Is Deontology a Moral Confabulation?

Forthcoming in *Neuroethics*

Emilian Mihailov
Romanian Academy, Iasi Branch
Research Center in Applied Ethics, Faculty of Philosophy,
University of Bucharest
emilian.mihailov@gmail.com

Abstract: Joshua Greene has put forward the bold empirical hypothesis that deontology is a confabulation of moral emotions. Deontological philosophy does not steam from "true" moral reasoning, but from emotional reactions, backed up by post hoc rationalizations which play no role in generating the initial moral beliefs. In this paper, I will argue against the confabulation hypothesis. First, I will highlight several points in Greene's discussion of confabulation, and identify two possible models. Then, I will argue that the evidence does not illustrate the relevant model of deontological confabulation. In fact, I will make the case that deontology is unlikely to be a confabulation because alarm-like emotions, which allegedly drive deontological theorizing, are resistant to be subject to confabulation. I will end by clarifying what kind of claims can the confabulation data support. The upshot of the final section is that confabulation data cannot be used to undermine deontological theory in itself, and ironically, if one commits to the claim that a deontological justification is a confabulation in a particular case, then the data suggests that in general deontology has a prima facie validity.

Keywords: confabulation, deontology, consequentialism, Greene, moral intuition

Introduction

Drawing upon his dual-process theory of moral judgement, Greene argues that "characteristically deontological judgments are preferentially supported by automatic emotional responses, while characteristically consequentialist judgments are preferentially supported by conscious reasoning and allied processes of cognitive control." [1: 699, 2, 3, 4] [1]

[1] For critical analyses of the claim that deontological judgments are predicted by emotional factors see Kahane and Shackel [5], Mihailov [6], Kahane [7].

He further develops the dual-process account to advance what has become known as the argument from irrelevant features, which aims to show that emotionally-driven deontological judgements respond to morally irrelevant features, such as spatial distance or personal force.[2] If he is right about this, then the reliability of deontological intuitions in making moral judgements is seriously called into question. His main contention is that we ought to distrust the automatic process of moral judgement, and favour consequentialism because it is preferentially supported by processes of cognitive control.

Yet, it seems clear that deontologists do engage in conscious moral reasoning. Even a cursory analysis of deontological arguments will show that their reflections do not seem to just be a heap of emotional exhortations. Greene's bold hypothesis, however, is that "deontological philosophy, rather than being grounded in moral *reasoning*, is to a large extent an exercise in moral *rationalization*." [2: 36] The upshot of this is that deontology stems from automatic emotional reactions, which are then justified by a post hoc rationalization. According to Greene, this explains why there appears to be a correspondence between what deontologists tell us to do, and what our emotions tell us to do: "What should we expect from creatures who exhibit social and moral behaviour that is driven largely by intuitive emotional responses and who are prone to rationalization of their behaviours? The answer, I believe, is deontological moral philosophy." [2: 62-63]

[2] Greene has changed his concept of personalness from being up close to the action to exerting personal force [8]. More recently he admitted that deontological judgements are sensitive to the distinction between intended and foreseeable consequences [9]. See Berker [10] for a critique of Greene's argument from irrelevant features.

It is important to distinguish between two senses of rationalization in order to carefully assess the claim that deontology is not based on genuine reasoning. On the one hand, Greene argues that "Kantian style of rationalizing" is intuition chasing [1]. The reasoning part consists in summarizing and organizing intuitive responses. While this "exercise in moral rationalization" has, arguably, no revisionary control over intuitions, it does not follow that it lacks objective merit in itself. Only when we presuppose the unreliability of intuitions i.e. that they track morally irrelevant features, does rationalization go astray. Suppose that intuitions were reliable, as some believe [11]. In this case, rationalization would not lack objective merit. Indeed, it would actually serve to organise a pattern of morally relevant features, as tracked by our intuitions. Rationalization understood as a process of finding intuitive patterns can, nevertheless, be poorly done, but it does not follow that the process *per se* is unreliable. Thus, calling deontology a rationalization in this sense is not a debunking characterisation unless one also proves the unreliability of intuitions. On another level, Greene describes the rationalization involved in deontological philosophy as "a kind of moral confabulation" [2: 63]. When people do not have introspective access to what triggers their responses, it is welldocumented that they tend to confabulate a plausible story as to why they reacted the way they did [12]. In the same manner, strong feelings issue prohibitions and it is not clear how to make sense of them. Greene says that "with the help of some especially creative philosophers" we make up a rationally appealing story to support our emotionally-driven commands [2: 63]. The argument from confabulation data is not supposed to undermine deontology on the basis that it is a rationalization only of moral emotions or that current rationalizations are not reliable, but rather, that it is more committed than consequentialism to a *type* of process that is not epistemically credible in itself.[3]

In this paper I will deal only with the charge that deontology is a confabulation type of rationalization and argue that the empirical evidence Greene offers does not support this hypothesis. First, I will highlight several points in Greene's discussion of confabulation, and identify two possible models. Then, I will argue that the evidence does not support the relevant model of deontological confabulation. In fact, I will show that deontology is unlikely to be a confabulation because alarm-like emotions, which allegedly drive deontological theorizing, are resistant to be subject to confabulation. I will end by drawing attention to some implications of the data, which indicate that they cannot be used to cast doubt "on deontology as a school of normative moral thought" [2: 36].

Clarifying Greene's model of deontological confabulation

In many places, Greene uses rationalization and confabulation interchangeably, an indication that he also focuses on the unconscious tendency to invent reasons for belief or action. In his recent book, *Moral Tribes*, he states that "the moral equivalent of confabulation is

[3] When Richard Dean argues that deontological theories are not just "rationalizations" of emotional reactions, he assumes that emotional reactions are troubling, not the rationalization itself [13: 49]. Berker also takes Greene to be objecting to deontology because of its emotional basis, but he is reluctant to attribute the argument on a more charitable interpretation [10: 315]. Moreover, Wielenberg suggests that Greene's argument implies the need for better deontological rationalization [14: 126]. They have in mind the sense of rationalization as summarizing intuitive responses, whereas Francis Kamm also points out Greene's usage of rationalization as confabulation, in the debunking sense of lacking objective merit in itself [15].

rationalization. The confabulator perceives himself doing something and makes up a rational sounding story about what he's doing and why. The moral rationalizer feels a certain way about a moral issue and then makes up a rational-sounding justification for that feeling." [16: 300] The striking feature of confabulators is that they "seem to believe their claims, and the consensus among those who study them is that they are not deliberately lying". [17: 1] People's behavior is often determined by unconscious influences which are hard to spot through introspection. As we do not have conscious access to the causes of our responses, there is a tendency to confabulate reasons either for an internal need of personal coherence, or to meet social expectations [12]. Just as the confabulator creates an ill-grounded explanation for his behaviour because is not aware of the unconscious influences, the moral rationalizer creates an ill-grounded post hoc justification for a "feeling" of wrongness, being unaware of what actually activated it.

According to Greene, two conditions are needed to spot a rationalizer: "First, you have to find a factor that predicts the rationalizer's judgments. Second, you have to show that the factor that predicts the rationalizer's judgments is not plausibly related to the factors that according to the rationalizer are the bases for his or her judgments." [2: 67-68] To illustrate this model, Greene hypothesises a case relating to romantic choices. Alice goes on many dates and evaluates the people she likes as brilliant, kind, charming, and the ones she does not like as self-absorbed. At the same time, those who were evaluated positively are exceptionally tall, while the ones rejected are less than six-foot-four. A statistical analysis reveals that height is a near-perfect predictor of Alice's preferences. However, she believes that her romantic choices are based on personality traits, ruling out the idea that height mattered. Thus, Alice's talk about personality traits "is mere rationalization." [2: 67]

The first condition Greene posits is straightforward, as one need to investigate robust patterns of behaviour by statistical analysis, or identify the influencing factors in controlled experiments. The second condition is not straightforward. How we should understand the idea that the factors are not plausibly related? Why should this be the case? It is clear that height is not related to personality traits. But it does not follow from this alone that the factors that Alice believes are the bases of her judgments are actually a mere rationalization. What does follow from this is that two kinds of factors are the bases of her judgements, and that she invokes only one kind, being unaware of additional influencing factors. This is not quite a mere rationalization. The second condition needs to prove that the rationalizer's criteria do not in fact have a significant enough causal contribution to be a predictive factor, contrary to what the rationalizer believes. Therefore, 'plausibly unrelated' must refer to a relation of causal influence on belief or action. Understanding it this way makes more sense because it is line with the first condition which requires identifying a factor that is the main causal influence and with the confabulation characteristics used by Greene that the rationalizer's factors do not drive *ex ante* his judgements. Greene is committed here to a minimal causal premise because he does not want just to point out a mere correlation or association. For example, when he claims that utilitarian judgments are driven by cognitive areas of the brain and deontological judgments by emotional areas of the brain, his aim is to find direct evidence that cognitive and emotional processes do play a causal role in *ex ante* moral judgment [18]. Explaining one's choices by highlighting features which had no significant influence on the actual choice, while not being unaware of the actual causal factors, is what a confabulation is. Greene's model of confabulation can be stated as:

(1) Alice believes that plausible factors X, Y, Z preferentially support her romantic choices.

(2) Alice's romantic choices are preferentially supported by W.

(3) Alice is unaware and does not believe that W drives her choices.

(4) X, Y, Z do not preferentially support Alice's romantic choice.

(5) Therefore, Alice's citation of factors X, Y, Z is a confabulation.[4]

To answer the question of how deontological philosophy is done, Greene claims that "we must appeal to the well-documented fact that humans are, in general, irrepressible explainers and justifiers of their own behaviour. Psychologists have repeatedly found that when people don't know why they're doing what they're doing, they just make up a plausible-sounding story" [2: 61]. It seems that for Greene the cognitive processes involved in confabulation are mechanisms that work in general for both moral and non-moral judgement. In support for this claim he invokes the dual-process theory about our brain works: "Our automatic settings gives us emotionally compelling moral answers, and then our manual modes go to work generating plausible justifications for those answers, just like the manual modes of amnesiac patients trying to explain what they're up to." [16: 300][5] Greene, thus, suggests that

[4] My reconstruction uses some features from Hirstein's model of confabulation [17] and departs from how Wielenberg [14] reconstructs Greene's argument.

[5] Greene emphasizes this when he says that the lesson to be drawn from cognitive neuroscience is that „we're *all* confabulators, and those of us with healthy brains are just better at it." [16: 300]

something similar to Alice’s thought process reflects the way deontological philosophy is done. Deontologists seem to be confident that certain factors, such as thoughts of duties and rights, generate deontological judgments, but in fact something else is decisive to issuing these judgments. Similarly, the argument against deontology has the following form:

(1) Deontologists believe that plausible factors X, Y, Z preferentially support characteristic deontological (CD) judgments.

(2) Characteristic deontological judgments are preferentially supported by W.

(3) Deontologists are unaware and do not believe that W drives CD judgments.

(4) X, Y, Z do not preferentially support CD judgements.

(5) Therefore, deontologists’ citation of X, Y, Z is a confabulation.

Premise (1) expresses the deontologist’s belief that the rational requirements of duties and rights generate CD judgments. For example, the doctrine of double effect is supposed to explain why people decide to sacrifice one person to save five in the switch dilemma, but not in the bridge dilemma. Premise (2) is Greene’s first condition to identify a factor that preferentially generates CD judgements. The evidence for (2) is supposed to be taken directly from confabulation data and indirectly from Greene’s dual process theory that CD judgements are preferentially supported by automatic emotional processes. Premise (3) expresses Greene’s usage of confabulatory characteristics when he says that “when we don’t know why we feel as we do, we make up a plausible-sounding story and go with it.” [16: 298] Thus, deontologists may be unaware of the influential factors which generate CD judgements. Moreover, confabulators tend to deny the unconscious influence when it is explicitly suggested [12]. In some places, Greene puts this point in stronger terms: “By definition, a rationalist cannot say that that some action is right or wrong because of the emotions we feel in response to it." [2: 68] This implies that deontologists deny the emotional engagement influence not only as a psychological feature manifested in confabulatory behaviour, but also as a matter of conceptual constraint. Here one may choose between the weaker empirical claim and the stronger conceptual claim. Premise (4) represents Greene's second condition and it is crucial for the argument because the reasons invoked by confabulators do not preferentially drive their judgements, neither are they derived in any way from the predictive factor. A deontologist, as Greene puts it, "makes up a rational-sounding story about what he's doing and why" [16: 300] Therefore, it is important for the model that the reasons explicitly invoked by a rationalizer to be a post hoc story, not related to the main *ex ante* causal factors. Here one may also choose between different versions of the premise. In one version, the rationalizer's reasons are not the main factors which influence his/her judgements, but they do have a meaningful contribution. In another version, the rationalizer's reasons are not the main factors which influence his/her judgements and they also have no significant contribution. Depending on which version of premise (4) one endorses, the argument will entail either that deontology is based on both genuine reasoning and rationalization, or only on rationalization. I will call this the neutral confabulation model (NC) because Greene's two conditions of rationalization do not stipulate the nature of the predictive factor W.

Greene insists that the existence of moral emotions in conjunction with confabulatory tendencies gives rise to deontological philosophy [2: 60]. However, this can also be applied to consequentialism. Why should all moral philosophy not be a confabulation of moral emotions? As Hume argued [19], all moral judgements are dependent on emotional input. To avoid the objection, Greene draws the distinction between currency emotions, specific to consequentialism, and alarm-like emotions, specific to deontology. Currency emotions announce only what is relevant in a weighing process, whereas alarm like emotions are highly salient, blunt, simple and almost forces one to issue strong commands such as “Don’t do it!” or “Must do it!” With this distinction at hand, Greene claims that consequentialist moral judgment is more cognitive and independent from emotional constraints, in contrast with deontological moral judgment which is driven by alarm-like emotions [2: 64-65]. Even though consequentialism is not emotionless, the emotions involved in consequentialist reasoning are less salient and more subtle, merely signaling what to factor in. We now get what I will call the alarm-like emotion based confabulation model (AEC), in which W is qualified as an alarm-like affective factor in premise (2).

The confabulation data

Due to Greene’s dialectics about the psychological profile of deontology and the fact that he does not stipulate the nature of the influencing factor in his two conditions for rationalization, I have identified two possible models of the argument from confabulation data: neutral confabulation (NC) and alarm-like emotion based confabulation (AEC). Next I will present Greene’s evidence and then assess which model is best suited to accommodate the data. Note that the data has to exhibit the alarm-like model of confabulation because this is characteristic of the psychological profile of deontology. If NC explains most cases then it seems that deontology is actually driven by a phenomenon which Greene hasn’t correctly identified.

Haidt and colleagues [20] devised scenarios (the incest story and the cannibalism story) to elicit responses that involve two separate processes: an aversive emotional reaction followed by a post hoc justification. Subjects were presented with stories that elicit strong intuitive responses and that are hard to reason about. In the incest story, Julie and Mark, who are brother and sister, decide to make love. Julie was already taking birth control pills and Mark used a condom. After hearing the story most subjects immediately judged the action to be wrong, and began to justify their response by pointing out psychological and inbreeding risks.

Nisbett and Wilson [21] carried out several experiments to see if people are able to report accurately their cognitive processes. In one experiment subjects had to choose one of several pairs of pantyhose displayed in a row. Explaining their preference, most people invoked quality features such as superior knit or elasticity, despite the fact they all were actually identical. In fact, what really caused the choice was an unconscious preference for items on the right side of the display. When asked whether their choices have been influenced by a left to right position effect, almost all participants denied it.

Other experiments, which aimed to challenge the trial and error or association by similarity view of problem solving, found out that people are unconsciously influenced by subtle cues [22]. Subjects were given the task to tie the ends of two strings together, hanging from a ceiling and reaching the floor in length. One hung near a wall, the other in the middle of the room. Since the cords were too far apart from each other, the solution was to swing one cord like a pendulum. Participants received a subtle cue from the experimenter, which involved casually moving one of the cords like a pendulum. The subjects were unaware of the influence and attributed their solution to a different cue.

Dutton and Aron [23] tested the hypothesis as to whether an attractive female is seen as more attractive by males who encounter her while experiencing a strong emotion (fear), than by males who did not have this strong emotion. Male subjects had to cross a scary bridge, after which they met the attractive female experimenter. In the arousal-inducing condition, participants were more than twice as likely as the control subjects to call the experimenter later and ask her out. Many subjects believed that their increased arousal was due to an increased sexual attraction to the female experimenter.

Confabulation has also been observed in patients with mental conditions such as Korsakoff's amnesia and related memory disorders. Stuss et al. [24] describe a patient who reported that he was "in an air-conditioning plant", when, in fact, he was standing near an air conditioner. The patient created a fabricated story which contained elements from the eliciting cue, without any awareness of its presence.

People acting on posthypnotic suggestions develop fascinating confabulations. Estabrooks reports the example of a hypnotized subject, who puts a lampshade on his head, kneels on the floor and calls "cuckoo" three times [25]. When he was asked to explain his behaviour, he answered: "Well, I'll tell you. It sounds queer but it's just a little experiment in psychology. I've been reading on the psychology of humour and I thought I'd see how you folks reacted to a joke that was in very bad taste."

According to Greene, one of the most striking examples of post hoc rationalization comes from studies of split-brain patients. In one study, researchers flashed a snow scene to the right hemisphere and a picture of a chicken claw to the left hemisphere. The patient picked the card with a shovel with his left hand and a card with a chicken with his right hand. When the patient was asked why he chose the shovel with his left hand he answered, “I saw a claw and picked a chicken, and you have to clean out the chicken shed with a shovel” [26]. He believed his explanation to be a statement of fact held with confidence.

In the same manner, Greene suggests, deontologists endorse emotionally driven deontological judgments by appealing to post hoc justification stories [2: 68]. People have strong negative emotional responses which tell them not to push the fat man in the footbridge dilemma, but in the trolley dilemma there is no such emotional reaction which, consequently, makes people believe it is permissible to sacrifice one life for five. Deontologists like Judith Jarvis Thomson [27] and Frances Kamm [28] appeal to a theory of rights to justify the permissibility of sacrifice in the trolley dilemma but not in the footbridge dilemma. People have strong emotional reactions to Singer’s drowning child scenario which makes them believe it is morally required to provide aid, but in scenarios with faraway needy children there is no such emotional reaction which, consequently, makes people believe the duty to give aid is rather weak. Deontologists like Colin McGinn [29] and Frances Kamm [30] appeal to a theory of duty to justify why there is a weaker obligation to aid faraway needy children. Kant finds masturbation disgusting which makes him believe it is morally wrong. To justify the wrongness of masturbation elicited by feelings of disgust, Kant argues that masturbation involves using ourselves merely as means [31].[6]

[6] Note that some examples misrepresent in part the deontologists’ views. For instance, Thomson’s main argument for the permissibility of sacrifice in trolley cases but not in footbridge cases appeals to the distinction between deflecting a threat and bringing about a new threat [27].

Which model does the evidence support?

Greene takes the presented cases of confabulation to parallel deontological theorizing. Deontologists have a strong emotional response, without knowing what triggered it, and make up reasons that comply with the emotional verdict. I will now analyze which model the evidence best supports.

The results from Haidt and colleagues seem to support the AEC model. The incest story elicits a judgement of moral condemnation which is driven by an alarmlike emotion such as disgust [32, 33]. However, subjects immediately abandon their justifications when features from the story are reinforced. They point out psychological and inbreeding risks, but after the experimenter reiterates the use of birth control, and the fact the story presupposes no psychological harm, they admit failure of justification. It is highly implausible that these results mirror the way in which deontological philosophy is done. Note that mere standards of rational criticism show the proposed justifications to be obvious failures. Despite the fact that subjects have a tendency to search post hoc for reasons, the proposed reasons are far from what we usually refer to as a plausible sounding or rational story. Haidt calls such rationalizations "moral dumbfounding" because it is immediately transparent that the reasoning involved is badly done. Subjects are neither confident about their justifications, nor do they provide at least *prima facie* plausible reasons. Haidt reports that after they retract the proposed justifications, subjects say something like "I don't know, I can't explain it, I just know it's wrong." This is indeed a confabulation driven by alarm-like emotions, but one that ultimately is not endorsed by the subjects themselves, and lacks *prima facie* plausibility. Therefore, this sort of confabulation violates premise (1) of the model.

The pantyhose and problem solving experiments support NC because the elicitor factor of behaviour is a non-affective cue. The right hand position effect, which influences the subjects' behaviour in the pantyhose experiment, is not an emotionally charged cue. Though, it is not clear why we have this bias, we can say that it is an unconscious order effect not related to emotionally valence cues, since all displayed items are identical. It might be a simple cognitive heuristic which indicates a preference for the most recent choice (the most right handed). Because there are no obvious relevant differences between items, the brain will send signals to pick the most recently analysed item, reducing the costs of choice. In the problem solving experiment, subjects look to the environment for bits of information relevant to the solution of the cords puzzle. As it works out the puzzle, the brain receives the subtle cue of pendulum-like movement, and processes it unconsciously as a possible solution. The cue is just suggestive information for problem solving, with no emotional valence.[7]

The posthypnotic suggestion experiment also fits the NC model because the result of fabricating a story is independent from the nature of the cue. As long as people do react to the posthypnotic suggestion, we can observe whether they manifest a tendency to confabulate. The operator could have hypnotized the subject to initiate an action irrespective of whether he saw a short emotional video or a short boring documentary. Many different types of cues (sound of a click, leg movement, etc.) can be used to trigger behaviour, and many different suggestions can be induced through hypnosis (movements, pain, hunger, thirst, muscle

[7] Though I take these findings at their face value, there are objections that they do not meet adequate criteria for awareness assessment. For example, the pantyhose experiment fails the relevance criterion which states that assessments of awareness should target only information relevant to the behaviour [34].

relaxation) [35]. The revealing point about posthypnotic confabulation is that it distinctively suggests a lack of reluctance to fabricate a story when we do not know what triggered our behaviour, while indicating that it does not matter what the triggering factor is, as long as it can be induced through hypnosis. Thus, the neutrality of the predictive factor from premise (2) pertains to the NC model.

Although the experiment devised by Dutton and Aron is based on an arousal inducing condition, it raises some questions. The experience of strong fear, while crossing the bridge, did made subjects more than twice as likely as the control subjects to ask the experimenter out, but the experiment implies that the subjects' choice was also based on the conscious cue of attractiveness. Control subjects also asked the experimenter out. The subjects misattributed the degree of attraction to the woman only *in part*, not the sexual attraction itself. The results show that there was a mixture of unconscious and conscious factors (attractiveness), which had important influences on subjects' choices. The design of the experiment builds in a predictive factor that is part of the subjects' reasons for actions, thus violating the stronger version of premise (4). But since subjects do misattribute part of the sexual attraction, this case conforms to the weaker version of premise (4), from which it follow that their judgement is based on both genuine reasoning and rationalization.

The use of reported cases of confabulation in patients with conditions such as Korsakoff's amnesia, related memory disorders and split-brain syndrome face a methodological problem. These are pathological cases which cannot be used to suggest something for non-pathological contexts such as philosophical theorizing. Greene needs to use everyday non-pathological cases of confabulation if we are to take his empirical hypothesis seriously. There is much indication that normal and pathological confabulations might not only differ in severity, but also in their mechanism [36, 37]. And even if we admit these pathological cases, they do not involve confabulations driven by alarm-like emotions.

I have argued that almost all of Greene's cases of confabulation illustrate the NC model. The other evidence is either irrelevant because the reasons invoked were not "a plausible sounding story", or weak due to the fact that the rationalizer's factor in part predicts his behaviour. At best, it can arguably suggest that deontology involves both genuine reasoning and confabulation when additional influencing factors are unknown and hard to be identified. Greene might reply that the confabulation argument can be used independently of whether the actual cause of confabulatory behaviour is emotional or cognitive. I did say that the evidence for premise (2) can also be taken from the dual-process theory of moral judgement. Thus, it is enough to show that if some ethical theory justifying some moral judgment appeals to factors that are very different from what actually caused that judgment, then that theory is likely to be based on an ill-grounded process, such as confabulation. If we grant Greene's direct evidence for the emotional source of deontological judgments, then one can argue that it does not matter so much that established cases of confabulation don't fit the AEC model. While this reply has some appeal, it only points out a puzzling difference between what causes CD judgments (emotional) and what deontologists use as factors (cognitive) to justify CD judgments, that can be explained by the confabulation hypothesis, but it does not diminish the plausibility of other explanations. The coincidence thesis can be explained as a co-variation between factors, or even though the evidence is direct, it does not rule out other (cognitive) sources of CD judgements.

Though this is not decisive criticism, it would have been important for Greene if paradigmatic cases did illustrate alarm-like confabulations, as this would suggest that confabulation could plausibly be involved in deontological arguments. Since the current research on confabulation tends to favour the neutral model, it is puzzling to expect outright alarm-like confabulations in philosophical theorizing. If paradigmatic cases illustrated the alarm-like model then it would not seem mysterious that it occurs among deontologists.

Conducive conditions for confabulation

However, I did granted earlier that it might be enough to show that if some ethical theory justifying some moral judgment appeals to factors that are very different from what actually caused that judgment, then that theory is likely to be based on confabulation. Now I will suggest a deeper reason as to why Greene's hypothesis is problematic, by analyzing the conditions that make confabulation likely to occur, and determining whether the psychological underpinnings of deontology favour, at the very least, the *tendency* to confabulate. We have to ask whether the tendency to confabulate matches well with the alarm-like profile of deontology.

Greene believes that the existence of moral emotions in conjunction with the irrepressible tendency to justify of our own behaviour gives rise to deontological philosophy, whereas consequentialism is not likely to be subject to confabulation because it is preferentially supported by cognitive processes [2: 60-63]. The suggestion is that alarm-like emotions favours confabulation, whereas cognitive processes do not. In this section, I will argue that there is an inherent tension between conditions which are conducive to confabulation, and Greene's psychological profile of deontology, which also explains why paradigmatic cases do not fit the alarm-like model.

There are many conditions that favour confabulation. Reliance on compromised information combined with assumed veracity can lead people to fabricate stories [38]. People can also confabulate when they are pushed to recall memories in more detail than actually stored [39, 37]. Confabulation is proven to be strongly related to the inability to withhold answers, and to monitor one's own responses [41]. Other experiments show that the associative strength between original pieces of information and slightly different ones not originally presented is significantly correlated with the production of memory confabulation [41, 37]. But what is of interest here is the conditions that favour confabulation in paradigmatic cases highlighted by Greene, such as the pantyhose experiment.

Nisbett and Wilson [21, 42] state that, in general, accurate reports about "knowing why" occur when (a) an influential stimulus is salient and (b) it is a plausible cause of the response. Vice versa, "knowing why" confabulation occurs when stimuli are not salient and are not plausible causes. In the pantyhose experiment, the position of the items on the display is not a salient influential stimulus because the subjects focus on the items themselves, and if there is nothing relevant about it, then people do not see any special reason to pay attention to it. Also, the position of the item is not a plausible cause for liking items because people usually base their preference on quality criteria or features that seem relevant to them. As Nisbett and Wilson acknowledge, "the position of stockings in an array does not seem a plausible reason for liking stockings" [42: 129]. These two conditions imply that the more an influential stimulus is salient, and a plausible cause, the less likely it is for confabulation to occur.[8] Correspondingly, the more an influential stimulus is a subtle cue, and an implausible cause, the more likely it is for confabulation to occur. Therefore, in sacrificial dilemmas in which Greene contrasts deontological and consequentialist judgments, we should expect people to be quite aware of their strong emotional reactions.

Let's analyse the implication for the scenarios used by Greene in his experiments. In the footbridge dilemma, contemplating pushing an innocent man with our bare hands elicits an alarm-like feeling of wrongness, which is very hard to override. To express this feeling, according to Greene, people use a deontological concept. It is wrong to push the fat man to his death because it would be a violation of his right to life. While the response is not caused by the thought that people have rights, the emotional reaction to pushing an innocent man to his death is highly salient in our phenomenal awareness. This is a case in which we can easily be aware that what causes our response is an emotional reaction to the aversive idea of killing an innocent human being with our bare hands. It is documented that people use what has been called the contact principle which states that using physical contact to cause harm to a victim is morally worse than causing equivalent harm to a victim without using physical contact. Cushman et al. [43] showed that subjects were typically able to articulate the contact principle, but were hesitant to endorse it as morally valid. The emotional contact factor is relatively accessible to consciousness because it is both a highly salient and a plausible cause. Alarm-like

[8] This explains why the case of sexual attraction confabulation, which involves an alarm-like emotion, can be misleading. The confabulation occurs because the fear experienced on the scary bridge is an implausible cause of heighten sexual attraction.

emotions may, in fact, facilitate accurate introspective reports about the bases of our judgments.

Consider now maybe the most emotionally salient dilemma in which Greene claims that CD judgments are caused by alarm-like emotions [4]. In the crying baby dilemma you are hiding with several other people from enemy soldiers. Your baby starts to cry loudly, and if you do not cover your baby's mouth the soldiers will find you and kill everyone. But if you cover your baby's mouth, that will kill him. Is it morally permissible to do this? Contemplating the action of covering the baby's mouth to his death elicits a strong aversion which makes people endorse the impermissibility of killing. In this case it is even much easier to articulate why knowledge, in comparison with people's ability to articulate the contact principle in the footbridge dilemma. Given the fact that an aversion to kill one's child is more salient than the aversion to kill an innocent person, we should expect people to be quite aware of the causes of their beliefs of actions. Between parents and children there is a special relationship of caring that makes the overwhelming majority of parents immediately pay attention to any possible harm that may endanger their children's wellbeing. This makes the influential stimulus a plausible cause of the response, thus satisfying the second condition for accurate reports about why knowledge.

It seems unlikely that highly salient cues, such as the aversion to bring harm to our own children, have no part in our intentional and phenomenal awareness when we report our inner mental workings. Therefore, it is implausible to expect people to have alarm-like emotions that are activated by tragic conditions and not know what the tragedy is about, and what the alarm-like emotion is referring to. When certain features of a situation capture our attention to an extensive degree and influence the way we respond, it is highly likely we can have accurately introspective reports about their presence and influence. Moreover, it should be easier to have accurate introspection if the emotional cues are fairly simple ones elicited by determinate features that do not involve decision making under uncertainty [40]. The cry baby dilemma, to use Greene's words, elicits simple emotions that are "blunt biological instruments" and there is no decision making under uncertainty (either we kill the baby and save the rest, or resist the killing option and, as a consequence all, will die). It seems there is no reason to expect a complete lack of accessibility to our emotional and cognitive processes, especially when we are experiencing one of the most alarm-like emotions from an evolutionary perspective, i.e. threat to offspring.

By contrast, we should expect confabulations to occur when the influential stimulus is a subtle cue and not a plausible cause of the response. Remember that the stimuli in the pantyhose experiment and cords puzzle are subtle cues which are not central to one's attention. In the pantyhose experiments, subjects focus naturally on the items themselves, having no special reasons to look for cues which are usually irrelevant for ones choices, such as the position on the display. In the cords puzzle, subjects focus naturally on assessing the explicit components of the puzzle, having no special reasons to pay attention to the experimenter actions since he is assumed to be separate from the activity.

Greene claims that "the kind of emotion that is essential to consequentialism is fundamentally different from the kind that is essential to deontology, the former functioning more like a currency and the latter functioning more like an alarm." [2: 41] The emotions that drive deontological judgments are less subtle and consist in blunt "alarm bells" that issue simple commands such as "Don't do it!" or "Must do it!" which dominate our decisions. On the other hand, the emotions that drive consequentialist judgments are more subtle and only influence our decisions by saying “Such-and-such matters this much. Factor it in.”, because, Greene argues, consequentialism relies primarily on the process of weighting harms and benefits, and is accountable in the end to evidence of what promotes the greater good [2: 41; 16: 304].

The analysis provided so far implies that occurrence of confabulation is at odds with the psychological underpinnings of deontology. Moreover, it implies that in some cases it is more likely for confabulation to be linked with the psychological profile of consequentialism. The influence “Factor it in” fits much better with what is happening in the cords puzzle than the command “Must do it!” As the subjects work out possible solutions, they are not aware of cognitive processes that analyse different sources of information, e.g. the experimenter’s actions. The brain picks the subtle cue from the larger environment and issues the unconscious suggestion to factor it in as a possible solution. If a highly salient cue dominates our decision-making, and is a plausible cause of action, then we are much more likely to consciously spot it and to report correctly the factors that influence our decisions. Also, defining consequentialist reasoning as a weighting process of harms and benefits and accountable to evidence of what promotes the greater good does not make it incompatible with confabulatory behaviour. Such an indeterminate process might actually enhance confabulation tendencies because it is very difficult to have precise introspective reports about to what degree certain cues influences decisions. Slovic and Lichtenstein [44] documented significant discrepancies between the explicit weights assigned post hoc by judges, and the implicit weights they placed on cues as shown by regression modelling. Judges tended not only to overestimate the importance of minor cues but also to underestimate their reliance on major cues.[9] It is a simplistic assumption to believe that evidence of what promotes the greater good is always clear cut, so that it forces reasoning to a certain conclusion. Maybe in simple cases, consequentialism is strictly dependent on evidence of best outcomes, but when dealing with more intricate social issues one is faced with decision making under uncertainty, evidence needs to be weighed against counterevidence, and so on.

It is not the existence of highly salient emotions which straightforwardly cause behaviour that gives rise to confabulation, but subtle, indeterminate and indirect cues, especially in uncertainty conditions. Our introspective awareness is not good at picking up on these and determining their influence. Understanding what is conducive to confabulatory behaviour suggests that it is incompatible with Greene's profile that the "psychological essence" of deontology consists in alarm-like emotions. Moreover, the likelihood of confabulation is not diminished by the cognitive profile of consequentialist reasoning. It is much easier to report a salient and straightforward cue that dominates our decisions, rather than a subtle cue that merely influences it. The upshot is that when someone experiences alarm-like emotions in reaction to cases that do not involve decision making under uncertainty, she has transparent access to what actually caused her judgment, thus allowing her to report correctly the factors that influenced him. She need not cite factors that are very different or unrelated from what actually caused her belief.

[9] Curiously, even though Nisbett and Wilson [21] present such cases, Greene does not mention them.

What kind of epistemic claims does the confabulation data support?

Greene would probably accept that consequentialist reasoning is not immune to confabulation. Consequentialism may be subject to confabulation, but only in how the principle is applied, as shown in the above example, not in fabricating the principle itself. In contrast, deontological theories of rights or duties are characterized as confabulations in themselves [2: 68]. Certain emotions compel us to believe that some actions are forbidden, and then, Greene claims, deontologists make up a rationally appealing story about rights: "There are these things called "rights" which people have, and when someone has a right you can't do anything that would take it away." [2: 63] It seems deontologists stipulate the existence of rights that correspond to our alarm-like feelings of wrongness. Greene describes the doctrine of double effect as a "normatively ugly bride" or Kant's categorical imperative as an "esoteric justification", challenging the validity of the principle itself [8; 16: 301]. In other places, he surprises the reader with expressions such as "appeals to rights function as an intellectual free pass" [16: 302]. This seems to suggest a different claim, namely that the appeal to rights to defend certain normative positions, not the rights *per se*, is ill-grounded or abusive.[10] However, the latter claim does not undermine deontology as a school of thought, as this seems to be Greene's main objective.[11]

[10] Deontologists are well aware of this, having criticized the abusive appeal to rights and, even, the inflation of rights [45].

[11] The appeal to consequentialist principles of maximizing the general utility can also function as an intellectual free pass. Authoritarian regimes often justify questionable policies by appealing to the priority of the overall good.

Having highlighted this diversity in Greene's points, in what follows I will clarify what kind of claims the confabulation data can support. More specifically, I will point out that the phenomenon of confabulation challenges the validity of a particular application of a body of knowledge not of the content of that knowledge itself, and that even if a justification is a confabulation in a particular case it does not follow that it lacks epistemic merit in general. The upshot of the clarification is that confabulation data cannot be used to undermine the validity of deontological theory *in itself,* and ironically, if one commits to the claim that a deontological justification is a plausible sounding confabulation in a particular case, then the data suggests that deontology has a *prima facie* validity.

Return to Greene's example of romantic choices. Alice believes that her choices are based on personality features such as wit and charm, but the prediction factor is a height preference. What is the epistemic status of her belief? Citing personality traits seems to be a reasonable explanation of romantic choices. Suppose that Alice holds with conviction that charm was an important factor, and that all her partners showed no signs of charm. Suppose, also, that her partners *did* show some signs of charm, but what made the difference among candidates was the unconscious influence of height. In the no-charm version, the explanation is not a plausible story, but in the charm version her justification is a plausible narrative.

Now, the two versions show a clear epistemic difference between Alice's beliefs. The first one is similar to pathological cases of confabulation in which the response may be coherent and sensible but it is obviously false. The patient who was sitting near an air conditioner confabulated that he was in an air-conditioning plant. The second version is similar to everyday confabulation in which the response is a plausible story, but it is not obviously false. In the pantyhose experiment, subjects explained their preferences in terms of quality features (superior knit, elasticity), even though all items were identical, but since it was difficult to see if the items were identical, they were pushed to assume subtle differences. The difficulty to spot subtle differences or no differences poses serious obstacles for both the subjects and third parties to see if the rationalizations are false.

Differences in "truthfulness" may be an explanation as to why abnormal confabulations are salient, and why normal confabulations go undetected [38]. In abnormal cases, confabulations are obviously false, whereas in normal cases, confabulations are *prima facie* plausible. The key question is why normal cases seem plausible enough to bypass immediate scrutiny. Where do people's truthful confabulations come from? They sound plausible because they are based on observations of covariation between one's responses and prior conditions, shared knowledge and norms [21, 42, 12, 38]. Most likely, people do not even try to investigate through introspection the causes of their behaviour, or the degree of influence of certain cues on their choices, but make, instead, causal judgments based on previous experiences and confirmed expectations [28, 21, 42]. For example, in one experiment, subjects were given an injection of adrenaline without their knowledge and this made them unable to sleep. They attributed the inability to sleep to nervousness about what will happen the next day [17]. The response could be based on previously similar situations in which the subjects experienced anxiousness the night before dealing with exams or work-related issues. Referring to Tversky and Kahneman's work [46], Nisbett and Wilson [21, 42] suggest that in such cases, subjects use the representativeness cognitive heuristic to assess whether an effect is similar to the stimulus, or is similar with the effects a given stimulus is expected to produce. The negative effect of anxiousness is thus assessed in light of the negative stimulus of stressful situations, or of similar effects usually caused by such stimulus. In normative contexts, a healthy confabulation includes shared norms and beliefs, which makes possible the coordination and dialogue within a community [38: 218]. Take for example Haidt's incest experiment. Subjects confabulate that incest is wrong because it involves inbreeding risks, even when contraception was used. Leaving aside the details of the scenario, almost all of us share the beliefs that inbreeding risks constitute a serious reason against incest.

This suggests that the *prima facie* plausibility of normal confabulations is based on shared knowledge and norms that have been proven accurate and appropriate, to a certain extent. If people's explanations and justifications were not as such, then it would be hard for normal confabulations to pass undetected as plausible-sounding stories. Therefore, it is the *deployment* of knowledge in *particular cases* that is ill-grounded in confabulation tendencies, not the *content* of the justifications or explanations in general. As Nisbett and Wilson put it, "Verbal reports relying on such theories will typically be wrong not because the theories are in error in every case but merely because they are incorrectly applied in the particular instance." [21: 248][12]

The clarification has ironic implications, as it shows that if we properly understand what confabulation consists in, then we cannot use confabulation data to claim that deontological philosophy is in general a faulty theory. If it were, the justification of particular deontological

[12] Although Kahneman is rather sceptical about intuitive impressions, he endorses similar lines: "Judging probability by representativeness has important virtues: the intuitive impressions that it produces are often – indeed, usually – more accurate than chance guesses would be." [47: 151]

judgements would not have passed as a plausible sounding story, and so become a confabulation. The evidence can only support an argument that deontology is incorrectly *applied* in particular cases. But in order for a deontological confabulation to be plausible-sounding story, it is crucial that people's judgments are guided by deontological norms in previous cases and that such norms are considered appropriate from a communal point of view. Admitting cases of deontological confabulation, therefore, implies accepting that, in general, deontology has some epistemic merit from which confabulations get their *prima facie* plausibility.

However, this does not rule out the possibility that people mistakenly believe in the epistemic merit of a body of knowledge. It might be the case, as it often has, that people endorse erroneous shared knowledge or that what is considered an appropriate norm turns out to be a faulty justification. But even though there is this possibility one needs separate evidence to prove it, not confabulation data. My analysis has focused on the implications which can be drawn from confabulation data alone. If a deontological confabulation is to sound like a plausible justification, then it has to involve some features which are *prima facie* valid, because the way healthy confabulation works is by picking up in its content features from shared knowledge and norms that people endorse in general. Thus, if Greene wants to challenge the validity of deontological theory then confabulation data is of little use. Also, if Greene wants to claim that a particular deontological justification is a plausible-sounding confabulation, then the data suggests a *prima facie* validity of deontological theory.

Acknowledgments I thank Guy Kahane, Gulzaar Barn and Bogdan Olaru for insightful comments and advice. I am also grateful to the two anonymous reviewers of this journal for thoughtful comments which helped to improve the manuscript. This paper is supported by the Sectoral Operational Programme Human, Resources Development (SOP HRD), financed from the European Social Fund and by the Romanian Government under contract number POSDRU/159/1.5/S/133675.

The final publication is available at Springer via http://dx.doi.org/ [10.1007/s12152-015-9244-5]

Before citing please consult the final publication.

References

1. Greene, J. D. 2014. Beyond Point-and-Shoot Morality: Why Cognitive (Neuro) Science Matters for Ethics. *Ethics* 124(4), 695-726.

2. Greene, J., D. 2008. The secret joke of Kant's soul. In Sinnott-Armstrong W. (ed.), *Moral psychology: Vol. 3. The neuroscience of morality*. Cambridge, MA: MIT Press.

3. Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. 2001. An fMRI investigation of emotional engagement in moral judgment. *Science* 293(5537), 2105-2108.

4. Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., and Cohen, J. D. 2004. The neural bases of cognitive conflict and control in moral judgment. *Neuron* 44(2), 389-400.

5. Kahane, G., and Shackel, N. 2010. Methodological issues in the neuroscience of moral judgement. *Mind & language*, 25(5), 561-582.

6. Mihailov, E. 2015. The argument from self-defeating beliefs against deontology. *Ethical Perspectives* 22 (4): 573-600.

7. Kahane, G. 2012. On the wrong track: Process and content in moral psychology. *Mind & language*, 27(5), 519-545.

8. Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., and Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, 111(3), 364-371.

9. Cushman, F., Young, L., and Greene, J. D. 2010. Our multi-system moral psychology: Towards a consensus view. *The Moral Psychology Handbook*, 47-71, Oxford: Oxford University Press.

10. Berker, S. (2009). The normative insignificance of neuroscience. *Philosophy & Public Affairs*, 37(4), 293-329.

11. Kamm, F. M. 1998. Moral intuitions, cognitive psychology, and the harming-versus-not-aiding distinction. *Ethics* 463-488.

12. Wilson, T. D. 2009. *Strangers to ourselves: Discovering the adaptive unconscious*. Harvard University Press.

13. Dean, R. 2010. Does neuroscience undermine deontological theory? *Neuroethics* 3(1), 43-60.

14. Wielenberg, E. J. 2014. *Robust Ethics: The Metaphysics and Epistemology of Godless Normative Realism*. Oxford: Oxford University Press.

15. Kamm, F. M. 2009. Neuroscience and moral reasoning: A note on recent research. *Philosophy & Public Affairs*, 37(4), 330-345.

16. Greene, J., D. 2014. *Moral tribes: emotion, reason and the gap between us and them*. Atlantic Books.

17. Hirstein, W. (2009). Introduction: what is confabulation?. In Hirstein, W. (ed.). *Confabulation: Views from neuroscience, psychiatry, psychology and philosophy*. Oxford: Oxford University Press.

18. Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., and Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, 107(3), 1144-1154.

19. Hume, David (1739/1978). *A treatise of human nature*. Oxford: Clarendon Press.

20. Haidt, J., Bjorklund, F., and Murphy, S. 2000. Moral dumbfounding: When intuition finds no reason. Unpublished manuscript, University of Virginia.

21. Nisbett, R. E., & Wilson, T. D. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological review*, 84(3), 231.

22. Maier, N. R. 1931. Reasoning in humans. II. The solution of a problem and its appearance in consciousness. *Journal of comparative Psychology*, 12(2), 181.

23. Dutton, D. G., & Aron, A. P. 1974. Some evidence for heightened sexual attraction under conditions of high anxiety. *Journal of personality and social psychology*, 30(4), 510.

24. Stuss, D. T., Alexander, M. P., Lieberman, A., & Levine, H. 1978. An extraordinary form of confabulation. *Neurology*, 28(11), 1166-1166.

25. Estabrooks, G. H. 1943. *Hypnotism*. New York: Dutton.

26. Gazzaniga, M. S., and LeDoux, J. E. 1978. *The Integrated Mind.* New York: Plenum Press.

27. Thomson, J. J. 1986. *Rights, Restitution, and Risk: Essays in Moral Theory*. Harvard University Press.

28. Kamm, F. M. 1993. *Morality and mortality. Vol. 1*. Oxford University Press.

29. McGinn, C. 1999. Our duties to animals and the poor. In Dale Jamieson (ed.), *Singer and His Critics*. Blackwell Publishers.
30. Kamm, F. M. 1999. Famine ethics: the problem of distance in morality and Singer's ethical theory. In Dale Jamieson (ed.), *Singer and His Critics*. Blackwell Publishers.
31. Kant, I. 1999. *Practical philosophy*. Cambridge: Cambridge University Press.

32. Haidt, J. 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review* 108(4), 814.
33. Kelly, D. R. 2011. *Yuck!: The nature and moral significance of disgust*. MIT Press.
34. Newell, B. R., and Shanks, D. R. 2014. Unconscious influences on decision making: A critical review. *Behavioral and Brain Sciences* 37(01), 1-19.
35. Oakley, D. A., and Halligan, P. W. 2013. Hypnotic suggestion: opportunities for cognitive neuroscience. *Nature Reviews Neuroscience*, 14(8), 565-576.
36. Hirstein, W. 2005. *Brain fiction: Self-deception and the riddle of confabulation*. MIT Press.
37. Schnider, A. 2008. *The confabulating mind: How the brain creates reality*. Oxford: Oxford University Press.
38. Wheatley, T. 2009. Everyday confabulation. In Hirstein, W. (ed.), *Confabulation: Views from neuroscience, psychiatry, psychology and philosophy*. Oxford: Oxford University Press.
39. Burgess, P. W., and Shallice, T. 1996. Confabulation and the control of recollection. *Memory*, 4(4), 359-412.
40. Mercer, B., Wapner, W., Gardner, H., and Benson, D. F. 1977. A study of confabulation. *Archives of neurology*, 34(7), 429-433.
41. Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. 2001. Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review* 8(3), 385-407.
42. Wilson, T. D., and Nisbett, R. E. 1978. The accuracy of verbal reports about the effects of stimuli on evaluations and behavior. *Social Psychology* 118-131.
43. Cushman, F., Young, L., and Hauser, M. 2006. The role of conscious reasoning and intuition in moral judgment testing three principles of harm. *Psychological science* 17(12), 1082-1089.

44. Slovic, P. and Lichtenstein, S. 1971. Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behaviour and Human Performance* 6:649–744.
45. O'Neill, O. 2005. The dark side of human rights. *International Affairs* 81(2), 427-439.
46. Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science* 185(4157), 1124-1131.
47. Kahneman, D. 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.