a fit candidate for being a rational process as well.

Call the claim that a mental process cannot be both automatic and rational, or, to put the same point differently, that automaticity excludes rationality, the *incompatibility thesis.* Despite the obvious counterexamples, the thesis is widely accepted.[3] We can, I argue, undermine this thesis by finding traces of reason in our emotional and intuitive reactions themselves. Since we have no metaphysical guarantee that our feelings and intuitive reactions will live up to that demand, their normative quality will have to be secured indirectly. My suggestion is to start from the notion of a "second nature", and draw on the idea that emotions and intuitions can be *educated*. In the process of moral upbringing, rational grounds become embodied in our intuitive thinking. The following paper makes a conceptual as well as empirical case for the claim that moral judgements are based on educated intuitions.

My paper has four parts. In the first part (1), I will set out the automaticity challenge in a little more detail and discuss the most important empirical evidence for the automatic character of moral judgement, using Jonathan Haidt's social intuitionist (henceforth: SI) model as an example. I shall argue that the SI model rests on assumptions which are complementary to the incompatibility thesis, and that it mistakenly ties the rationality of a mental process to the fact that it is or is not conscious. I shall then discuss recent philosophical attempts to free up conceptual space for the possibility of automatic yet rational forms of cognition, and explain how habits and the concept of a "second nature" figure in those attempts (2). I show that the incompatibility thesis has been endorsed by many philosophers of action as well, and why this move has seemed attractive to many. But

once we see that there is nothing conceptually dubious about automatic yet rational processes – and habits are the prime example here – we can provide empirical evidence for their significance for human cognition in general and moral judgement in particular. Part (3) explains in greater detail the distinction between *post hoc* reasoning and moral confabulation, and shows how the concept of an education of moral intuitions can help draw this distinction. Finally, I distinguish two different kinds of education of the intuitions (4), *ex ante* and *ex post* education, and show how educated intuitions can account for the central elements of a normative picture of human moral judgement and agency while leaving the central intuitionist insights intact. The relation moral reasoning bears to our moral judgements is not primarily mediated by episodes of conscious reasoning, but by the acquisition, formation and maintenance of rationally acquired – in short: educated – moral intuitions.

2. The automaticity challenge: intuition, reason and moral judgement

Faculty psychology is back with a vengeance: it is now becoming increasingly popular to understand the psychological underpinnings of human judgement and decision-making in terms of a dual process model of cognition (Kahneman 2003). Proponents of dual process models hold that judgement and behaviour are based on two mental subsystems (often referred to as Systems I and II) which are different in at least four important respects and work upon entirely different principles (for an overview, see Evans 2003, 2008). System I processes are said to be evolutionarily old (age), operate quickly and effortlessly (speed), their workings remain un- or preconscious (accessibility) and they process information holistically and often emotionally hot (mode of function) as opposed to the evolutionarily recent, controlled and effortful, conscious, analytical step-by-step reasoning characteristic of System II.

Which of the two systems is responsible for moral judgement? Simple – and, as I will argue below, overly simple – rationalist models[4] seem to suggest that moral judgement is based on deliberate reflection and the careful weighing of reasons: "one briefly becomes a judge, weighing issues of harm, rights, justice, and fairness, before passing judgment [. . .]. If no condemning evidence is found, no condemnation is issued". (Haidt 2001, 814). Haidt's SI model of moral judgement and reasoning challenges the empirical accuracy of that picture. It does so on the basis of two types of evidence:

(i) *Intuitive primacy*: Haidt stresses the fact that people generally arrive at their moral verdicts too quickly for it to be possible to engage in explicit reasoning upfront. Rather, the process of moral judgement formation works much more like perception. People simply "see" whether a particular action is morally wrong or not, and base their judgement on what intuition tells them. This process is often accompanied by quick flashes of emotional (dis)approval (Wheatley and Haidt 2005; Schnall et al. 2008).

(ii) *Moral dumbfounding*: That people do engage in explicit moral reasoning is an uncontroversial fact. The SI model, however, questions that these episodes of conscious deliberation are causally efficacious. Experimental vignettes that trigger a strong moral intuition yet render most possible justifications for moral condemnation pointless have shown that people do not suspend and/or change their moral beliefs if no appropriate reasons can be found. Rather, they enter a state of moral dumbfounding, the inability to articulate any good reasons for the moral intuitions they have.

These findings are taken to suggest that moral judgements are based on automatic System I processes. The intuitive primacy in moral judgement formation is said to give us reason to think that episodes of moral reasoning subjects engage in come *post hoc*. They provide mere rationalizations of the moral intuitions subjects already have. Call this the *post hoc thesis* (PHT):

> *PHT*. Conscious moral reasoning comes after the fact.

But the SI model makes an even stronger claim. Haidt argues that moral reasoning not only comes *post hoc*, but that typically, it is utterly superfluous in bringing about moral judgements (Saltzstein and Kasachkoff 2004; Clarke 2008; Musschenga 2008). The justifying reasons subjects might have play no role whatsoever in the formation of their beliefs (Prinz 2007, 31). Moral reasoning, on that account, is like confabulation (Hirstein 2005): it does not verbalize the reasons that actually lead subjects to adopt certain judgements. People do not make moral judgements because they reason from legitimate, morally salient considerations and well-founded principles to a judgement, but choose the reasons and principles *post hoc* that best justify the judgement they want to end up with anyway (Uhlmann et al. 2009). This can be called the *confabulation thesis* (CT):

> *CT*. Moral reasoning is a matter of causally ineffective confabulation.

The automaticity challenge – one prime example of which is Haidt's SI model – is supposed to attack the main tenets of rationalism concerning the psychology of moral judgement. Rationalism holds, among other things, that the normative reasons that justify people's moral judgements are also the causally effective explanatory reasons for why they make those judgements. Call this the *effectiveness thesis* (ET):

> *ET*. The justifying reasons subjects have for their (moral) judgments figure in true causal explanations for why they hold these judgments.

This is the main claim rationalist models of the psychology of moral judgement are committed to. It is important to carefully distinguish the *PHT* from the *CT*, because only the latter denies the *ET*. The former is compatible with the fact that in episodes of *post hoc* reasoning, people make explicit the reasons that really did play an effective role in how they arrived at their assessment of a given scenario, albeit unconsciously. The CT denies that this is the case.

The challenge the SI model poses towards the rationalist position is based on the idea that if one can show that moral judgements are largely based on automatic processes (emotional reactions and intuitive cognition), one has *eo ipso* shown that the ET does not hold – which means that the episodes of reasoning people do engage in must be confabulatory, as the CT states. Recently it has been shown, however, that this move is not as innocent as it may seem, but only follows on the basis of two further tacitly made assumptions (Horgan and Timmons 2007; Sneddon 2009; Sauer 2011b). For one thing, the SI model holds that unless the following accessibility requirement (AR) is fulfilled, the ET must be false:

> *AR*. A (moral) judgment M made by subject S counts as being based on reason only if S has internal access to the set of reasons {q, r, s, . . ., n} that justify M.

Second, and perhaps more importantly, the SI model not only requires that for moral reasons to be causally effective in bringing about a moral judgement, these reasons must be accessible to the judging subject; it also holds that for reasons to be operative in moral judgement formation, conscious reasoning on part of the judging subject must immediately precede the judgement: the "key part of the definition of reasoning is that it has steps, at least a few of which are performed consciously" (Haidt 2001, 818). We can call this the *causality requirement* (CR):

> *CR*. A (moral) judgment M made by subject S counts as being based on reason only if conscious consideration of the set of reasons {q, r, s, . . ., n} that justify M causes S to hold M.

The empirical evidence mentioned above – the intuitive primacy and the phenomenon of moral dumbfounding – clearly shows that in many cases, neither the AR nor the CR are met.

A closer look reveals that the two requirements are in fact restatements of the incompatibility thesis. They identify the rationality of a process with the accessibility and conscious awareness characteristic of controlled System II processes; conversely, the requirements rule out – by conceptual *fiat* alone – that automatic processes can be based on reason.

The incompatibility thesis is not a trivial assumption emotionist models of moral judgement could easily do without. It is a core commitment that does major work in their anti-rationalist argument. In fact, the thesis is so central that Jesse Prinz and Shaun Nichols, for instance, have straightforwardly included it in their definition of moral rationalism, which they characterize negatively as a position that holds that moral judgement can occur "in the absence of [. . .] emotions" (Prinz and Nichols 2010, 116). Emotions are a paradigm example for automatic System I processes, and rationality is defined in terms of the absence of these processes. For this definition to make sense, it must be assumed that a process cannot be both automatic and rational at the same time.

In what follows, I shall argue that we can account for the ET in a way that does justice to (i) the largely automatic and intuitive character of moral judgement and (ii) the *post hoc* nature of moral reasoning, but (iii) does not entail that moral reasoning is confabulatory and thus causally ineffective. I will argue that the automaticity of moral judgement can be squared with its rationality just in case moral judgements are based on patterns of moral reasoning that have, through a process of moral education, become habitual. I will show why it is legitimate to think of habits – acquired automatic processes – as processes that can be placed in the space of reasons, and I will make good on the claim that moral judgements are in fact based on such educated intuitions.

3. Habits and practical reason

3.1 *What are habits?*

Habits are "learned dispositions to repeat past responses" (Wood and Neal 2007, 843). One should understand the concept of a *disposition* used here as a placeholder for automaticity: habits are behavioural patterns whose execution is triggered in certain circumstances and becomes automatic over time. Habitualization can be plotted as an asymptotic curve, representing the relation between the number of repetitions of a given piece of behaviour and its degree of automaticity (Lally et al. 2010, 1002).

The very fact that automaticity is distinctive of habits, however, has led many modern philosophers of action to ignore their significance for a comprehensive theory of practical

reason. However, this has not always been the case. Schiller wanted to reconcile the strict Kantian dichotomy between "duty" and "inclination" in his concept of a beautiful soul, a person whose character traits lead her to automatically act in accordance with what is morally required (Schiller 2004); Hegel, at times, makes the even stronger claim that habituality is necessary for moral action, and that explicit deliberation about what to do is (at least potentially) "unethical [unsittlich]" (Hegel 1986, 323); and early twentieth-century philosophical anthropologists (Gehlen 1940; Scheler 2007) did not grow tired of emphasizing how important individual habits and structured, intersubjective habits – conventions and institutions – are for creatures with the degree of cognitive flexibility and plasticity humans enjoy. Recent findings support this insight. It has been established that one of the key benefits of habitual action is its cognitive efficiency: "Habits potentially free people to engage in other kinds of thoughtful activities such as rumination of past events and planning for future activities" (Wood, Quinn, and Kashy 2002, 1295). It is thus not only inevitable, but also pragmatically rational for agents to rely on habits.

3.2 *Intellectualism and the "reasons theory"*

Why is it, then, that habitual action has been so widely ignored in philosophy of action, or that the very possibility of genuine habitual action has been rejected in the first place? The "three categories of phenomena" which, according to Velleman, "philosophy of action must [. . .] account for", are "mere happenings, mere activities, and actions" (2000, 4). Neither the first nor the third category seems suitable for habitual behaviour. But the residual category of "mere activities", which consist in a "partial and imperfect exercise of the subject's capacity to make things happen" (Velleman 2000, 4) also does not seem quite up to the job.[5] Most habits allow for a degree of reflexive monitoring (Giddens 1984, 5ff.; Pettit 2001, 39) and intervention control (Pollard 2005b) that distinguishes them from mindless finger-tapping and compulsive nail-biting. (See Ryle 1949 and Winch 1958 for earlier accounts that have been similarly dismissive of habits.) That agents do not consciously initiate most of their everyday habitual actions – from brushing their teeth to making coffee – does not mean that they are not fully involved in these actions, and that we do not hold them fully accountable for them. We thus want to be able to distinguish (i) full-blown actions from mere behaviour and (ii) the reasons for which a subject merely *could have* acted from the reasons for which the subject *really did* act. (i) is the task of providing a theory of what genuine action is; (ii) is the task of showing how reasons-explanations of those actions work.

Gert (2003) and Pollard (2003, 2005b) have recently shown that philosophers of action as diverse as Donald Davidson, Jonathan Dancy, Warren Quinn, John McDowell, Joseph Raz and Thomas Scanlon have tried to achieve the above two tasks by making one and the same assumption, namely, that genuine actions, as opposed to mere mindless behaviour, are done for reasons; and that an agent only really does something *for* a reason if she acted "in the light" of that reason.

Now the way these philosophers spell out this "in the light of"-relation subjects need to bear to their reasons is strikingly similar to the AR and the CR. Remember that to get their challenge to rationalism about moral judgement off the ground, many empirical moral psychologists assume that for a judgement to be based on reason, conscious consideration (CR) of internally accessible (AR) reasons must bring a subject to accept a certain moral judgement. Philosophers of action, on the other hand, hold that for an action to be performed "in the light of" reasons, these reasons must somehow exert their motivational

and justificatory force upon the acting subject; and the way reasons do this is by being "present to the agent's consciousness" (Dancy 2000, 129) or by "reveal[ing] the favourable light" of the "projected action" to the agent (McDowell 1978/1998, 79). In order "for a consideration to be an operative reason for me", Scanlon writes, "I have to believe it" (Scanlon 1998, 65). Pollard and Gert have coined different names for this requirement, from "judgment thesis" (Gert 2003) to "conception constraint" or simply "reasons theory" (Pollard 2005b) of rational action. According to the reasons theory, task (i) is easily achieved: genuine actions are guided by reasons, whereas mere behaviour is not much more than a brutish, mindless response to internal or external stimuli. And the theory has no problem with task (ii), either: an agent's real reasons, and the ones which do not only justify a subject's action from an outside perspective, but explain her actions in terms of what she saw in it, are those that the agent is consciously aware of.

In the psychology of moral judgement, the challenge to the ET is based on the very same intellectualist assumption philosophers of action make: for reasons to be operative in my arriving at a moral verdict, and thus for these reasons to render my judgement rational, I must be aware of them, and consciously consider them in advance. But, so the challenge goes, as an empirical matter of fact, this is not what happens. Most of the time, my moral judgements are based on quick and effortless intuitions (Haidt's "intuitive primacy") and episodes of reasoning come after the fact (the *post hoc* thesis); therefore, reasons are not causally effective in how people arrive at their judgements (the CT). It is clear, however, that the automaticity challenge only delivers this anti-rationalist result on the basis of the aforementioned intellectualist assumptions concerning the nature of reasons, and what it means for them to be operative in subjects' judgement and behaviour. Luckily, we do not have to accept these assumptions.

3.3 *Reason, habits and second nature*

Can we sketch a theory of action and practical reason that avoids this implausible dismissal of automatic action, leaves room for its rationality and thus helps to meet the *automaticity challenge* to moral rationalism? Can we develop an alternative picture of how reasons can become effective in moral judgement in a way that efficiently bypasses subjects' conscious awareness?

I have argued earlier that there are two different ways in which this can be done. The direct way is to argue for cognitivism about automatic processes, preferably emotions, but I have explained why one should resist the temptation to choose this strategy. The indirect way is to pay attention to the malleability of automatic processes: if one can find evidence for complex cognitive processes which used to be executed by System II, but – over the course of habitualization and education – have become automatic over time, we have no reason to think that these processes have changed as far as their rationality is concerned. The idea here is structurally similar to the *parity principle* made prominent in the context of the "extended mind" hypothesis (Clark 2010). If we would not hesitate to classify, on the basis of its functional characteristics, an event or operation as mental if it went on "in the head", but as it happens, it does not, then we should not hesitate to classify the operation as mental. Similarly, we can propose a *parity principle for automatic processes* (PPA):

> *PPA*. If we would not hesitate, on the basis of its functional characteristics, to call a process "rational" were it performed consciously and effortfully, but, as it happens, it has become habitual and automatic over time, then we should not hesitate to call it "rational".

This principle suggests that there is no reason to think that consciousness is a necessary condition for rationality, or that the two are congruent.

Stretching the concept of rationality so widely that the notion of automaticity can be squeezed in will not do here, and it will not convince those who have different conceptual intuitions. Therefore, my argument hinges on the idea that if, as Kahneman and Fredrick (2002) put it, "complex cognitive operations eventually migrate from System 2 to System 1 as proficiency and skill are acquired" (51), and if there is empirical evidence that this is the case in the domain of moral judgement and reasoning, then we have no reason to think that simply because the *modus operandi* of a mental process has changed from "controlled" to "automatic", its status as a rational or non-rational process must have changed as well.

In recent years, this idea – the idea that a great many of automatic processes, especially habits, can qualify as rational – has become increasingly popular. John McDowell has arguably most influentially championed the idea. He argues that habits are the key to bridging the gap between the space of causes and the space of reasons, and has defended the view that conceptual capacities – whose involvement seem to be necessary to anything that can be subject to rational justification – can become embodied in the "habits of thought and action" which become "second nature" (McDowell 1994, 84). Following McDowell, Sabina Lovibond argues that the capacity to make moral judgements and to act on them essentially consists in the initiation into a culture, and the acquisition of habits and traits that come with the participation in social practices:

> Over time, our participation in these activities [. . .] gives rise to a "second", or acquired, nature. This second nature is manifested in behaviour which, though learned, is largely unreflective [. . .]; and which if we do make it into an object of reflection, usually produces in us a sense of inevitability. From one point of view, the dispositions that constitute our second nature are passive, for they are dispositions to be affected in a certain way: ideally, to register the "proper force and necessity" of reasons for judgment [. . .]. However, it is a feature of human socialization [. . .] that one is led not just to receive and process sensory input from one's environment, but to recognize the state of the world as imposing rational constraints on one's thinking. (Lovibond 2002, 25f.)

One of the main goals of moral upbringing is to equip subjects with the capacity to make good moral judgements. And if moral judgements, as the empirical evidence suggests, largely depends on intuitive processes, then an understanding of how the education of moral intuitions works is of foremost importance.

Before I turn to the empirical question how the education of (moral) intuitions really works, and which types of education of the intuitions there are, let me briefly discuss the problem what it is that renders a habit rational from yet another perspective. The rejection of the "reasons theory" of rational action as well as the ARs and CRs on rational moral judgement does not relieve the friends of habitual actions from the obligation to account for the difference between action and mere behaviour (task (i) from above) and between the reasons that really were effective in bringing about an action from those that merely could have been (task (ii) from above). How can this be done?

4. From *post hoc* reasoning to confabulation

4.1 *Rational habits: making it explicit or confabulation?*

Pollard (2005a) has argued that the concept of *Bildung* (education) makes it possible to "naturalize" the space of reasons in a way that allows us to reconcile the habituality of

an action with its rationality. It is clear that habitual judgements and actions do not satisfy the AR and the CR, because "habitual behaviours [. . .] are automatic, which is to say that they do not seem to be preceded by deliberation of any sort" (Pollard 2005a, 74). But how can there be reasons in action if the requirements of the "reasons theory" are not met? How can reasons be operative, if they are not in any way "present" to the subject? In my idiom: how can the ET hold if AR and CR are not satisfied? There are two alternative ways to solve this problem.

A first way to explain away this tension in the concept of habitual action would be this: habitual action does not draw on controlled cognition. To a large extent, the actualization and execution of habits goes on at a subconscious level that is typically not accessed by the acting subject. This does not mean, however, that the agent's practical reasons cannot be present *on this subconscious level*. On that account, subjects can act on reasons in ways that are not mediated by conscious awareness and explicit deliberation. But this is not to say that these play no role at all:

> On the contrary, her reasons could be brought to consciousness were she, or somebody else, to enquire about why she did what she did. The key thought is that her reasons are, as it were, already in place when she acts, and are thereby ready to be discovered afterwards should the need arise. On this conception of reason giving, success is marked by these hitherto subconscious states being made explicit. (Pollard 2005a, 79f.)

The above quote makes clear why, as I have argued above, we must carefully distinguish the *post hoc* from the *CT*. Episodes of *post hoc* reasoning need not be confabulatory. In most cases, we have no reason whatsoever to think that when people give reasons for their moral beliefs, they are confabulating. Despite the *post hoc* character of moral reason*ing*, genuine moral reasons are effective in how people arrive at their verdicts: they figure effectively in the acquisition, formation and maintenance (that is, the education) of subjects' moral intuitions, and make a psychologically real difference to people's moral beliefs. Effective moral reasoning requires nothing more than this.

A second way to explain the effectiveness of reasons in habitual action is this: we can show, the suggestion goes, that an action is rational if the acting subject can come up with a narrative that reconstructs how a given piece of behaviour fits into the agent's overall world view and character. What decides whether an action is based on reasons or not, then, is not whether these putative reasons have been always already in place, albeit subconsciously, but whether the behaviour in question can plausibly be made to fit into and cohere with the agent's overall system of desires, intentions, goals and values (Pollard 2005a, 80). But this second option, I argue, is not what we are looking for.

4.2 *Rational habits: education and goal-dependency*

Which of the two solutions is the correct one? Pollard seems to think that both options are equally legitimate. But this is clearly not the case, and the reason for this is that the second solution does not do the job it has been hired for: it conflates the distinction between *post hoc* reasoning and confabulation, and makes episodes of explicit reason-giving which are accurate indistinguishable from cases in which subjects merely come up with a "coherent fiction" (Snow 2006, 559). Social psychologists have shown that sometimes we construct stories that seem to make sense and come up with reasons for our responses which, though plausible, are demonstrably false (Nisbett and Wilson 1977, 1978). So a version of the first solution must be true, if such a thing as genuine, non-confabulatory *post hoc* reasoning is to be possible. At this point, however, it remains obscure what it means that reasons are

subconsciously present at the time a subject acts, waiting to be made explicit. Not surprisingly, the concept of education sheds light on this issue as well.

Remember that habits are behavioural patterns which become, through repeated enactment, routinized over time. The "parity principle" proposed above suggests that this process is rationality-preserving, because what changes in this process of automatization is not the *nature* of the process, but merely its *modus operandi*. But why is it that agents acquire some habits, and not others? Is this a matter of mere luck? And, perhaps more importantly, why is which habits are acquired and which are not a completely random thing? Consider this example: I am riding home from work on my bike, and I do so, as it were, on autopilot. My unlocking the bike, my leaving the lot, my using the handle bar are all entirely automatic. But, of course, this sequence of automatic actions is not pointless, and it is not irresponsive to the tiny environmental features that change every day. Rather, these atomic actions all serve my goal – arriving at home. In fact, that I have this goal is why I have developed this particular sequence of habitual actions in the first place:

> The reason that these automatic, habitual actions are performed is to serve the agent's chronically accessible goals. Thus, habitual, automatized goal-dependent actions are purposive. The agent's reason for acting – to serve a chronic goal – is not present to her consciousness at the time of acting. Nevertheless, it is operative in her psychological economy. It is a motivating factor that explains her actions. (Snow 2006, 552; see also Snow 2010).

In the course of an agent's education, her practical reasons become embodied in her automatic judgemental and behavioural responses. These reasons are thus both internal to a subject's psychology and external to her conscious awareness and initiation control at a given point in time. Making explicit the reasons that brought me to adopt my after work routine – namely that I want to go home – is an enterprise that comes entirely *post hoc*. But it need not be confabulatory; indeed, it would be ludicrous to suppose so.

4.3 *Varieties of post hoc reasoning*

In meeting the automaticity challenge against rationalism about the psychology of moral judgement, the rationalist must explain why there is nothing intrinsically dubious about *after the fact* justifications. But she must be careful to avoid saying that there is nothing even potentially dubious about *post hoc* reasoning. Sometimes, people irrationally hold on to their convictions; sometimes, reasons are just rationalizations; and sometimes, people are guilty of confabulating.

This distinction is important, and it must be drawn neatly. The idea that moral intuitions can be educated seems to allow for reasons to become effective somewhere more "upstream": it is alright, I have argued, for intuition rather than reasoning to be the proximate cause of moral judgement, as long as the reasons one cites after an intuition is arrived at did figure in the acquisition, formation and maintenance of that intuition. But the more "upstream" these reasons become causally efficacious, the more the distinction between genuine *post hoc* reasoning and mere confabulation is blurred. This shows that more needs to be said about the distinction, and it also shows that the concept of education alone, though important, is not sufficient to draw the line between "good" *post hoc* reasoning – which cites the reasons that helped educate your intuitions – and "bad" confabulation – which cites reasons that are entirely disconnected from the causal genesis of your judgements.

There are two different kinds of causally ineffective *post hoc* reasoning, and it is an intricate question at what point they become confabulatory. Either, one comes up with reasons that really do justify one's judgement (they are "good" reasons) yet these did not play a causal role. You used to think that torture can, under certain circumstances, be right. Now you have changed your mind – you think it is always wrong to torture people because it violates their dignity as persons. Let us suppose, for the sake of the argument, that this judgement is correct, and it is correct for precisely those reasons. But those good normative reasons are not the effective motivating reasons for your change of mind. Rather, there is this girl you fell in love with who is very strongly against torture, and you started to share her opinion out of affection. This causal path led you to the correct belief and reasoning – that torture is wrong because {q, r, s, . . ., n} – just by accident. Or, one comes up with a bunch of bad reasons – reasons that are fallacious or factually inaccurate – but those were not the genuine causes of your judgement either. You believe that cannibalism is wrong because it is unnatural. However, not only are your reasons bad, they are also not causally responsible for why you believe that cannibalism is wrong. You just have this spontaneous response of disgust and horror towards it, which you are trying to rationalize after the fact. Should we treat these two cases differently? Or are they equally confabulatory?

It seems to me, perhaps somewhat unsatisfyingly, that whether the former type – the one that is illustrated by the torture example – ought to be considered confabulation is a matter of degree. That does not mean, however, that there are no criteria for this degree: it is how recalcitrant your willingness to confabulate reasons is, and how willing you are to let yourself be driven into a state of moral dumfounding and defend your position nonetheless. People differ with respect to how willing they are to justify hopeless beliefs, or with respect to how happy they are to give up their judgement under the uncoercive coercion of the better argument. What about people who cite good reasons for their otherwise emotionally driven intuitions and are thus "confabulating" in this first round, but are happy to revise their judgement in the light of undermining reasons? If they were merely confabulating, and their reasons did not play any causal role *whatsoever*, how could the fact that those allegedly confabulated reasons have been undermined cause them to change their point of view?

This suggests that there is a second way for reasons to become effective: the first one lies "upstream", and recruits our capacity to educate our intuitions. The second one lies more "downstream", and depends on how subjects' *would* react – counterfactually – to legitimate challenges to their intuitions and reasoning. On that account, whether one is confabulating is not only a question of whether one cites the reasons that played a formative role for his or her intuitions. It is also a question of whether one treats one's reasons, once they are on the table, *as if* they did play such a role, and whether one is willing to reconsider his judgement when one's reasons are debunked. If your good, but merely co-opted reasons function in such a reason-responsive way, you are not confabulating, because in that case, there is a real causal connection between your judgement and your reasoning, just an indirect and delayed one. If your co-opted reasons are not like that, then you *are* confabulating, because this reveals that there really was nothing but, say, your affection to the girl that made you change your mind, and this affection was not coupled (not even counterfactually) with any rational insight. Moral reasons are free-floating in that sense: one can grab hold of them if one has to, and even though they might have played no significant role at first, they might do so in the future, when it comes to defending one's position in a moral conversation, or to altering it in light of a new case.

5. Moral judgements as educated intuitions

5.1 *Intuitive judgement and moral education: experience and teaching*

Intuitions are, like most other kinds of judgement and behaviour, educated through experience and teaching (Hogarth 2001).

The first of the two ways – experience – is nicely reflected by Kohlberg's (1969; see also Darley and Schultz 1990) stages of moral development. These do not only correspond to internal cognitive developments on behalf of the subject, but to external changes in the social environment subjects are confronted and have to deal with. It is hardly a coincidence that the patterns of reasons subjects come to master over the course of their moral development tend to mirror the social, interactive environment subjects typically find themselves in and most depend on. Children's first interactive context is their parents. Accordingly, the reasons they put forward for their judgements refer to authority and punishment. As peers and friends become more important, their moral reasoning typically refers to the rules of a specific community. When adolescents have developed a stable identity of their own, and start to endorse values and norms independently of any particular social context, their moral reasoning refers to universal rights and abstract norms of justice and fairness. A crucial part of "education" through experience consists in joint action with other people, which is impossible without a shared background of moral norms. Accordingly, feedback from unsuccessful, interrupted collective action will have an influence on the moral norms one is equipped with in future attempts to act jointly with other people. To put it in the form of a slogan: "Moral stages are not structures of thought. They are structures of action encoded in thought" (Reed 2008, 373). Experience is the process by which structures of action become encoded in our intuitive responses.

Explicit teaching is a profoundly important means of educating children's and adult's capacity to make moral judgements. In fact, it is part of what makes children understand what makes a judgement a moral one that they be made familiar with different categories of reasons that bear on the validity of different kinds of (social) norms and on why their transgression is prohibited. In a series of studies, Smetana (1984, 1989) found that children learn the distinction between moral and conventional norms on the basis of the different kinds of reasons their parents put forward when either the former or the latter are violated. Parents tend to refer to considerations of social order and abstract rules in the case of conventional transgressions; in the moral case, parents will request children to take the perspective of others, think about other children's needs and feelings, or refer to considerations that pertain to their rights and entitlements. Children thereby learn which norms it is appropriate to associate their moral emotions with.

Like any other kind of education, the education of (moral) intuitions is a process of habitualization. The level of automaticity with which an intuitive judgemental response is triggered increases with the number of repetitions, a process that consists in a "migration" of controlled and effortful cognitive processes into an agent's effortless, perception-like intuitive system. Dreyfus and Dreyfus' (1986, 1991; see also Musschenga 2009) model of intuitive skill acquisition contains an apt description of the stages this migration typically involves. For the *novice* (stage 1) who learns how to perform a task (here: of moral judgement formation), the elements of this task need to be decomposed such that he can become familiar with them. This is, as mentioned above, what parents do when they teach their children what they ought and ought not do, and why this is the case. The following stages – from *advanced beginner* (2) to *competence* (3) – involve an increase in what might be called *normative automaticity*: the agent not only acquires higher levels of automaticity in dealing with certain tasks, but manages to perform them with greater reliability and a

more autonomous and flexible understanding of her subject. Competent subjects have acquired mastery of moral concepts and implicit knowledge of the reasons that count in the context of moral discourse. They are in a position to teach the practice of moral judgement to novices by being an example. At the level of *proficiency* (4) and *expertise* (5), the agent need not rigidly follow the rules she has been taught anymore at all, but has acquired a perception-like, intuitive ability to evaluate which response a particular situation calls for. A proficient moral judge will be more original, creative, independent and reflective in his application of moral concepts and have the ability to improve his or her web of moral beliefs from within. An expert moral judge, then, is a proficient moral judge with meta-knowledge about normative and meta-ethical theories concerning the nature of the practice of moral judgement and reasoning she engages in.

5.2 *Ex ante education*

The education of the intuitions is supposed to improve the quality of our intuitive responses to morally salient scenarios. For the most part, rationalist accounts of moral judgement have focused on how reflection and deliberation can regulate our moral emotions and intuitions after the fact. But intuitive education is possible *ex post* and *ex ante*: the latter type is antecedent-focused, the former response-focused. *Ex ante* education is concerned with the conditions under which a moral intuition is generated:

> Prior reasoning can determine the sorts of output that emerge from intuitive systems. This can happen through shifts in cognitive appraisal, as well as through conscious decisions as to what situations to expose oneself to. In both of these regards, prior controlled processes partially determine which fast, unconscious, and automatic intuitions emerge. (Pizarro and Bloom 2003, 194; see also Gross 2002, 282)

Ex post education is concerned with how an intuition, once it has been generated, can and should be dealt with, and how controlled *after-the-fact* reflection feeds back into an agent's intuitive system: a "closer examination of the interaction between automatic and controlled reflective processes in moral judgment [. . .] makes room for the view [. . .] that genuine moral judgments are those that are regulated or endorsed by reflection" (Kennett and Fine 2009, 78). This type of *post hoc* deliberation often results in an improvement of a subject's intuitions on future occasions.

Subjects can influence their moral intuitions *ex ante* by selecting the situational input they are confronted with (Pizarro and Bloom 2003). Situation selection and input control can be very general and far-ranging: many academics, I suppose, have the lingering suspicion that a successful career in business will require them to get their hands dirty, and that it might alter their character in unwanted ways. Avoiding situations one can expect to trigger unwanted intuitive responses gets to the root of problems like these, because relevant factual knowledge often does not suffice to eliminate automatic behavioural responses, whether they are character traits, stereotypes, prejudices or racial biases.

Input control does not have to be negative, and consist in the avoidance of certain situations, but can also be about selectively exposing oneself to wanted situational stimuli. People who want to become vegetarians often start reading about the horrors of factory farming, and people who register, and want to get rid of, their unwanted racist attitudes can deliberately engage with people of a different race, a strategy that is known to be very effective (Brandt 1979). Monteith and Mark (2005) have studied, for example, the various strategies people employ to monitor and influence their own racial biases and prejudices. They have shown that under suitable circumstances, people can be very good at

registering so-called "should/would discrepancies" in their automatic responses to racially significant stimuli. Subjects can establish – via retrospective and prospective reflection – automatically activated cues which serve to control and help inhibit unwanted judgemental responses, and thus bridge the gap between how people think they would and should behave.

Ex ante education of (moral) intuitions need not be focused on external conditions. There are ways to directly influence the formation of intuitive responses by altering internal processes underlying judgement formation. One striking example for this comes from research on so-called "implementation intentions" (Gollwitzer 1999, Gollwitzer et al. 2009). These *if–then* plans to respond in certain ways upon encountering anticipated stimuli can create "instant habits" (Gollwitzer 1999, 499). Gollwitzer et al. found that implementation intentions can significantly shape people's automatic emotional reactivity in desired ways. In one study, a group of participants managed to reduce automatic responses of disgust and fear towards external stimuli by forming an if–then intention to stay calm and relaxed when confronted with fear-inducing images.

Arguably the most spectacular evidence for how subjects can educate their intuitive judgements comes from research on social prejudice and stereotype activation. Rudman, Ashmore, and Gary (2001) showed that repeated exposition to suitable stimuli can dramatically influence people's implicit as well as explicit racial prejudices and stereotypes. After taking a class about social prejudice taught by an African-American professor, students were less likely to automatically associate typical "black" names with negative concepts such as *laziness* or *hostility*. And this is not a peculiarity only a bunch of scientists are interested in: a quick look at the history of the last 200 years tells us that mechanisms like these have had a profound impact on our society, and helped do away with irrational responses, ill-founded prejudices, and intolerance.

5.3 *Ex post education*

Ex post education of moral intuitions recruits the human capacity for metacognition: the ability to reflectively monitor one's cognitive operations and alter them according to standards of rationality or reliability deemed appropriate by the reflecting subject. This is not so much an empirical hypothesis – although it is also that – as a constraint on the concept of a moral judgement: genuine moral judgements are responsive to episodes of rational reflection (Fine 2006; Jones 2006; Sauer 2011a). These play a role in the education of our emotionally charged moral intuitions to the extent that they feed back into subjects' intuitions, and thereby exert a corrective influence. The idea that episodes of moral reflection feed back into our intuitive responses and help educate them undermines the incompatibility thesis from yet another angle. Often in the debate between sentimentalists and intuitionists about moral judgement on the one side, and rationalists on the other,

> there is no further discussion of the extent to which moral intuitions are amenable to modification as the result of reflection. The [. . .] clear focus on the idea that moral judgments are driven by *either* one system or the other, and the need for reflective processes to override intuitive responses, suggests that the scope of reflective modification of moral intuitions is assumed to be minimal. (Craigie 2012, 60; see also Besser-Jones 2012)

But this is not so, as the empirical evidence shows.

People use their capacity to make intuitive judgements in utterly different areas – from moral issues to assessments of probabilities, logical relations and judgements about their

own well-being. As diverse as these areas are, the mechanisms subjects recruit to make judgements are often, though not always, quite similar. The influence of *post hoc* reflection on these mechanisms is well documented, especially in cases where people correct for so-called "mental contamination" (Wilson and Brekke 1994). Schwartz and Clore (1983) have shown that people are both willing and able to take into account irrelevant circumstantial factors that might distort their judgement and to discount them if necessary. In one of their experiments, subjects were asked to evaluate their quality of life under different conditions (on a sunny or rainy day, respectively). When they were given information about how that fact might have affected their evaluation, subjects discounted it, trying to counterbalance the extraneous influence.

The habitualization of judgemental patterns in the wake of *ex post* reflection is not a pious hope, but results almost inevitably from the repeated execution of moral reflection:

> [. . .] the more frequently people perform a behavior, the more habitual and automatic it becomes, requiring little effort or conscious attention. One of the most enduring lessons of social psychology is that behavior change often precedes changes in attitudes and feelings. Changing our behavior to match our conscious conception of ourselves is thus a good way to bring about changes in the adaptive unconscious. (Wilson 2002, 212)

Ex post moral reasoning exerts a rational pressure on subjects to modify their moral intuitions in accordance with the reasons that become available them, or to give up their intuitions if there are not any. The perception of should/would discrepancies, the motivation to overcome them and the evidence that those rational "shoulds" do become effective in shaping a person's automatic responses are prime examples here. These processes cannot but have an influence on a person's moral mindset, however subtle, mediated and delayed it may be.

For those who remain unconvinced by this reassurance, here is one more line of evidence for the malleability of automatic processes of judgement formation. The phenomenon of moral dumbfounding seems to suggest that a fair amount of our moral judgements are based on irrational automatic intuitions subjects are not able to justify. But, as Levy (2007; also see Haidt et al. 1993) observes, this is not true; subjects who have a better education are able to do so, or do not attempt to if they are not:

> Haidt's work on moral dumbfounding [. . .] actually demonstrates that dumbfounding is in inverse proportion to socio-economic status (SES) of subjects [. . .]. Higher SES subjects differ from lower not in the moral theories they appeal to, but in the content of their moral responses. In particular, higher SES subjects are far less likely to find victimless transgressions – disgusting or taboo actions – morally wrong than lower. (Levy 2007, 307)

The only explanation for this seems to be that

> the differential responses of higher and lower SES subjects demonstrate that in fact moral intuitions are amenable to education. The greater the length of formal education, the less the likelihood of subjects holding that victimless transgressions are morally wrong [. . .]. (Levy 2007, 307)

The model developed here can explain that fact, the SI model cannot.

Moral judgements, I have argued, are typically made intuitively. But moral reasoning, even of the explicit kind, is typically performed habitually, too. Haidt exploits this latter fact in order to produce the phenomenon of moral dumbfounding: if subjects were not used to putting forward certain reasons in support of their verdicts in ordinary situations, they

would not attempt to do so – misguidedly – in the extraordinary cases they are given in his experiment. This shows that the education of moral intuitions has two different objects. It is not only particular intuitions with a particular morally salient content – wrongdoing ought to be punished, promises must not be broken, honesty is a virtue – that are being acquired over the course of a subject's moral education, but always also a particular set of reasons that bear on those intuitions. In Haidt's incest case, subjects have a particular moral intuition: *incest is wrong*. But they are also equipped with a set of reasons they have learned to cite as considerations relevant for their judgement: the harmfulness of inbreeding, the value of family relations and so forth. Moral education is about both: an improvement of one's intuitions as well as the reasons one has for them.

The educational processes just described, especially those of *ex post* reflection, can be performed monologically or dialogically, that is, either by interacting with and talking to fellow moral reasoners, or by interacting with and talking to oneself. The System I model misrepresents the nature of intersubjective moral reasoning when it describes it as a process of two or more parties reciprocally persuading and being persuaded by one another. Quite the contrary: Lapsley and Narvaez (2004) put great emphasis on how episodes of intersubjective reasoning shape moral intuitions, which they call "chronically accessible cognitive-affective moral schemas" (24). They argue that as children interact and communicate with the moral "experts" around them – usually their parents – they develop ever more sophisticated abilities of intuitive moral perception. These episodes of dialogical moral reasoning then become integrated into children's identity, and "enable children to organize events into personally relevant autobiographical memories, which provides, in the process, as part of the self-narrative, action-guiding scripts [. . .] that become overlearned, frequently practiced, routine, habitual, and automatic" (26). Such dialogical moral reasoning is not only a genuine source of improvement of one's moral intuitions. It is also a demand of morality itself to maintain the conditions under which it is possible.

6. Conclusion: reason and its limits

Let me sum up. Our moral intuitions are educated by different *means*: experience and teaching. There are different *types* of it: education operates on our intuitions *ex ante* and *ex post*. The *objects* of education are the content of particular moral judgements as well as the reasons that bear on them, and this whole process can be *monological* as well as *dialogical* in form.

None of the things just said are meant to deny the shortcomings or detrimental effects explicit reasoning can have. The phenomenon of "choice blindness" (Johansson et al. 2005) illustrates just how poorly people sometimes reason about their decisions. Wilson et al. (1989) found that the inability to accurately introspect the factors that influence our choices can lead to cases of severe attitude/behaviour dissonances: the reasons people can articulate to themselves can influence their *assessment* of states of affairs, but leave their *behaviour* almost entirely unaffected. It gets even worse when conscious reasoning not only does not adequately represent what influences subjects' beliefs and decisions, but starts to actually reduce their quality.[6] Wilson and Schooler (1991) have found, for example, that people's assessments of the quality of different brands of jams became worse, compared to experts' opinions, when they were asked to analyse why they felt the way they did. Conscious deliberation can sometimes be disruptive. These examples, however, are exceptions rather than the rule.

As far as their *post hoc* nature is concerned, moral reasoning and ordinary reasoning are "companions in guilt" (Lillehammer 2007), and the SI model uses a double standard for the

two. The suggestion made by the model is that moral reasoning is prone to confabulatory after the fact rationalizations, because our moral intuitions are rooted in intense emotions, and we are strongly motivated to hedge them as good as we can. But this cannot be the correct explanation. Many psychologists nowadays (Wegner 2002; Wilson 2002) claim that all reasoning, not just the moral kind, is a *post hoc* enterprise, even though those types of reflection typically do not serve to protect dearly held ethical convictions. Reasoning comes after the fact due to simple restrictions of information processing capacities. But humans have to struggle with these restrictions in moral and non-moral cases alike. Unless one wants to defend the idea that there is no such thing as genuine reasoning at all, whether it is moral or non-moral reasoning, the insight that all reasoning is "epiphenomenal" in this sense should not be particularly troubling. In fact, it is hard to see how it could be otherwise.

For the same reason, the primary function of explicit moral reflection is not to directly precede and thereby cause people's moral judgements, but to feed back into people's intuitive responses and to improve, shape and inform them. It is an ongoing process that creates a chain of feedback loops, with each one influencing the following one. The reason why the SI model mistakes this process as confabulation is that if one looks at only one of those loops, it is indeed the case that the underlying intuitive process is prior to subjects' conscious reasoning: for each loop at a time, the automatic intuition comes first. But if one steps back and takes a look at the whole chain of feedback loops, what used to look like idle confabulation suddenly starts to look like an extremely efficient way of managing one's intuitions. This, I have argued, is what the education of our intuitions is all about.

Acknowledgements

I wish to thank Tom Bates, Jeanette Kennett, Pauline Kleingeld, Jesse Prinz, Markus Schlosser, and Maureen Sie for many helpful suggestions and valuable feedback on this paper. Research on this paper was funded by The Netherlands Organisation for Scientific Research (NWO).

Notes

1. Obviously, not all emotions are automatic, and not all automatic mental processes are emotional in nature. For the purposes of my argument, however, this fact can be set aside, because the challenge to rationalist models of moral judgement I discuss relies on evidence that moral judgements are based on emotion *or* intuition *or* both.
2. Throughout this paper, I do not always properly distinguish between *emotion* and *intuition*. There are two reasons for this seeming conceptual sloppiness: first, in the research I discuss, the concepts "emotion" and "intuition" are often, though not always, used interchangeably. "Emotion" highlights the dimension of feeling and "hotness" of a given process of automatic judgement formation, whereas "intuition" (which is indeed often taken to be emotionally charged) highlights its speed and conscious inaccessibility. Second, the problem I am interested in in this paper is whether we are entitled to say that moral judgement does not have a rational foundation on the basis of the evidence that most of our moral judgements are the upshot of automatic processes. Emotion and intuition are both types of automatic processes. In that respect, most of my argument can remain insensitive to the distinction, and remains unaffected by the obvious differences between emotion and intuition.
3. My argument will not in any way rely on the claim that philosophers of action and empirical moral psychologists *explicitly endorse* the incompatibility thesis (although some might). Rather, it seems to be the case that many researchers are often implicitly, and sometimes maybe even inadvertently, *committed* to the thesis, because most available theories of rational action and moral judgement are, as a matter of fact, unable to account for the possibility of automatic yet rational thought and behaviour.

4. It is of course questionable whether there are any rationalists out there who would recognize their position in Haidt's description of rationalism. I do not think, however, that Haidt is arguing against a strawman. There is at least some plausibility in the idea that we think before we judge morally, and there is an undeniable tendency among people to overestimate the extent to which that happens.
5. In his more recent writings, Velleman (2010) seems to be aware of this.
6. An excellent philosophical discussion of cases where emotions rather than explicitly articulated moral principles do a better job at latching onto the features of the world that constitute (moral) reasons for acting can be found in Arpaly (2003).

Notes on contributor

Hanno Sauer is a PhD student at the University of Leiden/University of Groningen, the Netherlands. His research is located at the intersection of moral psychology, metaethics, practical reason and normative ethics.

References

Arpaly, N. 2003. *Unprincipled virtue: An inquiry into moral agency*. New York: Oxford University Press.

Bargh, J.A. 1994. The four horsemen of automaticity. In *Handbook of social cognition*, ed. R.S. Wyer and T.K. Srull, 1–40. Hillsdale, NJ: Erlbaum.

Bargh, J.A., and T.L. Chartrand. 1999. The unbearable automaticity of being. *American Psychologist* 54, no. 7: 462–79.

Besser-Jones, L. 2012. The role of practical reason in empirically informed moral theory. *Ethical Theory and Moral Practice* 15, no. 2: 203–20.

Blair, J., D. Mitchell, and K. Blair. 2005. *The psychopath: Emotion and the brain*. Oxford: Blackwell.

Blair, R.J.R. 1995. A cognitive developmental approach to morality: Investigating the psychopath. *Cognition* 57: 1–29.

Brandt, R.B. 1979. *A theory of the good and the right*. Oxford: Oxford University Press.

Clark, A. 2010. Memento's revenge: The extended mind, extended. In *The extended mind*, ed. R. Menary, 43–67. Cambridge, MA: MIT Press.

Clarke, S. 2008. SIM and the city: Rationalism in psychology and philosophy and Haidt's account of moral judgment. *Philosophical Psychology* 21, no. 6: 799–820.

Craigie, J. 2011. Thinking and feeling: Moral deliberation in a dual-process framework. *Philosophical Psychology* 24, no. 1: 53–71.

Damasio, A. 1994. *Descartes' error: Emotion, reason, and the human brain*. London: Penguin Books.

Dancy, J. 2000. *Practical reality*. Oxford: Oxford University Press.

Darley, J.M., and T.R. Schultz. 1990. Moral rules: Their content and acquisition. *Annual Review of Psychology* 41: 525–56.

Dijksterhuis, A. 2004. Think different: The merits of unconscious thought in preference development and decision making. *Journal of Personality and Social Psychology* 87, no. 5: 586–98.

Dijksterhuis, A. 2006. On the benefits of thinking unconsciously: Unconscious thought can increase post-choice satisfaction. *Journal of Experimental Social Psychology* 42: 627–31.

Dreyfus, H.L., and S.E. Dreyfus. 1986. *Mind over machine: The power of human intuitive expertise in the era of the computer*. New York: Free Press.

Dreyfus, H.L., and S.E. Dreyfus. 1991. Towards a phenomenology of ethical expertise. *Human Studies* 14: 229–50.

Evans, J.S.B.T. 2003. In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Science* 7, no. 10: 454–59.

Evans, J.S.B.T. 2008. Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology* 59: 255–78.

Fine, C. 2006. Is the emotional dog wagging its rational tail, or chasing it? Reason in moral judgment. *Philosophical Explorations* 9, no. 1: 83–98.

Gehlen, A. 1940. *Der Mensch. Seine Natur und seine Stellung in der Welt*. Berlin: Junker und Dünnhaupt.

Gert, J. 2003. Brute rationality. *Nous* 37, no. 3: 417–46.
Giddens, A. 1984. *The constitution of society: Outline of the theory of structuration*. Cambridge: Polity Press.
Gollwitzer, P. 1999. Implementation intentions: Strong effects of simple plans. *American Psychologist* 54, no. 7: 493–503.
Gollwitzer, P., I. Schweiger Gallo, A. Keil, K.C. McCulloch, and B. Rockstroh. 2009. Strategic automation of emotion regulation. *Journal of Personality and Social Psychology* 96, no. 1: 11–31.
Greene, J.D., L.E. Nystrom, A.D. Angell, H.M. Darley, and J.D. Cohen. 2004. The neural bases of cognitive conflict and control in moral judgment. *Neuron* 44, no. 2: 389–400.
Greene, J.D., R.B. Sommerville, L.E. Nystrom, H.M. Darley, and J.D. Cohen. 2001. An fMRI investigation of emotional engagement in moral judgment. *Science* 293: 2105–8.
Gross, J.J. 2002. Emotion regulation: Affective, cognitive and social consequences. *Psychophysiology* 39: 281–91.
Haidt, J. 2001. The emotional dog and its rational tail. *Psychological Review* 108, no. 4: 814–34.
Haidt, J. 2007. The new synthesis in moral psychology. *Science* 316: 998–1001.
Haidt, J., and S. Kesebir. 2010. Morality. In *Handbook of social psychology*, ed. S. Fiske, D. Gilbert and G. Lindzey, 797–832. Hoboken, NJ: Wiley.
Haidt, J., S.H Koller, and M.G. Dias. 1993. Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology* 65, no. 4: 613–28.
Hegel, G.W.F. 1986. *Phänomenologie des Geistes*. Frankfurt am Main: Suhrkamp.
Hegel, G.W.F. 1902. *Philosophy of History.* Trans. J. Sibree. New York: P.F. Collier & Son.
Hirstein, W. 2005. *Brain fiction: Self-deception and the riddle of confabulation*. Cambridge, MA: MIT Press.
Hogarth, R.M. 2001. *Educating intuition*. Chicago: Chicago University Press.
Horgan, T., and M. Timmons. 2007. Morphological rationalism and the psychology of moral judgment. *Ethical Theory and Moral Practice* 10: 279–95.
Johansson, P., L. Hall, S. Sikström, and A. Olsson. 2005. Failure to detect mismatches between intention and outcome in a simple decision task. *Science* 310: 116–9.
Jones, K. 2006. Metaethics and emotions research: A response to prinz. *Philosophical Explorations* 9, no. 1: 45–53.
Kahneman, D. 2003. A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist* 58, no. 9: 697–720.
Kahneman, D., and S. Fredrick. 2002. Representativeness revisited: Attribute substitution in intuitive judgment. In *Heuristics and biases*, ed. T. Gilovich, D. Griffin and D. Kahneman, 49–81. New York: Cambridge University Press.
Kennett, J., and C. Fine. 2009. Will the real moral judgment please stand up? *Ethical Theory and Moral Practice* 12: 77–96.
Koenigs, M., L. Young, R. Adolphs, D. Tranel, F. Cushman, M.D. Hauser, and A. Damasio. 2007. Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature* 446, no. 7138: 908–11.
Kohlberg, L. 1969. Stage and sequence: The cognitive-developmental approach to socialization. In *Handbook of socialization theory and research*. ed. D.A. Goslin, 347–480. Chicago: Rand McNally.
Lally, P., C.H.M. van Jaarsveld, H.W.W. Potts, and J. Wardle. 2010. How are habits formed: Modelling habit formation in the real world. *European Journal of Social Psychology* 40: 998–1009.
Langer, E., A. Blank, and B. Chanowitz. 1978. The mindlessness of ostensibly thoughtful action: The role of 'placebic' information in interpersonal interaction. *Journal of Personality and Social Psychology* 36, no. 6: 635–42.
Lapsley, D.K., and D. Narvaez. 2004. A social-cognitive approach to the moral personality. In *Moral development, self and identity*, ed. D.K. Lapsley and D. Narvaez, 189–212. Mahwah, NJ: Erlbaum.
Levy, N. 2007. *Neuroethics: Challenges for the 21st century*. Cambridge: Cambridge University Press.
Lillehammer, H. 2007. *Companions in guilt: Arguments for ethical objectivity*. Basingstoke: Palgrave Macmillan.
Lovibond, S. 2002. *Ethical formation*. Cambridge, MA: Harvard University Press.
McDowell, J. 1978/1998. Are moral requirements hypothetical imperatives? In *Mind, value, and reality*, 77–94. Cambridge, MA: Harvard University Press.

McDowell, J. 1994. *Mind and world*. Cambridge, MA: Harvard University Press.
Monteith, M., and A. Mark. 2005. Changing one's prejudiced ways: Awareness, affect, and self-regulation. *European Review of Social Psychology* 16, no. 1: 113–54.
Musschenga, B. 2008. Moral judgement and moral reasoning. In *The contingent nature of human life: Bioethics and the limits of human existence*, ed. M. Düwell, C. Rehmann-Sutter and D. Mieth, 131–41. Dordrecht: Springer.
Musschenga, B. 2009. Moral intuitions, moral expertise and moral reasoning. *Journal of Philosophy of Education* 43, no. 4: 597–613.
Nichols, S. 2004. *Sentimental rules: On the natural foundations of moral judgment*. New York: Oxford University Press.
Nisbett, R.E., and T.D. Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84, no. 3: 231–59.
Nisbett, R.E., and T.D. Wilson. 1978. The accuracy of verbal reports about the effects of stimuli and behavior. *Social Psychology* 41, no. 2: 118–31.
Pettit, P. 2001. *A theory of freedom: From the psychology to the politics of agency*. Oxford: Oxford University Press.
Pizarro, D.A., and P. Bloom. 2003. The intelligence of the moral intuitions: Comment on Haidt (2001). *Psychological Review* 110, no. 1: 193–6.
Pollard, B. 2003. Can virtuous actions be both habitual and rational? *Ethical Theory and Moral Practice* 6, no. 4: 411–25.
Pollard, B. 2005a. Naturalizing the space of reasons. *International Journal of Philosophical Studies* 13, no. 1: 69–82.
Pollard, B. 2005b. The rationality of habitual actions. In *Proceedings of the Durham-Bergen Philosophy Conference*, 1, 39–50.
Prinz, J. 2007. *The emotional construction of morals*. New York: Oxford University Press.
Prinz, J., and S. Nichols. 2010. Moral emotions. In *The moral psychology handbook*, ed. J. Doris and The Moral Psychology Research Group, 111–47. New York: Oxford University Press.
Reed, D.C. 2008. A model of moral stages. *Journal of Moral Education* 37, no. 3: 357–76.
Rudman, L.A., R.D. Ashmore, and M.L. Gary. 2001. 'Unlearning' automatic biases: The malleability of implicit prejudice and stereotypes. *Journal of Personality and Social Psychology* 81, no. 5: 856–68.
Ryle, G. 1949. *The concept of mind*. London: Penguin Books.
Saltzstein, H.D., and T. Kasachkoff. 2004. Haidt's moral intuitionist theory: A psychological and philosophical critique. *Review of General Psychology* 8, no. 4: 273–82.
Sauer, H. 2011a. Psychopaths and filthy desks: Are emotions necessary and sufficient for moral judgment? *Ethical Theory and Moral Practice* 15, no. 1: 95–115.
Sauer, H. 2011b. Social intuitionism and the psychology of moral reasoning. *Philosophy Compass* 6, no. 10: 708–21.
Sauer, H. Forthcoming. Morally irrelevant factors: What's left of the dual-process model of moral cognition? *Philosophical Psychology*, 1–29.
Scanlon, T.M. 1998. *What we owe to each other*. Cambridge, MA: Harvard University Press.
Scheler, M. 2007. *Die Stellung des Menschen im Kosmos*. Bonn: Bouvier.
Schiller, F. 2004. Über Anmut und Würde. In *Sämtliche Werke*: *Erzählungen/Theoretische Schriften*, Vol. 5, ed. P.A. Alt, A. Meier, and W. Riedel, 433–89. München/Wien: Carl Hanser Verlag.
Schnall, S., J. Haidt, G.L. Clore, and A.H. Jordan. 2008. Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin* 34, no. 8: 1096–109.
Schwartz, N., and G.L. Clore. 1983. Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology* 45, no. 3: 513–23.
Sie, M. 2009. Moral agency, conscious control, and deliberative awareness. *Inquiry* 52, no. 5: 516–31.
Singer, P. 2005. Ethics and intuitions. *The Journal of Ethics* 9 nos. 3–4: 331–52.
Smetana, J.G. 1984. Toddlers' social interactions regarding moral and conventional transgressions. *Child Development* 55, no. 5: 1767–76.
Smetana, J.G. 1989. Toddlers' social interactions in the context of moral and conventional transgressions in the home. *Developmental Psychology* 25, no. 4: 499–508.
Sneddon, A. 2009. A social model of moral dumbfounding: Implications for studying moral reasoning and moral judgment. *Philosophical Psychology* 20, no. 6: 731–48.

Snow, N.E. 2006. Habitual virtuous actions and automaticity. *Ethical Theory and Moral Practice* 9, no. 5: 545–61.
Snow, N.E. 2010. *Virtue as social intelligence: An empirically grounded theory.* New York: Routledge.
Uhlmann, E.L., D.A. Pizarro, D. Tannenbaum, and P.H. Ditto. 2009. The motivated use of moral principles. *Judgment and Decision Making* 4, no. 6: 476–91.
Valdesolo, P., and D. DeSteno. 2006. Manipulations of emotional context shape moral judgment. *Psychological Science* 17, no. 6: 476–77.
Velleman, D. 2000. *The possibility of practical reason*. New York: Oxford University Press.
Velleman, D. 2010. There are no 'reasons for acting'. Unpublished manuscript.
Wegner, D. 2002. *The illusion of conscious will*. Cambridge, MA: MIT Press.
Wheatley, T., and J. Haidt. 2005. Hypnotic disgust makes moral judgment more severe. *Psychological Science* 16, no. 10: 780–84.
Wilson, T.D. 2002. *Strangers to ourselves: Discovering the adaptive unconscious*. Cambridge, MA: Belknap Press of Harvard University Press.
Wilson, T.D., and N. Brekke. 1994. Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin* 116, no. 1: 117–42.
Wilson, T.D., D.S. Dunn, D. Kraft, and D.J. Lisle. 1989. Introspection, attitude change, and attitude-behavior consistency: The disruptive effects of explaining why we feel the way we do. In *Advances in experimental social psychology.* ed. L. Berkowitz, 287–343. San Diego, CA: Academic Press.
Wilson, T.D., and J.W. Schooler. 1991. Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology* 60, no. 2: 181–92.
Winch, P. 1958. *The idea of a social science and its relation to philosophy.* London: Routledge & Kegan Paul.
Wood, W., and D.T. Neal. 2007. A new look at habits and the habit-goal interface. *Psychological Review* 114, no. 4: 843–63.
Wood, W., J.M. Quinn, and D.A. Kashy. 2002. Habits in everyday life: Thought, emotion, and action. *Journal of Personality and Social Psychology* 83, no. 6: 1281–97.