Neuroethics: A Renewed View of Morality? Intentions and the Moral Point of View

8

Bernard Baertschi

Abstract

In the traditional view of morality, intentions play a central role: they define what a typical action consists of and allow for the assignment of both blame and praise. Actions are intentional bodily movements, and if actions are morally assessed, it is first and foremost because they are intentional. Recently, several psychologists have investigated the neural basis of these mental phenomena. Although many studies confirm the traditional view, others point in the opposite direction: intentions play only a subordinate role in morality. For Joshua Knobe, intentionality is not central but depends on ascriptions of responsibility, far from grounding them. For Joshua Greene, moral judgement is based on intentions only when we rely on alarm emotions. If these studies are found to be convincing, it would oblige us to modify our view of morality: responsibility would be linked with outcomes rather than with intentions. On the legal level, the doctrine of *mens rea* would also be modified, and perhaps even abandoned. Neuroethics would then be a field that purports to offer a renewed view of morality. However, I think that a careful examination of the data and their interpretation shows that this conclusion is mistaken: intentions remain at the centre of morality even if it is not easily noticed in some situations, especially when side effects are involved.

B. Baertschi
Institute for Ethics, History, and the Humanities, University of Geneva, Geneva, Switzerland
e-mail: bernard.baertschi@unige.ch

E. Racine, J. Aspler (eds.), *Debates About Neuroethics*, Advances in Neuroethics,
DOI 10.1007/978-3-319-54651-3_8

8.1 Introduction

In the traditional view of morality, intentions play a central role: they define what a typical action consists of and allow for the assignment of both blame and praise. Actions are *intentional* bodily movements, and if actions are morally assessed, it is first and foremost *because* they are intentional. For example, accidental harm does not arouse anger, but intentional harm does, and only the latter causes moral reproach. With the progress of neuroscience and brain imaging, several psychologists have investigated the neural basis of these mental phenomena. Many studies confirm the traditional view, but others point in the opposite direction: intentions might only play a restricted and subordinate role in morality. This constitutes a challenge to the traditional view that has been raised in particular by Joshua Knobe and Joshua Greene, two influential neuropsychologists. If these studies are found to be convincing, it would oblige us to modify our view of morality, perhaps profoundly. Neuroethics would then be a field that purports to offer a renewed view of morality. However, are these studies convincing? This is the question I try to answer in this paper.

It demands first that I carefully spell out what the traditional view of morality says about intentions. Sections 8.2 and 8.3 are devoted to this task. In the first section, I argue that intentions are an essential part of what constitutes an action: strictly speaking, an unintentional action is a contradictory expression. Actions are morally judged, but intentions are also crucial to assign responsibility, as I show in Sect. 8.3: in ordinary circumstances, we blame and praise what is done intentionally, not what happens accidentally. In Sect. 8.4, I present several empirical studies confirming the traditional view: intentions are really at the centre of morality, and when people judge that they are not, it is often because they are afflicted by some psychiatric or neurological conditions. In the two following sections, I examine challenges based on experiments to this traditional view of morality. Joshua Knobe (Sect. 8.5) has claimed that intentionality does not ground attribution of responsibility: on the contrary, we see an action as intentional only if we already consider its agent as responsible (this phenomenon is called the *Knobe effect*). For Joshua Greene (Sect. 8.6), our morality does not represent a seamless whole and is formed of two parts: the first (System 1) is intuitive and assigns a central role to intentions, but the second (System 2) does not – it is rational and places most of the moral weight on outcomes (i.e., the important factor is not whether we behave intentionally, but that some good is done).

Knobe and Greene's arguments place the traditional view of morality in jeopardy, and they suggest it is in need of revision. But are their arguments convincing? In Sect. 8.7, I argue that they are not, since a thorough reinterpretation of the studies they rely on suggests that their evidence has only limited validity: it concerns not all actions – only actions with side effects. Consequently, it is not the intentionality of the actions themselves that is concerned but only the intentional character of the side effects (i.e., an action can have several effects, and not all of them are willed – some are even unforeseen). In the last section, I add some precisions in discussing the *doctrine of double effect*, a view that has been proposed in traditional morality to manage actions with side effects.

8.2 Intention, the Criterion of Action

"I had intended to visit my mother after lunch, so I could not attend your meeting". Such an utterance is commonplace, and it is not difficult to imagine a context where it would be completely appropriate. An intention is a mental act, akin to a decision.[1] In ordinary language, an intention is often less strong than a decision: I am disposed to reconsider what I intend to do, much more than what I have decided to do. But both can be described as mental acts aiming at a goal – here, a visit to my mother.[2]

"I went to a meeting, when I happened to run into my mother. It was not intentional because I did not know that my mother was in town". This encounter with my mother is fortuitous and, consequently, not intentional. Of course, when I stumbled upon her, I was fulfilling an intention – to attend a meeting – but I did not aim to see my mother there.

Intentions are about something (i.e., we intend to do something). *To be about something* is what philosophers call "intentional", and Daniel Dennett warns us: "This *aboutness* that, for example, sentences, pictures, beliefs, and (no doubt) some brain states exhibit, is known in philosophical jargon as *intentionality*, an unfortunate choice as a technical term, since outsiders routinely confuse it with the everyday idea of doing something intentionally" (Dennett 2013: 62). It is particularly misleading for actions, because they are intentional in both senses: they are about something and accompanied or shaped by an intention.

What interests me in this paper is the dual property of typical actions: to be about something while being directed by an intention.[3] This dual property in fact covers two properties, since they can be separated in some pathological conditions like *anarchic hand syndrome*, where the patient observes his hand moving purposely (it aims at an object), but without having had any intention to move it: he discovers his hand's "action" (Marcel 2003). However, what I am interested in here are "normal" actions, and not what would be better named "pseudo-actions". Such normal or typical actions must also be recognised as *mine*; this sense of ownership is crucial (Forest 2014: 99–105).[4]

---

[1] Mental acts like decisions have been extensively studied by neuropsychologists since Benjamin Libet in the debate concerning free will (Fried et al. 2011). My subject focuses on another important point in action theory, unrelated to the free will debate.

[2] Here, I follow Franz Brentano, who said: "Each mental act is primarily directed to an object" (Marek 2013). For a mental act, to be directed to an object and to have a goal are synonymous.

[3] Rigato and col. notice: "What philosophers call 'intentional,' neuroscientists call 'goal-directed'" (2014: 181). However, everything that is goal-directed is not intentional in the relevant sense, but this terminological difference is not important for my argument.

[4] There exist other kinds of nontypical actions, like impulsive actions, which seem to be intentional only in the sense of aboutness, actions performed under coercion or actions made while sleeping, during an episode of REM sleep behaviour disorder (Maoz and Yaffe 2015; Cerri 2016). I will not investigate them.

This dual property is constitutive of actions, which Donald Davidson noticed.[5] To see this, let us imagine four situations where a roofer causes the death of a pedestrian.

- John is working on a roof. Suddenly, he trips and falls on a pedestrian. The pedestrian softens the fall, saving John's life, but is killed.
- Andrew is working on a roof. When he passes a tool to a workmate, he loses his balance and falls on a pedestrian. The pedestrian softens the fall, saving Andrew's life, but is killed.
- Paul is working on a roof. Suddenly, he has a vision: God orders him to kill a nearby pedestrian. Paul throws himself off the roof and lands on the pedestrian, who is killed.
- Peter is working on a roof. He sees his worst enemy walking down the street. Peter throws himself off the roof and lands on his enemy, who is killed.

Intention – and responsibility, as we soon will see – is understood differently in each of the four cases. Only Paul and Peter fall intentionally; Andrew's fall is a non-intentional effect of a former intention: to pass a tool, and what John does contains nothing intentional at all. We are even tempted to say that John does nothing and that things happen to him against his will: John does not perform an action. Notice that I use here the expressions "to perform *X* with an intention to do it" and "to perform *X* intentionally" as synonymous; I will generally follow this use, in agreement with what has been named the *Simple View,* that is, the thesis "that anyone who *A*'s intentionally intends to *A*" (McCann 1991: 205).

John's case allows a distinction to be drawn between an *action* and an *event*. What distinguishes them is the presence of an intention in the former. Every action is intentional, which means, as Elisabeth Anscombe suggests, "intentional under some description that we give (or could give) of it" (Anscombe 1963: 29).[6] What John does cannot be described as intentional, even if we try hard. We can also say that the goal of an action should be identical with the content of the corresponding intention: the goal of Peter's action and the content of his intention are that his enemy will be dead. This analysis is meaningless for John.

8.3 Intention, the Prime Bearer of Responsibility

What happens in the four cases above is bad, given that a man dies as a result. Therefore, questions of ethics are relevant. More precisely, when someone is harmed, the question of who bears responsibility arises, and to answer that question,

[5] "I follow a useful philosophical practice in calling anything an agent does intentionally an action" (Davidson 2002: 5).

[6] Every bodily movement can be described in different manners. Consequently, an intentional action can be described without any reference to the intention, but of course, it does not deprive it of its intentional character: a bodily movement is intentional – it is an action – if there exists a description of it mentioning an intention; otherwise, it is (a part of) an event.

it is necessary to take intentions into account. John and Andrew do not kill the pedestrian intentionally, because they do not aim at the death of the pedestrian. Therefore, they are not morally responsible for the pedestrian's death. They are causally responsible for it indeed, but not morally.[7] Paul acts intentionally (he wants to kill the pedestrian), but without being morally responsible for his actions because of a cognitive impairment (i.e., he suffers from a hallucination): intentionality is not the only condition that needs be met when determining someone's responsibility. Only Peter is fully responsible and, in this case, blameworthy.

In order to blame or to praise someone, we have to consider intentions, because we are first and foremost responsible for what we aim at consciously and willingly, and moral responsibility is a prerequisite for blame and praise. Sometimes, we hear that consequentialists disagree, because they only take outcomes into account. But this is false, as John Stuart Mill adamantly stated: "There is no point which utilitarian thinkers (and Bentham pre-eminently) have taken more pains to illustrate than this. The morality of the action depends entirely upon the intention – that is, upon what the agent *wills to do*" (Mill 1991: 150, n. 1), but he adds that we should not confuse intention with motive, that is, we should not confuse what we intend to do with its cause. If I rescue a famous drowning person, my intention is to save her, but my motive can be that I want to be rewarded. I will insist on this distinction in Sect. 8.5.

In criminal matters – but not in civil law – the lesson is the same: the doctrine of *mens rea* in common law is an expression of it. A defendant cannot be culpable if he did not have a bad intention when he acted. Michael Treadway and colleagues showed in a study that graphic descriptions of harmful acts amplify amygdala activity and willingness to punish, but only if the act is intentional. They conclude that "Justice […] requires that punishment takes into account not only the negative emotions elicited by harm, but also an evaluation of the transgressor's intent" and comment: "An actor's mental state – whether it is purposeful, knowing, reckless, negligent or blameless – can markedly affect how severely he or she is punished for the harm committed" (Treadway et al. 2014: 1270).[8]

The centrality of intention in morality has a developmental signature: studies have shown that children are very soon aware of the distinction between what is intentional and what is not. Chris Frith, recalling how important it is to discriminate between intentional and non-intentional movement, adds that very young children can do just that:

> "It is important for us to make a distinction between deliberate actions and accidents. If my arm movement accidentally spills the wine on you then everyone is very compassionate about my embarrassment. But if, with much the same arm movement, I do it deliberately, the action is meant, and taken, as a severe insult. Infants as young as 9 months can distinguish between deliberate and accidental actions made by other people, for example, whether the toy was withheld deliberately or dropped accidentally" (Frith 2010: 13).

[7] I leave here the question of negligence or carelessness aside.

[8] Notice that *lex talionis* (an eye for an eye, a tooth for a tooth) and many ancient laws focus on action's effects rather than on intentions.

Here, Frith uses the expression "deliberate". As we can already see, many words express the same idea: willed, voluntary, intentional and deliberate. We can also add: knowingly, purposely and probably others.[9] Frith contrasts deliberate with accidental, but it can also be an antonym for impulsive, opening another semantic repertoire. Depending on the context of an utterance, these expressions are strictly synonymous or not. We will see in Sect. 8.5 that this has some importance, when I examine and discuss the *Knobe effect*.

8.4 Empirical Confirmations

The central place of intention in our traditional view of morality has been highlighted and buttressed by several experimental studies. I have already mentioned some of them, and I will now examine three others more thoroughly.

The first has to do with the *Ultimatum Game*, a well-known economic experiment assessing our sense of fairness. Someone (*B*) receives a sum of money from *A*, but *B* can keep it only if he transfers a part to a third person (*C*) and if *C* accepts the gift. If *C* refuses, the sum is entirely given back to *A*. Usually, *C* refuses when *B* gives him less than 30% of the sum.

Berna Güroglu and colleagues have introduced a modification to the game, in order to test the influence of intentions. *B* is instructed to give 20% of the sum. However, *C* is told that *B* can choose between three different offers: to give 20%, 50% or 80% of the sum, so that *C* believes that *B* makes the 20% offer freely. Consequently, *C* refuses very often (75% of the time). Later, *C* is informed that *B* has had no choice and did not intend to give the lesser sum; consequently, C's refusal drops (only 30%). The authors of the study conclude that the refusal is, in large part, a consequence of the perception of *B*'s intention. The moral quality of the intention counts more than the fairness of the result (the offer): "Information that highly influences fairness judgments is intentionality, that is, perceptions of fairness are influenced by the intentions of the interaction partner. A seemingly unfair act might evoke less negative affect if one believes that it was not done intentionally" (Güroglu et al. 2010: 414). These findings are robust, since they have been confirmed by other studies (Güney and Newell 2013).

The second experimental study concerns patients with ventromedial prefontal cortex (VMPFC) damage, who are individuals that suffer from a brain lesion responsible for a serious neuropsychiatric condition: acquired sociopathy. It was conducted by Liane Young et al. (2010). They presented four scenarios to their subjects (VMPFC patients and typical people), where the intention of the agent and the outcome of the action vary in an ordered manner. Grace (the agent in the four scenarios) has two opposite intentions, one bad (to poison a friend) and one neutral (to offer him sugar), and her action has two opposite results, one bad (her friend is poisoned) and one neutral (her friend is fine). When combined, we obtain four possibilities:

[9] Legal systems consider some of these expressions to be synonyms (Zangrossi et al. 2015: 2).

1. Grace thinks the powder is sugar. It is sugar. Her friend is fine.
2. Grace thinks the powder is sugar. It is toxic. Her friend dies.
3. Grace thinks the powder is toxic. It is sugar. Her friend is fine.
4. Grace thinks the powder is toxic. It is toxic. Her friend dies.

Asked to judge Grace's act, typical people evaluate 4 (the successful attempt to harm) as the worst scenario, followed by 3 (the failed attempt to harm). Scenario 2 (accidental harm) is considered unfortunate, but not so bad. On a scale where 1.0 means strictly forbidden and 7.0 means completely permissible, they put the successful attempt to harm (Scenario 4) at 1.1, the accidental harm (Scenario 2) at 3.5 and the failed attempt (Scenario 3) at 2.2. These participants believed that it is morally worse to have bad intentions than to have good ones that lead to bad consequences.

With VMPFC patients, situation 4 remains the worst scenario, but 2 and 3 are assessed differently: accidental harm is ranked at 3.1 and the failed attempt at 5.0. Accidental harm is then worse than the failed attempt to harm for these patients. Young and colleagues comment: "Notably, VMPFC participants also judged attempted harms as significantly more permissible than accidental harms" (Young et al. 2010: 848), because they consider the result to be of greater moral importance – when a bad intention does not succeed, it is not so serious (even if good intentions followed by good results are judged better, 6.0 on the scale).

Both groups of subjects (brain-damaged patients and typical people) evaluate actions and intentions. For typical people, the value of the action is significantly determined by the intention of the agent and the intentional character of the outcome. For VMPFC patients, the intention plays a less central role. Young and colleagues comment: "Patients with bilateral damage to the VMPFC were more likely to deliver utilitarian moral judgments" (Young et al. 2010: 845). But utilitarianism is not a neurological condition! Does it mean that utilitarians have a mistaken conception of morality? A utilitarian would answer that an empirical study shows what people think, not what they ought to think and that ordinary people are perhaps wrong when they place so much importance on intentions in the assessment of responsibility and wrongness. I will address this question later, but for the moment, I continue with my analysis of traditional morality, which is summarised in the following way by Young and colleagues: "A fundamental component of normal moral judgment is the ability to blame those who intend harm, even when they fail to cause harm. […] The ability to blame for failed attempts not only features prominently in mature moral judgments but emerges quite early in development" (Young et al. 2010: 849). Moral education or maturation is nevertheless necessary to render normal human beings more charitable: bad outcomes tend to tip the balance unfavourably for children, but in the end, intentions win the game of responsibility assessment and consequently of blame and praise.

Chris Frith has suggested "that the cognitive basis for the feeling of responsibility is, first, a mechanism that binds intentions to outcomes" (Frith 2014: 139).[10] The

[10] See also Christensen and Gomila (2012: 1259), and Yoshie and Haggard (2013).

study by Young and colleagues confirms this claim, even if people with different capacities bind them differently.

A third study also supports this conclusion. This study has been conducted with Asperger's patients by Joseph Moran and colleagues. The participants read vignettes combining intention and outcome, as in Grace's scenarios (*no harm*, neutral intention and neutral outcome; *accidental harm*, neutral intention and bad outcome; *attempted harm*, bad intention and neutral outcome; *intentional harm*, bad intention and bad outcome). The participants had to pass moral judgement on the four types of scenarios. Moran and colleagues observed that Asperger's patients do not place the same importance on intentions in morality when compared to typical adults. If "neurotypical adults weigh a person's intention more heavily than the outcome of their action when evaluating the moral permissibility of an action" (Moran et al. 2011: 2688), it is not the same for Asperger's patients. They rate accidental harm lower on the scale of permissibility (1.0–7.0): 3.5 for Asperger's patients and 4.8 for typical adults.[11] Like VMPFC patients, Asperger's patients favour outcomes, but not in the same manner: if the former exculpates bad intentions when the result is neutral (i.e., when no harm is done), the latter blames innocent intentions if the result is bad. It reminds us of what Young and colleagues said of children, as Moran acknowledges: "In several respects the pattern of results displayed by the adults [with Asperger's] mirrors that displayed by typically developing children" (Moran et al. 2011: 2690).

The picture of morality emerging from these studies is on par with our intuitions: morality is centred on intention or the intentional aspect of our deeds. Intention is necessary for distinguishing actions from events: actions are purposeful, whereas events occur independently of goals. Some actions, but not all, relate to morality, notably actions involving a harm. However, every harm is not wrong: in order to be wrong, a harm should be intended and not forgivable. Consequently, intentionality is also an important ground for moral responsibility, blame and praise. Children and some patients have difficulties with this, but not "neurotypical" adults.

However, two different objections have been directed against this view. The first contends that a certain phenomenon (the Knobe effect) casts some doubt on it, whereas the second claims that intentions play a major role only in one kind of moral judgements. I examine these objections in the next two sections.

8.5 The Knobe Effect

Joshua Knobe (2003) imagined two scenarios where an assignment of intentionality is made; he has a surprising result for our traditional conception of morality and responsibility. In each scenario, the vice president of a company proposes a new program to the chairman of the board which will have two effects:

[11] When the outcome is not bad, Moran and colleagues did not observe any significant difference between the moral judgements of Asperger's patients and typical adults (Moran et al. 2011: 2690).

- First scenario: the profits of the company will rise (first effect), but the environment will be severely harmed (side effect).
- Second scenario: the profits of the company will rise (first effect), but the environment will be benefited (side effect).

In both scenarios, the chairman's answer is the same: he does not care about the environment and only wants to increase the company's profits. Then, Knobe asks the participants if the chairman intentionally harmed/benefited the environment. With regard to traditional morality, it seems that the answer will be the same in both cases: the harm/benefit is not intentional, even if the harm is reckless. But the responses are different, as Neil Levy noted. The harm (first scenario) is judged to be intentional by 82% of the participants, and the benefit (second scenario) is assessed as non-intentional by 77% of the participants: "Surprisingly, altering the moral valence of the side effect dramatically alters subjects' perception of its intentionality: The majority of subjects now judged that helping the environment was not intentional" (Levy 2011: 7). Both answers are surprising, the first because the chairman's goal is not aiming at modifying the environment and the second because it is not consistent with the first answer.

However, there is an easy way out, emphasised by Levy: "If judgments of intentionality are sensitive to moral considerations, then it might be because people judge the intentionality of a side effect on the basis of its moral permissibility, rather than judging the permissibility of an action on the basis of the intentionality (or unintentionality) of the side effect" (Levy 2011: 7). Permissibility or responsibility comes first; intentions are only second.

If this interpretation is correct, then the traditional view of morality is in jeopardy: intentions are not so crucial. In the experiments examined in the previous section, intentions were central, at least for typically developing mature adults. Are people confused and inconsistent? Are they ambivalent? For many authors, we can explain these reactions by distinguishing two processes at work in our mind: Systems 1 and 2, the first being fast and intuitive, the second being slow and rational. A recent study suggests this reading: it was observed that participants who were good at the *Cognitive Reflection Task*, which measures a person's capacity to suppress spontaneous responses and to reflect on the task at hand, are less prone to the *Knobe effect*, "suggesting that the Knobe effect may arise from a System 1 process" (Ngo et al. 2015: 2).[12]

I will examine this interpretation in the next section. Here, I will offer two other comments on the Knobe effect, establishing that it is in fact not a challenge to the traditional conception of morality.

The first is semantic. As I have said, many expressions are used when we speak of an action as intentional. I have added that, depending on the context, some of them can be synonymous or not. In my opinion, the Knobe effect suffers from such ambiguities. In ordinary language, it is clear that responsibility for bad results is linked with intentionality. Think of this frequent reproach: "I am sure he did it on

[12] Ngo and colleagues have nevertheless not been able to confirm the results (Ngo et al. 2015: 5).

purpose"; a reproach that has no counterpart for good actions. However, it is not easy to apply this claim to the chairman, because his purpose is not specifically to harm the environment.[13] But think of: "I am sure he did it knowingly", or "consciously", or "willingly" or "deliberatively". The context authorises the application of such sentences to the chairman's behaviour. Moreover, the same context shows that "to have the intention to do *X*" and "to do *X* intentionally" are not always synonyms.[14] Our ordinary language is subtle, but sometimes too much so for philosophical precision, especially since intentions create complicated situations – in this respect, traditional morality is not a fully unified theory; but it is not a surprise.

The asymmetry between bad and good results – responsibility is usually only invoked in the case of the former – also suggests that "responsible" is very often used as a synonym for "blameworthy" or "culpable". As Vladimir Chituc and colleagues note: "Judgments in many domains are distorted by a motivation to blame" (2016: 22), even if, philosophically speaking, we are morally responsible for good results too. The Knobe effect is also linked with this semantic fact, which impacts other moral judgements too; it highlights a kind of imbalance in folk morality (Doris et al. 2007).

Another ambiguity consists of a frequent confusion between intentions and motives or reasons: Henry Sidgwick observes that "the distinction between 'motive' and 'intention' in ordinary language is not very precise" (Sidgwick 1981: 202). When we look for more accuracy, we notice that if both are sometimes identical (a goal can be a motive or a reason), it is not always the case. It is not surprising because they are conceptually different: motives and reasons are causes, preceding the actions, whereas intentions are mental acts embodied in actions and aiming at a goal. For instance, if I see someone drowning and I help him, my intention is to save his life, but my motive can be very different: it could be that I want to be at peace with my conscience or that I hope to be rewarded. In the stories imagined by Knobe, the reason the chairman has to begin the new program is precisely the goal of it: to increase the company's profits. The fact that his decision is motivated by a reason that appears to be morally suspect for many – his reason could be described as "increasing profits even at the expense of the environment" – and that this reason is also the goal aimed at could explain why the bad consequences were considered to have been intended (the good ones are not, because their value is at odds with the morally dubious motive).

Secondly, I observe that the Knobe effect impacts morality only at its periphery. It concerns side effects only, not primary effects: nobody ever doubted that the chairman's project to increase the company's profits was intentional, that he had the intention of increasing them and that he was making a genuine action. The problem focuses on the side effects – the Knobe effect is also named "the side-effect effect" – and, as another study has shown, it also focuses on the means: if the vice president suggests to the chairman that they shorten the worker's coffee break (a bad means)

[13] See nevertheless (Leslie et al. 2006: 425).

[14] Joshua Knobe (2004) acknowledges this. Some authors also emphasise that "intentionally" has several meanings; see, for example, Cova et al. (2012).

in order to increase productivity, the participants will likely consider the chairman to have acted intentionally in shortening the break – but not if he gives the workers a 1-hour nap break (a good means) for the same purpose (they will be in better health, and so will work more) (Cova and Naar 2012). Édouard Machery observes that "people take the costs that are incurred in order to reap some benefits to be intentionally incurred" (Machery 2008: 166).[15]

It is not difficult to see that both findings are linked to a psychological problem, known as the direction of intention. Take side effects first: when our action has several effects, it is often difficult to say which of those effects are willed and which are not. Classic debates about euthanasia (to kill someone in order to stop his suffering), abortion (to destroy a foetus in order to save the mother's life) and warfare (to bomb a strategic bridge where children are playing) are full of such difficulties[16]: can the agent confidently say that he does not aim at the bad effect, even if he knows for sure that his action will cause it? But if we are causally responsible for bad effects, and if we cannot claim that they are mere unintentional by-products, should we instead argue that we intended to bring them forth and so are fully morally responsible for them? Here, intentions and responsibilities are intermingled; it is not surprising that both notions contaminate each other and that the Knobe effect reflects this.

The case of means is a little different, because they are necessarily chosen. Consequently, we are not surprised "that means are generally judged more intentional than side effects" (Cova and Naar 2012: 837). But this judgement seems to be valid only for bad means and not for good ones. Why? Since the scenarios tested by Cova and Haar are inspired by the Knobe effect, the same mechanisms are probably at work. But here too, all the discussion takes place because we are faced with a genuine action, which is an intentional one.

In such complicated cases where effects and means are not easily evaluated, we seek simpler ways out, and one solution is to focus on anomalies and resort to characters: if we think that the person is greedy or benevolent, we tend to judge her acts accordingly. Grant Gillett notices: "In the normal course of events, human behaviour is, more or less, explicable on the basis of character or personal narrative" (Gillett 2008: 122). Peter Railton has imagined two scenarios where an olive tree owner sprays them against pests, knowing that his neighbour's goats will be harmed/benefited by the product, and he observes:

> "Intuition makes use of whatever evidence it can, and given our experience, someone who 'doesn't care at all' about whether he harms his neighbours is, happily, statistically rare, while someone who 'doesn't care at all' about whether he helps his neighbours is, perhaps regrettably, much more common. Statistical learning systems pay special attention to anomalies, since they carry more information than events that are more predictable" (Railton 2014: 853).

[15] However, Machery seems to understand "intentionally" as "deliberately", since he says: "Because [people] believe that costs are intentionally incurred, they judge that harming the environment is intentional" (Machery 2008: 177).

[16] They have been tackled by the *doctrine of double effect*. I will make some observations on this doctrine in the last section.

Such a disregard for the interests of others is rare and often revealing of a more general attitude: to act on bad intentions or at least on not good ones. Railton concludes: "What may matter in such intuitive social attributions of intent with respect to side effects is the fit of the action with the causal-attitudinal-intentional model of the agent we tacitly construct in light of his or her behaviour – for example, as 'anti-social' versus 'self-concerned' versus 'prosocial' – rather than the moral quality of the side effect itself" (Railton 2014: 854). A bad guy generally has bad intentions; consequently, the bad side effects of his actions will (probably) be intentional and considered by an observer to be so. Add in the asymmetry of good and bad, and you will arrive at the Knobe effect.

As we see, in the end, the Knobe effect does not constitute a challenge to the picture of traditional morality I have presented: it can take place within it, because it presupposes genuine actions, i.e., intentional actions. However, some authors link intentions with intuitions. Are intentions consequently tied to a System 1 process? And if this is the case, what impact does being tied to System 1 have?

8.6 A Dual Process Approach

Joshua Greene was the first to study moral dilemmas with the aid of neuroimaging in order to better understand moral judgement and moral decision-making (Greene et al. 2001). The dilemma that interested him most was *the trolley problem*, where subjects are asked if it is permissible to divert a threat (an out-of-control trolley) with the effect that only one person is killed instead of five. The responses are surprising, since the participants think it is permissible if the agent turns a switch, but forbidden if he has to push a fat man onto the rails to stop the trolley. However, in both cases one man dies and five are saved.

Greene also observes that different brain areas are mobilised in each case: rational ones (VMPFC) in the case of the switch, emotional ones (amygdala) in the fat man's version. In order to interpret these data, he turns to a dual process theory, inspired by Daniel Kahneman and Amos Tversky. When we make a moral judgement or a moral decision, we have two different resources: an intuitive, emotional and swift one (System 1) and a rational and slow one (System 2). Greene uses the metaphor of a camera to illustrate them: "The human brain is like a dual-mode camera with both *automatic settings* and a *manual mode*. A camera's automatic settings are optimised for typical photographic situations ('portrait', 'action', 'landscape'). […] A dual-mode camera also has a manual mode that allows the user to adjust all of the camera's settings by hand" (Greene 2013: 133). In other words, in usual situations, we resort to System 1, since it is efficient and quick, but in unusual ones, when the situation is not clear, System 2 is more appropriate: rational deliberation is better here than intuitions and emotions.

Sometimes, especially in intricate situations such as moral dilemmas, we are fooled. In the *trolley problem*, it is better if five people survive instead of one; however, if we have no problem with turning the switch (i.e., we judge that it is morally permissible), why are we reluctant to push the fat man? Because System 1 enters the

game and arouses emotions in order to prevent us from causing harm through personal physical force. Of course, such an emotion is usually very appropriate from a moral point of view: we should refrain from personally harming people, but in the fat man case, it seems to deliver a wrong answer.

In order to better understand what is at stake in System 1 judgments, Greene and colleagues have investigated some other versions of the *trolley problem*. They conclude that "harmful actions [are] judged to be less morally acceptable when the agent applied *personal* force to the victim. [… However] the personal force factor only affects moral judgments of intended harms, while the intention factor is enhanced in cases involving personal force. Put simply, something special happens when intention and personal force co-occur" (Greene et al. 2009: 369). This result is not surprising: harmful actions are morally problematic, especially when they are intended, since intention is at the centre of responsibility. However, it is not in this traditional sense that Greene sees the matter, since the intentional factor depends also on the personal force, and permissibility is lower when force increases, even if intention remains constant. System 1 appears then to have a certain complexity and a holistic function: it is not only an emotional immediate response (sometimes, Greene speaks of "alarm emotions") because the automatic setting

> "responds to harms that are *specifically intended*. Second, it responds more to harm caused *actively*, rather than passively. And, third, it responds more to harm caused directly by *personal force*, rather than more indirectly. It seems that these are not three separate criteria, employed in checklist fashion. Rather, they appear to be intertwined in the operation of our alarm gizmo, forming an organic whole" (Greene 2013: 246).

In Greene's dual process theory, intention counts, but only within System 1 processes, where it combines with personal force in an active pattern. System 2 focuses exclusively on outcomes: it is morally better if only one man dies. Moreover, from the point of view of ethics, System 1 and System 2 are not on the same footing, because when they conflict, Greene claims that we should rely on System 2, which is rational, and put System 1 on hold. Such a move has a price and might not be psychologically possible in some circumstances or for some of us, since we also read in *Moral Tribes*:

> "If you don't feel that it's wrong to push the man off the footbridge, there's something wrong with you. I, too, feel that it's wrong, and I doubt that I could actually bring myself to push, and I'm glad that I'm like this. What's more, in the real world, not pushing would almost certainly be the right decision. But if someone with the best of intentions were to muster the will to push the man off the footbridge, knowing for sure that it would save five lives, and knowing for sure that there was no better alternative, I would approve of this action, although I might be suspicious of the person who chose to perform it" (Greene 2013: 251).

If Greene were a supporter of virtue ethics, such ambivalence could be a genuine difficulty, but he is not, and for him, despite all psychological difficulties, pushing the fat man is the correct and ethical decision. Morality as it is and morality as it should be do not always coincide, as is well-known.

Consequently, for Greene, the traditional view of morality and responsibility is mistaken, because it is one-sided: the intentional component of an action has and should only have a secondary weight in morality. In usual situations, it is only one part in a shortcut allowing us to determine what is right and where responsibility lies; and in unusual or intricate ones, it is irrelevant, since we have to focus on outcomes, as utilitarianism and consequentialism are urging us to do (Greene 2008).

8.7 The Death of Traditional Morality?

If Greene's conclusions about the role and place of intentions are correct, this will result in an upheaval of our traditional view of morality. Even Mill would be thrown overboard because, as we saw, he also placed intentions at the heart of ethics. Morality as traditionally understood would come to an end and be replaced by something else, a new morality conceived perhaps as a kind of social management of misfortunes or a social control for bad behaviours – a move probably not unwelcome for Bentham and some other utilitarians. Nevertheless, at first glance, this seems to be absurd: John, Andrew, Paul and Peter, the protagonists of my four stories, ought then to be judged alike, as murderers, because they have caused the death of a human being. But what appears to be absurd for our traditional view of morality may well not be for rational morality even if we are psychologically unable to switch completely to this new stance – at least, it would take time. Greene has already argued for such a move concerning punishment in a paper written with Jonathan Cohen (Greene and Cohen 2004): retributivism should be replaced by consequentialism. Instead of punishing a criminal in a spirit of revenge or of debt paying, we ought to aim at social reintegration and rehabilitation.[17] Notice that, for Greene and some other authors, there is a link between intentions and retribution, and they contend that "our criminal justice system should change radically in the light of new neuroscience as it is imprudently concerned with an agent's intention" (Gkotsi and Gasser 2016: 63).

The question should nevertheless be frankly faced: is our traditional view mistaken when it makes intentions central to morality? Greene has given his reasons; can we find other ones?

The fact that children perceive intentions very early on can be an argument for linking them to System 1, which is a kind of "primitive" system. An experiment conducted by Jean Decety and Stephanie Cacioppo also shows that the perception of intention comes very early – as a kind of intuition – when we watch others' behaviour and that it is linked with emotional arousal: "We demonstrate for the first time how intention understanding […] and then affective processing occurs in very early stages of moral cognition processing […]. These results support the view that intentionality judgments both precede and guide moral cognition" (Decety and Cacioppo 2012: 3071). However, these data are inconclusive, because they also confirm the traditional view, for which intentions come first in order to guide moral assessment.

[17] Their argument presupposes hard determinism and is presented in the context of the free will debate.

Other studies cast doubts on Greene's interpretation. Michael Koenigs and colleagues have shown that sociopaths tend to push the fat man more frequently than typical subjects (Koenigs et al. 2007), and we already know that they assess intentions differently too. However, it seems difficult to say that they practise a better morality than we do. Greene nevertheless bites the bullet. After having reaffirmed his view under the name *The No Cognitive Miracles Principle* (NCMP), stating that "when we are dealing with unfamiliar moral problems, we ought to rely less on automatic settings (automatic emotional responses) and more on manual mode (conscious, controlled reasoning), lest we bank on cognitive miracles", he adds: "A corollary of the NCMP is that we should expect certain pathological individuals – VMPFC patients? Psychopaths? Alexithymics? – to make better decisions than healthy people in some cases" (Greene 2014: 715).[18] It remains nevertheless to be investigated if these better decisions are due to more rationality or less emotivity. As Jean Decety and Jason Cowel observe: "Are individuals who make utilitarian judgments in personal situations more rational and calculating, or are they simply colder and less averse to harming others?" (Decety and Cowel 2015: 10).

Let us return to intentions. For Greene, it is probably a mistake to say that psychopaths assess intentions differently: rather, they discard them for the benefit of outcomes. They put System 1 aside or at least rely less on it. Paradoxically, since they suffer from severe conditions, the judgements of these patients do not show the importance of intentions but their unimportance.

However, I think that all this reasoning is grounded in a misunderstanding: intentionality is not absent from System 2 but is as ubiquitous and as important there as in System 1. When someone judges that an agent should hit the switch or push the fat man, he presupposes that this agent has the intention of saving the life of the five persons. He does not act by accident, but deliberately, and this element is taken into account to determine the agent's responsibility. Consequently, Greene's argument cannot be generalised but is only valid (if it is) in a specific situation, when harms are caused as side effects – it was already the case with the Knobe effect.

More precisely, what is illustrated by the *trolley problem* is a situation where two effects are caused, a good one (to save five lives) and a bad one (the death of one person). Greene contends that in such situations, only the outcome should count and not the fact that the bad effect (the death of the victim) is willed as a means or permitted as a side effect of the good one. In contrast, traditional morality tends to suggest that in such situations we intend means, but not side effects, or at least not in the same manner: such effects are intentional or deliberately accepted, but not intended in the sense that they were the object of an intention and aimed at. Consequently, it is morally wrong to push the fat man but right to turn the switch. Greene objects to this thesis – and in my mind, he is partly right, as I will argue[19] – but this has nothing to do with the place and importance of intention in morality generally.

[18] The connection between utilitarianism and psychopathy is nevertheless weak and probably misguided; see Jaquet (2015).

[19] See also Baertschi (2013: chap. 1–2). Contrary to Greene, I conclude that we should refine our conception of intention and of its direction when several effects, good and bad ones, are present and not throw intentions overboard.

8.8 The *Doctrine of Double Effect* and Traditional Morality

Some classical moralists have tried to systematise such complex situations with the help of the *doctrine of double effect*. This doctrine consists of several rules saying that we should intend the good effect only, that the bad effect should not be the cause of the good one and that we have to take proportionality into account (e.g., a small bad effect is wiped out in relation to a larger good effect). This doctrine is not a general principle of morality and so should only be used in intricate situations where it is not easy to see the right option because a bad effect that cannot be avoided is involved (Goffi 2004: 238).[20] In brief, it has been conceived as a rational tool when our intuitions are confused or even muted, whereas for Greene it is "just an (imperfect) organising summary of the intuitive judgments. […]. It's our moral intuitions that justify the principle" (Greene 2013: 223). We can indeed object to the summary and be suspicious of a doctrine that fits with our intuitive judgements; but once more, it does not cast doubt on the intentional character of actions and the importance of intention for responsibility – Greene himself acknowledges that these questions are raised in the frame of an "action plan" (Greene 2013: 247), and what is planned is willed.

One of Greene's moral concerns is System 1's emotional blindness to distant and impersonal harms. He states: "When we harm people (including future people) by harming the environment, it's almost always as a side effect, often passive, and never through the direct application of personal force to another person. If harming the environment felt like pushing someone off a footbridge, our planet might be in much better shape" (Greene 2013: 253). Here, System 2 is necessary to avoid catastrophic effects, but is difficult to mobilise. I completely agree with him on this point: reflection is necessary in such situations; however, once again, it has no relevance for intentionality. It has to do with the relationships between agency, causality and responsibility: Am I responsible for the acts I intentionally perform and for voluntary omissions? Yes, of course, but I am not responsible for those I do or allow to happen accidentally or inadvertently. Am I responsible for *all the effects* of my deliberate acts and omissions? No, of course not, because these effects can be unforeseeable. And if they are foreseen, but not willed? Here the difficulties begin, which the *doctrine of double effect* tries to clarify. If intentions are at the centre, then responsibility cannot extend in the same manner to all the effects: the fact that they have been willed or not counts, and here, the doctrine has merits.

Utilitarianism provides another answer: we have to balance foreseeable good and bad effects and choose the best option. Accordingly, it objects to the *doctrine of double effect*, suggesting that it wrongly dismisses our responsibility for bad effects when they are not intended. Sidgwick made this proposal, linking responsibility to intentions: "For purposes of exact moral or jural discussion, it is best to include under the term "intention" all the consequences of an act that are foreseen as certain or probable (Sidgwick 1981: 202)". Utilitarianism keeps up Mill's lesson on the

[20] Notice that in the case of the Knobe effect, the *doctrine of double effect* considers that the chairman's project is not permissible: harming the environment is too high a cost.

moral importance of intentions, and it is necessary if it wants to remain a *moral* theory: our intentions always matter. However, morality is not alone in the struggle toward happiness, since, for many bad effects and undesirable states of affairs, we can instead turn to risk management, social security and political action.

Conclusion

Neuroethics, with the aid of neuropsychology, questions several traditional moral views. One of them is the central place given to intentions: for traditional morality, they characterise what actions are and ground the attribution of responsibility. Several empirical studies confirm this view, but others seem to refute it: the Knobe effect suggests that the intentional character of our actions is not central but derived from allocations of responsibility. For the dual process theory (at least Greene's version), intentions are important only when we rely on moral intuitions. I have argued that these two charges are not conclusive, mainly because they are at most valid only for actions' side effects. Consequently, traditional morality can stand firmly on its foundations, at least with regard to the place and role of intentions.

Traditional morality abides, of course, through many debates, and there exists pressure to modify parts of it. One place of disagreement concerns responsibility when an action has several effects. Here, utilitarianism and the *doctrine of double effect* are in tension, showing that this part of traditional morality is not on a firm ground. It is to their credit that Knobe and Greene's arguments have brought this difficulty to the forefront of neuroethics and to discussions about moral reasoning.

Acknowledgements I wish to thank Florian Cova, Ugo Gilbert Tremblay, Jean-Yves Goffi, François Jaquet, Pierre Le Coz and Yves Page for their helpful comments on a first version of this paper.

References

Anscombe E (1963) Intention. Blackwell, Oxford

Baertschi B (2013) L'éthique à l'écoute des neurosciences. Les Belles Lettres, Paris

Cerri M (2016) The ethical ghost in the brain: testing the relationship between consciousness and responsibility in the special case of REM sleep behavior disorder. In: Lavazza A (ed) Frontiers in neuroethics. Cambridge Scholar Publishing, Cambridge, pp 117–133

Chituc V, Henne P, Sinnott-Armstrong W, De Brigard F (2016) Blame, not ability, impacts moral "Ought" judgments for impossible actions: toward an empirical refutation of "Ought" implies "Can". Cognition 150:20–25

Christensen J-F, Gomila A (2012) Moral Dilemmas in cognitive neuroscience of moral decision-making: a principled review. Neurosci Behav Rev 36:1249–1264

Greene J, Cohen J (2004) For the law, neuroscience changes nothing and everything. Philos Trans R Soc Lond B 359:1775–1785

Cova F, Naar H (2012) Side-effect without side effects: the pervasive impact of moral considerations on judgments of intentionality. Philos Psychol 25(6):837–854

Cova F, Dupoux E, Jacob P (2012) On doing things intentionally. Mind Lang 24(4):378–409

Davidson D (2002) Essays on actions and events. Clarendon Press, Oxford

Decety J, Cacioppo S (2012) The speed of morality: a high density electrical neuroimaging study. J Neurophysiol 108:3068–3072
Decety J, Cowel J (2015) Empathy, justice, and moral behavior. AJOB Neurosci 6(3):3–14
Dennett D (2013) Intuition pumps and other tools for thinking. Penguin Books
Doris J, Knobe J, Woolfolk R (2007) Variantism about responsibility. Philos Perspect 21:183–214
Forest D (2014) Neuroscepticisme. Ithaque, Montreuil
Fried I, Mukamel R, Kreiman G (2011) Internally generated preactivation of single neurons in human medial frontal cortex predicts volition. Neuron 69:548–562
Frith C (2010) Consciousness, will and responsibility: what studying the brain tells us about the mind. ICR Monograph Series 58. www.thebrainandthemind.co.uk/. Accessed 11 June 2016
Frith C (2014) Action, agency and responsibility. Neuropsychologia 55:137–142
Gillett G (2008) Subjectivity and being somebody. Human identity and neuroethics. Imprint Academic, Exeter
Gkotsi G, Gasser J (2016) Neuroscience in forensic psychiatry: from responsibility to dangerousness. Int J Law Psychiatry 46:58–67
Goffi J-Y (2004) Le principe des actions à double effet. In: Goffi J-Y (ed) Hare et la philosophie morale. Vrin, Paris, pp 231–248
Greene J (2008) The secret joke of Kant's soul. In: Sinnott-Armstrong W (ed) Moral psychology. MIT Press, Cambridge, pp 35–79
Greene J (2013) Moral tribes. Atlantic Books, London
Greene J (2014) Beyond point-and-shoot morality: why cognitive (neuro)science matters for ethics. Ethics 124(4):695–726
Greene J, Sommerville B, Nystrom L, Darley J, Cohen J (2001) An fMRI investigation of emotional engagement in moral judgment. Science 293:2105–2108
Greene J, Cushman F, Stewart L, Lowenberg K, Nystrom L, Cohen J (2009) Pushing moral buttons: the interaction between personal force and intention in moral judgment. Cognition 111:364–371
Güney S, Newell B (2013) Fairness overrides reputation: the importance of fairness considerations in altruistic cooperation. Front Hum Neurosci 7(252):1–123
Güroglu B, van den Bos W, Rombouts S, Crone E (2010) Unfair? It depends: neural correlates of fairness in social context. Soc Cogn Affect Neurosci 5(4):414–423
Jaquet F (2015) Les conséquentialistes ne sont pas (des psychopathes), Implications philosophiques. http://www.implications-philosophiques.org/actualite/une/les-consequentialistes-ne-sont-pas-des-psychopathes/. Accessed 11 June 2016
Knobe J (2003) Intentional action in folk psychology: an experimental investigation. Philos Psychol 16:203–231
Knobe J (2004) Intention, intentional action and moral considerations. Analysis 64:181–187
Koenigs M, Young L, Adolphs R, Tranel D, Cushman F, Hauser M et al (2007) Damage to the prefrontal cortex increases utilitarian moral judgments. Nature 446:908–911
Leslie A, Knobe J, Cohen A (2006) Acting intentionally and the side-effect effect. Psychol Sci 17(5):421–427
Levy N (2011) Neuroethics: a new way of doing ethics. AJOB Neurosci 2(2):3–9
Machery É (2008) The folk concept of intentional action: philosophical and experimental issues. Mind Lang 23(2):165–189
Maoz U, Yaffe G (2015) What does recent neuroscience tell us about criminal responsibility? J Law Biosci 3(1):120–139
Marcel A (2003) The sense of agency: awareness and ownership of action. In: Roessler J, Eilan N (eds) Agency and self-awareness: issues in philosophy and psychology. OUP, Oxford, pp 48–93
Marek J (2013) Alexius Meinong. Stanford encyclopedia of philosophy. http://plato.stanford.edu/entries/meinong/. Accessed 29 Aug 2016
McCann H (1991) Settled objectives and rational constraints. Am Philos Q 28(1):25–36
Mill JS (1991) Utilitarianism. Oxford World's Classics, Oxford

Moran J, Young L, Saxe R, Lee SM, O'Young D, Mavros P et al (2011) Impaired theory of mind for moral judgment in high-functioning autism. Proc Natl Acad Sci U S A 108(7):2688–2692

Ngo L, Kelly M, Coutlee C, Carter M, Sinnott-Armstrong W, Huettel S (2015) Two distinct moral mechanisms for ascribing and denying intentionality. Sci Rep 5(17390)

Railton P (2014) The affective dog and its rational tale: intuition and attunement. Ethics 124:813–859

Rigato J, Murakami M, Mainen Z (2014) Spontaneous decisions and free will: empirical results and philosophical considerations. Cold Spring Harb Symp Quant Biol LXXIX:177–184

Sidgwick H (1981) The methods of ethics. Hackett PC, Cambridge

Treadway M, Buckholtz J, Martin J, Jan K, Asplund C, Ginther M et al (2014) Corticolimbic gating of emotion-driven punishment. Nat Neurosci 17(9):1270–1277

Yoshie M, Haggard P (2013) Negative emotional outcomes attenuate sense of agency over voluntary actions. Curr Biol 23:2028–2032

Young L, Bechara A, Tranel D, Damasio H, Hauser M, Damasio A (2010) Damage to ventromedial prefrontal cortex impairs judgment of harmful intent. Neuron 65:845–851

Zangrossi A, Agosta S, Cervesato G, Tessarotto F, Sartori G (2015) "I Didn't Want to Do It!" the detection of past intentions. Front Hum Neurosci 9(608). eCollection 2015