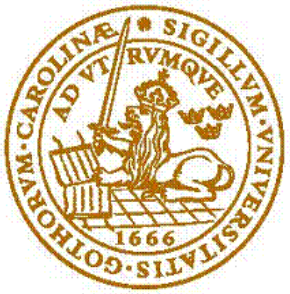

LUND
UNIVERSITY

Implicit and Explicit Moral Reasoning

*

Potential implications to ethical theory given a "Dual-process theory" of mind.

Bachelor's exam paper, Practical Philosophy
FPRK01:2, Lund University, VT 2012
The Philosophical Institution

Author: Anders Persson (fpr10ap2@student.lu.se)
Tutor: Dan Egonsson

Abstract

In this paper I examine potential implications of how a "Dual-process theory" of the mind can influence ethical theory. It is a theory dividing the brain in two system, an (1) *implicit system*, as intuitive, automatic and unconscious, and an (2) *explicit system*, as a deductive and conscious process. Empirical studies seem to indicate that this is how we can understand *moral reasoning*. It has been proposed that much of our reasoning is executed by our implicit system, and it is argued that the nature of deontology "at it's core" is implicit and intuitive. More specifically, it is claimed that a Naturalistic Fallacy is committed, that one treads over the line distinguishing 'ought' from 'is'. Ultimately, I will claim the criticism of deontology is at least partly held and that what can be deemed faulty is when one draws explicit knowledge out of implicit. I suggest that the implicit knowledge is approximations, and the fault is committed when deriving definite and exact explicit rules from these. As an ethical theory that takes a "Dual-process theory" of mind into account, appears to be "Two-level utilitarianism". I will also propose an added emphasis on virtue ethics, and try to show how implicit principles can be taught with explicit control. Finally, an increasingly important question seems to be to find an answer to the threat of having an "inhuman" moral theory. I will agree with arguments that a theory must have the possibility to be completely counter-intuitive, but that the human limitation also must lead to that at least some "moral wrongs" cannot be blameworthy.

Index

Introduction

What is moral reasoning? What do we do when we morally deliberate about a problem? How do we conclude a judgment deeming something good or bad? What kind of processes, mechanism and functions does our brain use? How do we process information about moral problems and dilemmas, to derive a judgment of a moral nature?

It is largely still a big mystery how Mind's processes function. However, during the past decade a flurry of neurological data has arisen, and we seem to just get a closer and closer look behind the curtain about what actually is going on in there. The contemporary narrowed down field of "Neuroethics" have focused on the question of what the "the moral brain" is like, and several models have been proposed. In this paper, I do not have the scope to try and represent the whole field and all the models proposed, rather I will concentrate on one of them, and my purpose will be three-fold.

First, I will try to explain how a proposed "Dual-process theory" of mind is argued to function, for moral as well as regular decision making. I will give background on how the theory have developed, present neurological studies of support, give concrete examples of dilemmas and review specific terms as "cognitive control" and "conflict detection". I will also present a useful analogy on an Elephant and a Rider (Haidt, 2006), which may help in yielding new predictions of moral behavior (chapters 1-3).

Secondly, I will review the criticism directed towards deontology and intuitions in general, which is seen like a kind of confabulation, or "honest lies" (Greene, 2007). Ultimately, I will argue that the criticism which is mainly aimed at Kant may at least partly be warranted. And I will suggest a perhaps more precise explanation what "the fault" may consist of, that it is a *definite rule* concluded from an *approximate assumption* (chapters 4-8).

And thirdly, I review what kind of ethical theory a "Dual-process theory" of mind may lead us towards. I will suggest a support for a "Two-level utilitarianism" (Hare, 1981) as a general framework, with an added emphasis on virtues (chapters 9-11).

1. Reasoning Agents

Traditionally, human reasoning was postulated to be of a logic and analytic nature. The model of Edwards (1954) explained human problem solving and decision making by using the principle of *"Expected Utility*". In other words, to account for probabilities of maximized utility, to always do what is most likely the best outcome. The model above was later extended to deductive reasoning, where the agent analyzes, evaluates as well as constructs premises to deductive arguments.

However, occasionally we as humans seem to fail at making the rational, logical and probabilistic calculations that this model describes. Consider the example of the "*9/11 Traveler*": one who is scared of flying on the 11th day and the 9th month of the year. Even if the chance of a new terrorist attack is minimal and the risk of fatal traffic accidents is much higher, they may decide that taking the car must be safer (Gigerenzer, 2004).

Another example is what I will call the "*Pantyhose Choice*" experiment: several pairs of pantyhose were displayed in a row for female subjects to choose from. They gave sensible answers to their choices, as the preference of a specific knit, elasticity etc. But the items were identical, and the result from the experiment showed that what people actually preferred were the items on the right-side of the display (Nisbett & Wilson, 1977).

Studies also show that we humans seem to not have a "natural sense of numbers", just as animals. Rats and pigeons can only consistently discriminate between 1-4 object, but not amounts greater than that (Shettleworth, 2010, pp.344). Experiments show that infants in a similar way have a hard time discriminating between objects greater than four (Shettleworth, 2010, pp.352). An Amazonian tribe has also been found which do not have numbers at all for amounts greater than four, the rest is a fuzzy "many" or "more" (Pica et al. 2004). This seems to indicate that even grown up humans does not have a "natural sense of numbers".

Empirical evidence like these can be found in great mass, and seem to falsify and refute the model of deductive human reasoning as not "descriptively adequate" (Kahneman & Tversky 1982, Wason & Shapiro 1971). A contemporary solution is then to adopt a "Dual System Theory".

2. Dual Systems of Reasoning

A Dual system theory of human reasoning can have many specifications, perhaps depending on what kind of problems that are to be explained. But there is usually a general distinction between an (1) Implicit system and an (2) Explicit system.

The *Implicit* system is best describes as fast, effortless, automatic and unconscious, which is "highly contextualized, personalized and socialized" (Stanovich & West, 2000, pp.658-659). In other words, it accounts for the automatic contextualization of problem solving. As for example the "9/11 Traveler"; in that particular context, on that special day, it would rather travel by car instead of by airplane. It is also related to implicit ("procedural") memory, which is gradually learned by fine tuning habits (Sun, 2001), as for example the skill of driving a car. It is an ability filled with unconscious, automatic, yet coordinated movements, such as shifting gear, steering and looking in the side-mirrors. Kahneman (2003) also describes this system as "intuition", because of a strong emotional connection with the process.

The *Explicit* system can be described as slow, effortful, active and conscious, which serve more controlled processes that can "decontextualize and depersonalize problems" (Stanovich & West 2000, pp.658-659). In other words, the traditional deductive model can be interpreted to be retained in this second system (Sahlin et al. 2010, pp.135). It accounts for the conscious, but often slow and effortful, deductive and logical reasoning. Explicit memory is on the contrary to implicit, based on the idea of "one-shot explicit rule learning" (Sun, 2001), such as facts. It is knowledge held and retrieved to the working memory of the brain, where it generally is proposed that active, cognitive information processing takes place.

It might be difficult to draw a distinct line between these systems, due to the fact that they constantly interact. For example a "*Stroop task*" (Gazzaniga et al. 2008, pp.116): a set of words of colors with a mismatching color of ink is presented to the subject, whom is given the task of saying the color of the ink. For example the word "RED" in blue ink, or "GREEN" in purple ink. Most subjects will have the impulse of reading the word, but this is not the given task. Hopefully, a *conflict detection* will happen, a process that has been associated with a specific region of the brain, the anterior cingulate cortex (ACC). In that case a *cognitive conflict* will occur, between (1) the automatic impulse to read the word and (2) performing the "explicit" task.

Another test is the "Cognitive Reflection Test" (Fredrick 2005), which contains small but intuitively problematic questions, often of a mathematical nature. For example *the Bat and the Ball*: "A bat and a ball cost $1.10 in total. The bat cost one dollar more than the ball. How much does the ball cost?" Fredrick (2005) notes that most people answer "10 cents", which seems like an intuitively neat solution, and many sticks to this answer even after long deliberation. To arrive at the correct answer, one has to for example put it into some mental equation (X is the ball, 1.10 = X + [X+1.00] --> 1.10-1.00 = 2X --> X = 0.05). In other words, this would mean one had to use explicit resources as the working memory, for the active and conscious information processing.

To get the right answer in the "Cognitive Reflection Test" above, as well as to succeed in the "Stroop task", two things need to happen. First, the cognitive conflict has to be detected, and secondly, the explicit resources have to be available. What these kinds of tasks are supposed to test then, is the subject's ability for *cognitive control.*

To illustrate how the two systems of a Dual-process theory interact, Jonathan Haidt (2006) proposes a useful analogy:

The Elephant and the Rider: "I'm holding the reins in my hands, and by pulling one way or the other I can tell the elephant to turn, to stop, or to go. I can direct things, but only when the elephant doesn't have desires of his own. When the elephant really wants to do something, I'm no match for him. [...] The controlled system [can be] seen as an advisor. It's a rider placed on the elephant's back to help the elephant make better choices. The rider can see farther into the future, and the rider can learn valuable information by talking to other riders or by reading maps, but the rider cannot order the elephant around against its will. [...] The elephant, in contrast, is everything else. The elephant includes gut feelings, visceral reactions, emotions, and intuitions that comprise much of the automatic system. The elephant and the rider each have their own intelligence, and when they work together well they enable the unique brilliance of human beings." (Haidt, 2006)

The understanding Haidt wants to present is how the unconscious, automatic intuitions of the *Implicit* system (the elephant), is related to the controlling, conscious, deliberating and deductively reasoning of the *Explicit* system (the rider). To me, the analogy effectively highlights the symbiosis of the two systems, not as separate systems, but as highly dependent on each other.

3. Moral Reasoning

Similar doubts on our deductive reasoning capability have been raised towards moral judgments. Haidt (2001) presented people with the following dilemma:

The Brother and the Sister: "Julie and Mark are brother and sister. They travel together in France on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide it would be interesting and fun if they tried making love. At the very least it would be a new experience for both of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy making love, but decide not to do it again. They keep that night as a special secret, which makes them feel even closer together. What do you think about that? Was it OK for them to make love?" (Haidt, 2001, pp. 814)

Individuals answering are most of the time instantly sure it was not OK. When asked why, they may point out the danger of inbreeding, but then be reminded of the double birth control. They may then object it may leave emotional scars, but the story seems to make it clear it does not. In the end, Haidt et al. (2000) summarizes that people are often still very sure that it is wrong, but cannot give an acceptable argument why it is so. Haidt (2001) proposes a "Social Intuitionist Model"; that judgments are quick moral intuitions and reasoning most of the times come afterwards, as *post hoc rationalizations*. In other words, this is contrary to what we might think, that our judgments are caused by (deductive) reasoning.

A contemporary proposition to account for this is a "Dual-process theory", also for moral reasoning. Greene et al. (2001 & 2004) examined participant's brains in fMRI (functional magnetic resonance imaging) as they give moral judgments to a battery of moral dilemmas. They distinguished between two types of dilemmas, which are captured in two originally proposed by Judith Thomson (1986).

Trolley: "A runaway trolley is headed for five people who will be killed if it proceeds on its present course. The only way to save them is to hit a switch that will turn the trolley onto an alternate set of tracks where it will kill one person instead of five. Should you turn the trolley in order to save five people at the expense of one? Most people say yes." (Greene et al. 2004, pp. 389).

Footbridge: "As before, a trolley threatens to kill five people. You are standing next to a large stranger on a footbridge spanning the tracks, in-between the oncoming trolley and the hapless five. This time, the only way to save them is to push this stranger off the bridge and onto the tracks below. He will die if you do this, but his body will stop the trolley from reaching the others. Should you save the five others by pushing this stranger to his death? Most people say no." (Greene et al. 2004, pp.389).

The difference Greene et al. (2001 & 2004) wants to distinguish between is that *impersonal* dilemmas, like Trolley, is damage to an individual merely deflected from saving the five. Whereas in the *personal* dilemma, like Footbridge, bodily harm is made by the moral agent himself, and is an "up close and personal" (Greene et al. 2001, pp.2106) type of violation.

The fMRI-results from these studies suggest two distinctively different brain structures to be active depending on the answers. People answering "Yes" to the Trolley dilemma activate structures associated with working memory and cognitive control (dorsal lateral prefrontal cortex), in other words with the *explicit* system. People answering "No" to the Footbridge however, activate structures associated with emotion (ventro medial prefrontal cortex and amygdala), in other words the *implicit* system.

Even more interesting are the people answering "Yes" to the Footbridge dilemma, to push the man from the bridge. They activate the same emotional structures, but also what is associated with *conflict detection* (anterior cingulate cortex), as with the Stroop task (see previous chapter).

Further support for the distinction above can be found in four different categories of people, which all are more likely to answer "Yes" to Footbridge; (1) patients with damage to the emotional structures of the brain (Koenigs et al. 2007), (2) people with good scores on Cognitive Reflection Tests (Hardman, 2008), (3) people with an expressed high "need for cognition" and low "faith in intuition" (Bartels, 2008), (4) individuals with an unusual high working memory capacity (Moore et al. 2008). These examples seem to assume, for different reason, why the explicit system would probably be stronger in this these individuals, and correlate to "more positive answers" on Footbridge type dilemmas.

Greene et al. (2008) also performed experiments where participants had to do a cognitively demanding task, at the same time as answering to Footbridge-type dilemmas. The task would "steal" cognitive resources, from working memory, so if used on the moral dilemma, it should be harder or talk longer time to answer. The results indicated that *negative answers*, the intuitive

option, was not affected, but *positive answers*, took longer time and was expressed harder. Other studies also showed that if pressed by time, more people will give negative answers, as if the "cognitive decision" was not finished.

All in all, the data presented above seems like support for a Dual-process theory of moral reasoning. On the one hand, there would be the conscious, controlled and active processing of information in the Explicit system, using the limited resources of the working memory. And on the other hand, we have the automatic, unconscious and intuitive process of the Implicit system, often resulting in post hoc rationalizations.

4. Psychological Natural Kinds

Greene (2008) suggests that the empirical data he presents explain why we have developed the two moral philosophies of (1) "*deontology*" and (2) "*consequentialism*" . The debate on which is right and which is wrong, may perhaps be explained with *psychological natural kinds* (Greene, 2008, pp.37). In other words, Greene proposes the distinction of the implicit and the explicit system, which he calls the "intuitive" and the "cognitive" system.

Deontology can as Greene do, be defined "by its emphasies on moral rules, most articulated in terms of *rights* and *duties*" (Greene, 2008, pp.37). In the Footbridge dilemma the prototypical deontologist would answer "No". They could for example refer to the principle that no one is to be used as "a mean" to help somebody else (Kant, 1785/1959).

Consequentialism however, has the emphasis on consequences alone, which all are summed up, aggregated in a whole, as one single judgment. Greene summarizes: "For real-life consequentialism, everything is a complex guessing game, and all judgments are revisable in light of additional details." (Greene, 2008, pp.64). The prototypical consequentialist in the Footbridge dilemma would answer "Yes", that one dying is a better consequence than five dying.

As showed in the previous chapter, these answers correlate to the activation of the two distinct systems of reasoning. The implicit system resulted in negative answers, and the explicit activation in positive answers. Thus, Greene (2008) concludes, (1) deontology is, at least, mostly "intuitive" (implicit), and (2) consequentialism is, at least, mostly "cognitive" (explicit).

This would suggest that arguments and reasons from a deontologist often may be *post hoc rationalizations* (Haidt, 2001). This is a phenomenon that also could be called *confabulation*; verbal statements describing an inaccurate historical background (Dalla Barba, 1993). In other words, it is a kind of "honest lying" (Moscovitch, 1995), not saying the truth, but without the intention to lie. Perhaps the most extreme case of confabulation is "*split-brain patients*", people with separated hemispheres of the brain. Experiments can be set up so an instruction is received only by the right hemisphere. For example: "Stand up and walk over to the table." When the left hemisphere is asked why he is standing by the table, they often make up a reason, like: "I just wanted some water", and without any knowledge of the instruction. The reason for this result is that even though language-comprehension to some extent exists in both hemispheres, usually only the left hemisphere (for right handed people) is able to produce it. Gazzaniga & LeDoux (1978) argues that all humans to some extent confabulate in this fashion.

Greene (2008, pp.63) concludes that deontology seems like "a kind of moral confabulation". Greene points out that Kant might not even disagree, and presents a quote: "Two things fill the mind with ever new and increasing wonder and awe, the often erand more steadily we reflect on them: the starry heavens above me and the moral law within me." (Kant, 1785/1983).

Perhaps then the fault of Kant could be claimed to be a kind of *Naturalistic Fallacy* (Greene, 2008, pp.72). It is a term originally defined by Moore (1903), but is also related to Hume's (1939/1978) distinction of the factual *IS* and the normative *OUGHT*. A famous example is Social Darwinism (Leonard, 2009). It sought to apply the scientific conclusions of evolution to social norms, and to justify "survival of the fittest" as a principle in society. Also Kant identified the "*IS-OUGHT"* distinction. He concluded that we do have to use our experience to decide moral problems, for example the feeling of sympathy, but that experience alone could not decide an act to be moral or immoral. The attempt from Kant was to formulate a "Categorical Imperative", as an absolute, rational and unconditional requirement; to only act in accord to the maxim you could wish to be a universal law (Kant, 1785/1983). Maybe this then, can be seen as a post hoc rationalization. As Greene argues, anyone who claims to have such a theory, seems to have "crossed the infamous 'is' 'ought' divide" (ibid., 2008, pp.72).

It might seem a bit unclear exactly what kind of fault Kant and deontology is accused of. Deontology is considered confabulating, intuitive and to use the emotional, implicit system. At the same time, Kant explicitly denies the feeling of, for example, sympathy as solely relevant in

moral deliberation. By the looks of it, Kant seems to have shared many of the conclusions of the Naturalistic Fallacy.

Nonetheless, it is criticism with potentially severe consequences and has given rise to several attempts to refute the conclusions. I will in the next part of this paper present some of the criticisms, and I will distinguish between what I see as two possibilities to deny the conclusions. The first is to deny the empirical investigation, and the second to justify the use of the implicit system. By reviewing these objections, I believe it could be made clearer what the "fault" actually consists of. Ultimately, I will argue that the criticism from Greene (2008) remains, and that the "fault of Kant" can be explained as having derived at an explicit principle from an implicit background.

5. Deny the Empirical Conclusions

Several of Greene's critics doubt his empirical results. They question if judgments resulting from the explicit system always, or even most of the time, in fact is in line with consequentialism.

Consider the infamous example of the "*Axe-murderer*"; whom is knocking on Kant's door looking for Kant's friend, hiding in the basement (Kant, 1785/1983). *Should* he lie about the whereabouts of his friend? Kant denied the intuitive answer of telling the murderer his friend was not here. To do this, it seems one would exhibit a great deal of cognitive control, in other words activation of the explicit system.

Another example is the patients (as mentioned in chapter 3) with brain damages in areas associated with emotion. Their conclusions should then primarily be derived by the explicit system and be consequentialistic. However, in "Ultimatum Games" (see footnote[1]) they tend to punish people more than normal subjects (Kahane et al. 2010, pp.573). In other words, they seem to adhere to a deontological principle of Kantian "Retributivism"; that a proportionate punishment is morally acceptable.

[1] "*Ultimatum Game*" experiments consists of two players, and a sum of money preliminary given to one of them. The first has to propose how to divide the money, and the second has to accept the offers if any of them are to get any money at all. In other words, the first can propose an unequal division as 70/30, and the other can punish the first by not accepting. Often experiments are done as a series of rounds, or as a single offer.

In other words, some deontological judgments seem to be using the explicit system. To test this Borg et al. (2006) distinguished between two types of deontological principles. The *Doctrine of Double-Effect* (DDE) stipulates that it is more difficult to justify *intended* harm, either as end or as means, then it is to justify harm that was not (McIntyre, 2004). To hit the switch in the Trolley dilemma would be harm *merely deflected*, without the intention to hit the other, and would not violate DDE. But on the Footbridge, harm is done as a direct mean to help some other. The *Doctrine of Doing and Allowing* (DDA) stipulates that *doing* harm is worse than merely *allowing* harm to be done (Howard-Snyder, 2002). It can also be weighted, so for example to allow two to die instead of killing one, might be acceptable, but *not* to allow 1.000 to die without killing one. It could then be acceptable, according to DDA, to refrain from acting in both Trolley and Footbridge.

Borg et al. (2006) found the same tendency as Greene, that the implicit system was activated in DDE judgments, and that the explicit system was activated during consequentialistic judgments. However, DDA judgments were also the result of activation of the explicit system.

To test it further, instead of structures in accordance to deontological and consequentialistic judgments, Kahane et al. (2011) hypothesized a more fundamental function. They set up a new battery of dilemmas, as they claimed that the previous battery used by Greene et al. (2001 & 2004) had few dilemmas where the answer was intuitively in accordance to consequentialism. The intuitive (common sense) answer in for example Footbridge is deontological, to not push the man. It would then only be natural that the few that were not giving the intuitive answer activated the other system. They had expert judges distinguishing between intuitively deontological or consequential dilemmas, where the latter for example could be of the *Axe-murderer* type.

Kahane et al. (2011, pp.9) concluded that the results was a partial support for a more fundamental distinction of moral judgments, as either *intuitive* or *counterintuitive.* Regardless if the dilemmas were intuitively deontological or consequentialistic, implicit structures was activated when agreeing to the intuitive option. However, counterintuitive judgments activated the explicit structures in both cases, including the area of the brain for *conflict detecting* as noted for the Stroop task (see chapter 2).

To me, this seems like a plausible conclusion, and I would like to propose an exemplified explicit deliberation of a Footbridge dilemma. I will call it the *Graphical Image:"*He sees the

man screaming in fright as he falls through the air, with arms flailing uncontrollably. The big man lands on the railway track five meters below, with a muffled thud. One leg gives away and ends up in an ugly angle, twisted, half way up to his torso. The man shrieks in a sudden high pitch, as the pain reaches his awareness. He freezes briefly as he sees the locomotive rushing towards him down the track. Horror is mixed in with his shrieks. He frantically claws in the dirt, trying to get hold of something, anything, to pull himself away. He holds up both his arms in a final futile attempt to protect himself as the locomotive runs him over. The only thing that is left as it has passed is a dark trail of blood, starting where the man of the Footbridge once lay. The locomotive, hitting the brakes, stops a bit further down the track, just in time before hitting the five railway workers."

For the arguments sake I will deem this a plausible scenario and explicit mental images one could see. It seems to me then plausible and rational to (explicitly) switch back to the deontological option, to refrain from pushing the man. One could also imagine the relatives of the man, maybe a family of four that would lose their beloved father and husband. What would be their reaction, and what would be the reaction of the relatives to the five railway workers? If refraining from saving them, the latter would probably only deem it as a tragic accident, and not blame you, the agent, for not saving them by killing the one on the bridge. But the relatives to the man would likely scream "Murder!", and you would risk going to jail. One could then sum up all the potential consequences, and arrive at the conclusion that one would still do best consequence by refraining from pushing the man.

So what kind of conclusions can be drawn from this? I would suggest that the activation of the explicit structures of the brain does not automatically result in an answer in accordance to consequentialism. And the implicit structures are not automatically, or necessarily most of the time, answers in accordance to deontology. At least partly, there seem to be more fundamental functions, of counter-intuitiveness and intuitiveness. To me, the conclusion and judgment of the *graphical image*, as well as the consideration to consequences from relatives and society, would be of a "consequentialistic nature"; aggregating and summing up the consequences into a whole.

Nonetheless, perhaps this does not have to deny the conclusions of Greene (2008), that deontology "*at it's core"* is affective and implicit. The "nature" of deontology could perhaps still be claimed to be of to *justify* certain emotional *intuitions,* as specific rules and principles. And

the "nature" of consequentialism could perhaps still be claimed to be about *aggregating conclusions*.

6. Implicit Justifications

Another direction of responses to Greene (2008) is to justify the judgments from the implicit system, as warranted conclusions. They interpret the same empirical data, as Greene, to be in favor of a deontological account. An extensive research program has been formed by Hauser (2006) and Mikhail (2007) among others, going by the name of *Universal Moral Grammar* (UMG). It is an analogy with the Chomskyan project in linguistics, which claim language is embedded in our biology. It is said to be a set of syntax and grammar in us, that all languages make use of. Among other things, it is claimed to be a required explanation in linguistics, to account for how fast infants and kids learn new words (up to 20 per day), which often instantly are ready to be used.

The hypothesis in UMG is that we have a similar innate set of "syntax", rules or principles, also behind our moral judgments. Hauser (2006) concludes that at least one of these underlying principles seems to be DDE, the doctrine of double effect. Both Kahane (2010) and Borg (2006) agree that DDE is a *principal explanation* for the empirical result of the Trolley and Footbridge. It is a principle found across the world, and is independent of culture, education and class.

I would like to propose what I deem another plausible proposition. Weber's Law is a principle from *signal detection theory* of just noticeable differences. But it also happens to be a basic principle found in most animals, including humans, and predicts the *analogue* activity in the neurons of the brain (Shettleworth, 2010, pp.65). I will here concentrate on one prediction, that the sensory neurons are activated proportionally more to bigger stimuli. In other words, the closer an object is, or bigger, the individual neuron will react more. It could be seen as a very basic control of an organism to be more likely to go for a "*better option"*. An organism that see two apples, one big and one smaller, would without any extra effort be more motivated to reach for the bigger. Evolutionary I would claim it would make sense, as this type of neuron and organism would benefit in fitness and potentially get more energy and do less cognitive work. Perhaps then there is a "Weber's Moral Law"; that if X is perceived as closer than Y, all other

things equal, X is deemed more important. Even if this is just a hypothetical theory, remember the *Pantyhose Choice* (chapter 1). The participants reached for the product to the right. Most people are right handed. So it would probably be the most accessible option, and in a sense the *closest*.

Another interesting account is held by Gommer (2011). Much like Greene (2008), he proposes that intuitions are *biological facts* that have evolved as a result from interaction between genes and environment. It can for example give an explanation on why killing sometimes is deemed wrong, and at other times not, and how it correlates to our western Law. Let us assume Richard Dawkins theory of "The Selfish Gene" (1976); what makes *evolutionary sense*, is to maximize the own gene's fitness and survival. With this principle in mind, it would then make sense for our emotions to tell us that killing one's own child (genes), or mother, would be about the worst thing one could do. And killing some foreign threat as a "wild beast", or "evil terrorist", would be much more acceptable, perhaps even praiseworthy. This seems to correlate to our intuitions, and also to our (western) law and justification of for example killing when in war.

But do these proposed inner principles we may or may not have and acts upon, have anything to do with moral? I would claim, as Greene (2008), that there is an ever persistent naturalistic fallacy, if so is claimed. We can understand these facts, that killing babies would *feel* more wrong according to an evolutionary account. But they remain, as perhaps much of our emotions from an evolutionary heritage, often not relevant to the moral problem at hand.

These inner principles could perhaps all be called a kind of "*implicit bias*"; partialities, and aptitudes towards specific sets of rules or principles. There is a whole set of cognitive biases, even some very elaborate, that exists in most of us. Experiments have been done where comparably equivalent CV:s is sent out to reviewers, where only the name differs. An American study showed that an American-European name on the CV were 50% more likely to get a callback interview, compared to an African-American name (Bertrant & Mullainathan, 2004). In a Swedish study Swedish names was three times as likely to be offered an interview, compared to Arab names (Rooth 2007). The same ("discriminating") effect has been observed with female versus male names, where the latter more often was deemed "more capable" (Saul, 2011). This should be worrisome empirical data for anyone that holds *fairness* and *equality* as principles that should be defended.

In other words, independent of all the "facts" presented in this chapter of how the world *is,* we may still want to claim that the world *ought* to be in another way. And it may be claimed, that the fact that something *is* in some way, doesn't say anything about how it *should* be. This could include the *discriminating* implicit bias as above, as well as the *doctrine of double effect*.

7. Explicit and Implicit Principles

In the two previous chapters I have reviewed two directions of objections to Greene's critic of deontology. In the end, it may perhaps be a bit unclear how the implicit and explicit system actually correlates to both deontology and consequentialism. However, I will argue that from the different examples and empirical data presented, we can extract some general principles of how we acquire *moral knowledge*. I will distinguish between five ways in the following passage.

(1) *Innate to implicit.* The first is the proposal of innate principles. For example we have the Doctrine of Double-Effect, Weber's Moral Law, or in general our emotions as tuned by the evolution of our species. In other words, they are inner *presumptions* of how to act on certain situations, which can either be said to directly be *implicit knowledge,* or is transformed into it.

(2) *Experience to implicit.* The second alternative is experience from senses passively forming implicit knowledge about the world. For example the implicit bias of discriminating. I would propose this impression comes from the expose to the society we live in. For example in Sweden we rarely see Arab names or appearances in "high-ranking" positions, more often it is the other way around. In other words, the implicit system passively learns through exposure what to expect, and *not necessarily* with a conscious involvement. So, when in front of a choice of two candidates for a high-ranking position, the automatic judgment from our implicit system would then be that the non-Arab is most likely best for the position. Even after explicit deliberation and review of all the available information, as the CV, we might still have the "gut-feeling" that the Swedish looking candidate would most likely be the best option.

(3) *Explicit to explicit.* A third way is to use the explicit resources to arrive at explicit rules that can be used by the explicit memory. Such as facts, they are definite, and true or false. The main resource is the working memory of the brain, which is severely limited, but it can be used to perform simultaneous comparison of a few objects. Here we can choose between options, or

weigh consequences to each other and sum them up into a single resulting judgment. In other words, it is deductive reasoning and aggregating consequences into a whole.

(4) *Explicit to implicit.* A fourth alternative would be to have an explicit, cognitive control when learning, as for example learning to drive. We may have explicit knowledge, as laws and explanations for signs learned from the driver's book, held active in our working memory while performing the driving lessons. It will give us the ability to monitor our implicit acting, if we do all those explicit rules we have read in the book. Perhaps an even more obvious explicit monitoring system, is the driver's teacher in the seat beside us. The teacher can explicitly inform us when we forget something or risk making a mistake. At first, these sets of rules and advices will be explicitly instructed and monitored. But after time and training they will become part of the implicit system and automatically executed and used while acting in the world.

(5) *Implicit to Explicit.* A fifth identified way would be to interpret an implicit principle into an explicit, in other words, into a definite principle such as a fact. This explicit rule could in turn be used to cognitively control and monitor the implicit system in other situations. For example it could be to "listen" to the emotional "gut-feeling" like the one for killing "foreign threats", or rather the lack there-off perhaps. One could define this as an explicit rule, to be acceptable, in defense of the own entity. This should be contrasted to (3) explicitly weigh consequences against each other, but is instead to draw a conclusion from the implicit and automatic judgment. This is what *post hoc rationalization* would be, the implicit justification, to find explicit rules that correlates to the implicit.

Perhaps (1) and (2) could be bundled up to the same kind of function; both seem to adopt unconscious (implicit) knowledge into new implicit knowledge. I will suggest further support for this conclusion in the next chapter, and I believe this can give us an understanding of what the fault of Kant and deontology could be. In the preceding chapters I will also discuss what the division of acquiring moral knowledge proposed here, tells us of an ethical theory that we *ought* to have.

8. The Approximate and the Exact System

In his book "Supersizing the Brain", Andy Clark (2008) draws a similar picture of a "Dual-process theory" as presented so far. He also suggests an added distinction of the cognitive process, of an (1) *approximate* and an (2) *exact* process. The first (implicit) system is suggested to handle the complex scenarios, where generalizations are needed, and the second (explicit) managing limited but exact calculations. In other words, it can explain how we on one hand *interpret* complex contexts in the world, and on the other hand perform exact *calculations* as math or "logical conclusions". In this chapter I will try to analyze a few examples from this paper, to try and show how this seems to fit, and then suggest what it tells us of the "*implicit to explicit"* knowledge.

Let us assume Dawkins "Selfish Gene" (1976) again, as an example of innate implicit principle. The preposition in the gene should be a setting that in *most cases* has been a successful setting in gaining energy and reproducing the specific gene. For example, the individual specimen with the most likely presumption to "take care of one's own young", will have genes more likely to survive in this aspect. In other words, the heritage from our ancestors is settings to certain problems and contexts that are not *definitely* successful, but just *statistically most likely* successful.

How about the "discrimination bias"? My claim is that we get a generalization of the most likely character to a "high-ranking" position. This is done through social exposure of the current social distribution of "Swedish" and "Arab"-appearance. Most Swedes (including those with foreign heritage) will then implicitly deem the most likely good candidate as "most likely Swedish looking". But of course this is a generalization, there are deviations. The implicit principle is in other words what could be represented as a *statistical approximation*, what is most likely the "best fit". Just as innate, it is *not* a representation of the *exact distribution*.

Let us again consider the dilemma Haidt (2001) presented to people, *the Brother and the Sister* (see chapter 3). They are described to have had sex, just for one night, and the participants answering often pointed to the danger of inbreeding. In general this could be a real threat, evolutionary it makes sense to have this innate principle, and it would likely have bad effects. What it doesn't account for is not only the single, but double birth-control. Participants could then point to emotional damage. Also this could be a good approximation, a sensible general

principle. In many cases it would *probably* mix up the relational role of sister and brother. If made public, it would lead to difficult social reactions, and if held secret it may strain on the conscious of the siblings. Only in this unique case, it did *not* lead to emotional harm.

In all these cases there seem to be nothing "wrong" in the approximations *per se*, given the background they can rather be seen as *sensible* and *probable*. But they are not judgments derived from the exact case or problem in front of them. The principle deriving the judgment is from *approximations,* and it does not say anything definite or exact about the individual case. To me, this seems to be able to explain what kind of fault is done in the dilemma like the Brother and the Sister. It can also explain what kind of fault is done when we implicitly discriminate certain groups of people. Even if the probability for X is 90%, in a given situation S, it still does not necessarily say anything about the individual S. To be sure if S, as for example a certain problem or individual, is X, we would have to examine it.

It would then also be a mistake to draw explicit principles, as facts, from implicit approximations. So another aspect of the *fault* of Kant, is to formulate an "unconditional", *definite rule* and law as the Categorical Imperative, from an "inner" approximation. Or, at least to draw the further conclusion as he do, that lying is "categorically wrong". The actual mistake is perhaps committed at first when trying to apply this rule on new situations, in belief that it is definite when it actually "only" is an approximation. For example a hypothetical case of an *Axe-murderer* knocking at the door looking for your dear friend hiding in the basement. In this case you should lie, but in almost all other cases it would, also from a consequentialistic account, be bad to lie. This is one of the reasons why the more flexible explicit calculating ability is important, to solve these new, perhaps unique situations.

As Singer (2005, pp.349) points out, there have been many attempts to formulate a "first principle" of ethics, one that all else builds on. But most philosophers agree they have failed. What seems best, is to always have a revisable rule, in light of new facts or if facts are shown to be false.

A warranted question is to ask if consequentialism will not meet the same fate as deontology then. As for example Utilitarianism, is also some kind of assumed principle; "to maximize utility". As Singer (2005) discusses, it probably is necessary to assume some intuition, as for example the hedonistic approach with the ultimate goal of *happiness*. But perhaps it could be argued that deep down, at the "rock bottom", the same kind of definite assumption from an

implicit principle to *explicit* is done. If this is a problem that will come back and haunt utilitarianism, or *any* ethical theory, I will leave for more specific discussions of if we end up in nihilism or skepticism. Nevertheless, to me it appears to be another kind of distinction, when claiming one thing to be (1) justifying intuitions, and another (2) aggregating conclusions.

9. Virtue-learning the Elephant

The implicit approximations or quick automatic *assumptions* of given contextual situation, could perhaps be comparable to a person's character traits. They would be the action an individual normally automatically performs in given situations. In other words, it would be an important part of *virtue ethics*, what kind of character a moral agent should have, as *friendly, helpful*, or to have *courage*. These implicit character traits could be passively acquired from innate presumptions or learned from, for example, social exposure. But there was also an option proposed of "explicit surveillance", like learning to drive. The previous chapter, suggesting implicit principles as approximation, would also highlight the importance of explicit and "more exact" control.

I believe that Haidt (2006) have a concrete example of how this could be accomplished with moral character traits, virtues. He describes how Benjamin Franklin, one of the founding fathers of the United States, did to become virtuous. Franklin wrote in his memoirs how he noted "habit took the advantage of inattention" (Franklin, 1840, pp.33), as could be interpreted, how the implicit system took over control with the lack of the explicit concentration. His solution, was to list thirteen virtues he wanted to improve on, for example: "Temperance.---Eat not to dulness: drink not to elevation." (Franklin, 1840, pp.34). He made a spreadsheet, of 7 times 13 squares, for the thirteen virtues and seven days of a week. Every week he concentrated on one virtue, and every time he noted himself to have failed to uphold the virtue, he marked it with a black spot. After thirteen weeks had passed, he started all over again. In time, he noted how the sheet was less and less stained with black spots.

To me, this shows a possibility of explicit control when learning a new moral habit, comparable to learning to drive. One can perhaps marvel at Franklin's discipline and motivation, but perhaps it is also what it takes for someone to "reprogram", or relearn, one's own Elephant,

the implicit system. Potentially this would also highlight the importance of (explicit) education, either if it is the subject itself that monitors himself, or if it is by another. Like the driver's teacher, a supervisor, mentor or a teacher, that monitors the pupil.

The alternative to the active, explicit control, would be to trust the innate potential, or it's social exposure. Perhaps it could be argued that a passive learning system of an appropriate social context would be possible. As has been shown in this paper however, the potential for implicit biases like discrimination (inequality) and egoism (of genes), are real possibilities.

To me, this highlights the importance of the ability of an explicit system that monitors. And, there seem to be two ways leading to the same goal of "explicit conclusion". Either (1) an explicit system that calculates new situations the implicit system cannot handle, or (2) an explicit system that learns preset principles for the implicit system to make use of.

One could argue that the explicit control as a flexible system is *always* necessary, but there are big limitations to our *working memory*. It has been suggested the number of objects, or "*chunks*" of information, is 7 ± 2 (Miller, 1956), or only *4* (Cowan, 2001), that can be held simultaneously in working memory. Either way it is severely limited. It does then make sense to off-load this limited resource of attention to implicit principles that with less effort can attend to certain cognitive work.

There is also what can be called an *alternative version* of when one has "control", as in when one can stay true to the current task or goal, or "free will" if you like. So far I would claim it has been presented when one uses the explicit system, but the alternative version is that it when as much as possible is *implicit.* When most things are "automatic", the Rider on top of the Elephant only sit quietly and monitor, and do not actively need to calculate the result of actions in a given context. So, "most activation of the explicit system" would not have to be deemed "most control". In general, I would claim it might be interesting to note, that there is potentially two roads leading to the same goal.

I would suggest empirical support for this method can be found as well. In the Stroop task (chapter 2), in which participants had to perform the very cognitively demanding task of naming the color of the ink and not read the word. Spelke (1976) continued to train participants in the task. With about 85 hours of training, subject were essentially as fast as normal tasks at naming the correct color of the ink.

10. Dual-level Consequentialism

The Dual-process theory of the morally reasoning brain, could be seen to support a kind of "Two-level utilitarianism" as originally proposed by Richard M. Hare (1981). He distinguishes between (1) a set of *intuitive* moral principles and a (2) *critical* level of moral reasoning. The latter should take over in rare situations and with conflicting intuitive principles (ibid. pp.26), to avoid human (*cognitive*) errors and biases in moral decision making (ibid. pp.38). Hare proposes an analogy to an "archangel" and a "prole" (ibid. pp.40-46). The first is all knowing and can critically calculate the correct response to act upon in any given situation, as the latter is limited to only the intuitive level and principles.

The theory is often seen as a synthesis between "rule utilitarianism" and "act utilitarianism". The second is often deemed to be too demanding, or even impossible for an ordinary person. One would have be forced to calculate the maximized consequence of any given action and situation, with "imperfect knowledge" of the world (Hooker, 2004). Hare (1981) also argues that his theory could be seen as a synthesis between deontology and consequentialism. On one hand we have the set of intuitive moral rules as "prima facie rules" and on the other hand, the critical, consequentialistic aggregating of maximized utility.

To me, Hare (1981) seems to have identified the Dual-processes of the brain. The implicit system would be the set of intuitive moral principles, and the explicit system the critical thinker in rare situations. The objection of *imperfect knowledge* can also be compared to the conclusion I suggested, that implicit knowledge are *approximations*. It has also been questioned how the intuitive, "non-utilitarian", way of thinking ever could be deemed as a good idea, or even that these distinct systems actually existed (McNaughton, 1988, pp.180). The presented empirical data in this paper I would argue to support the existence of these two systems, and also to show how these two systems could function together in a symbiosis.

However, it could perhaps be argued that the importance of virtue ethics, as an alternative way to reach the same goal, is lacking in this theoretical framework. As it is often seen, it would also be something as in between the two extremes of deontology and consequentialism, and could explain how and in what manner to acquire new principles for the intuitive level.

It seems also that neither the implicit system nor deontology has to be excluded, even if it does not seem to be able to survive by itself. Granted, statistically it would most of the time make correct conclusions, but it would be difficult to justify the *incorrect* conclusions of all the *rare* situations not included in the approximation. Nor would the explicit system survive by itself. It would be *too cognitively demanding* for us to calculate every situation. Thus, it makes "cognitive sense" to make use of both the systems given to our brain.

Perhaps a warranted objection could be if this is not a Naturalistic Fallacy by itself: (i) Our brain, arguably, is best understood with a "Dual-process theory". (ii) "Two-level utilitarianism" is a theory that seem to correlate. (iii) Thus, we *ought* to use the discipline of "Two-level utilitarianism" in moral deliberation.

But what is argued is not: "*this* is the way it *IS*", so "*this* is the way it *OUGHT* to be". I would grant, that perhaps there is a consequentialistic approach presupposed, of the constant guessing game and change of judgment in light of new facts. But the reason for support in Two-level utilitarianism is more pragmatic I would argue. It is the "moral system" in our brain which we have to use to make judgments and consequently act upon. So, a consequentialistic conclusion *could* be that to maximize consequences using "our moral judgment system" given to us, the available facts seems to show we *should* use both systems.

To summarize, there seems to me to potential be support for all of the three major directions in moral philosophy, deontology, virtue ethics and consequentialism. Neither do they necessarily seem to be *mutually exclusive*, rather there could potentially be a very happy marriage between them. To be a moral agents, we could make use of all three disciplines, and the best result is, arguably, when all three are working together. The proposal of "Two-level utilitarianism" by Hare (1981) seems to have captured fundamental parts of how we as moral agents reason; using the "Dual-processes" of our brain. Perhaps a fitting name which sums up this proposal could be a: "*Dual-level Consequentialism*".

11. How Inhuman Could Moral Be

Is this a realistic theory of how to be a moral agent? It may still be criticized, as two-level utilitarianism has been, that it is too demanding on the human limitations. To be Hare's all-

knowing "Archangel" would be impossible. And to train your character traits as Benjamin Franklin, is perhaps neither something you can demand of everyone. It does seem like most people, for most of the time, actually are governed by this implicit system, and few have put down as much work as Benjamin Franklin.

Similarly, Greene (2008) ends his argumentation in "The Secret Joke on Kant's Soul", by stating an open question. "Where does one draw the line between the nearsightedness of human moral nature and obliterating it completely?" (Greene 2008, pp.76). Should we stop caring more about loved ones just because science tells us these impulses only exist because they have helped our ancestors spread their genes in evolution (Hamilton, 1964)? Can it really be deemed blameworthy to be more likely to help one's own child instead of another? I agree with Greene (2008), that this is a fundamental question, especially for the direction of an ethical theory as proposed in this paper. In this last chapter, I will in try to sketch a possible answer.

To me, it seems as separate questions; (1) what is humanly possible and (2) what a normative theory may demand. Singer argues that a normative theory should have the possibility to reject all common sense intuitions, to "[i]gnore all our ordinary moral judgments, and do what will produce the best consequences" (Singer, 2005, pp.346). I would argue that it also could be formulated as, if we want the possibility of an exact and correct answer, we need to have the *option* to reject all implicit approximations. In other words, an ethical theory should be able to demand something completely "*inhuman*", if the human condition is defined as including our intuitions. The alternative would become what Greene (2008, pp.75) notes as the *anthropocentric* view, an ethical theory conditioned and centered around the human intuitions. In a similar fashion as when Earth was deemed the center of the Universe.

The other question is what we can demand of people and what can we blame them for not doing. Jennifer Saul I believe to give a possible answer, as she theorizes about how to regard the implicit biases as discrimination of women. She suggests two reasons why not to blame and implicit bias. The first is what I would call the *honest intention*. "A person should not be blamed for an implicit bias that they are completely unaware of, which results solely from the fact that they live in a sexist culture." (Saul, 2011). It seems problematic to blame someone of an unconscious behavior. Perhaps the most obvious case is the "split-brain patient" confabulating about what he is doing by the table it was instructed to walk over to. He is making up a false reason and is telling a lie, but he is doing it without intent. On contrary, he has a pure and honest

heart. But as Saul (2011) continues, even if this behavior is deemed not to be *blameworthy*, it can still be deemed to be *bad*. It can still be deemed to be an unwanted consequence or effect of a specific behavior.

The second reason is more practical. "What we need is an acknowledgement that we are all likely to be implicitly biased—only this can provide the motivation for what needs to be done." (Saul, 2011). I believe this could be applied to moral in general, as it seems we have all kinds of implicit biases. The happy marriage between the major ethical theories proposed in the last section, would not be happy by claiming "one is bad and needs to be controlled", as in, the implicit system. The way to continue forward seems also to me, to only be possible if we acknowledge both systems and blame neither for their existence. In essence, the implicit system can be claimed it just *IS*, a fact, which if accepting the is/ought-distinction, does not entail any normative worth, nor blame.

A consequentialistic approach, I would suggest, would perhaps be to regard the emotion of blame as intuitive in itself, as well as the need to deem something morally blameworthy. They can be seen as evolutionary making sense, but remain approximations of situations. To be more correct and to take into account the "rare" situations, we need to justify these intuitions explicitly. Perhaps an *unintuitive* conclusion would be that the reason someone is doing what he/she is doing, is an *unconscious presumption,* the intuitive assumption seems to be that all things have a reason and intention behind it. This would be how children thinks, and most things happening to us and even more so if it is another agent, we have the intuition of intention behind the action. However, as the example with discriminating implicit biases, these type of unconscious presumptions may at least not be as rare as we would like to think they are.

It should be noted that just as the implicit bias and system is not *denied*, neither is the emotion of blame. Nevertheless, to me it appears that perhaps this will lead to *almost* entirely give up on the concept of blame and *blameworthiness* as part of an *ethical theory*. But myself I think I would be willing to accept that.

Summary

In this paper I have reviewed a "Dual-process theory" of mind. The traditional theory of a reasoning agent, as *deductive* and able to calculate conclusions using *expected utility* when solving problems, was deemed not to be sufficient. In the contemporary model the deductive reasoning was retained, as an *Explicit* system, but an intuitive and emotional system was introduced as the *Implicit* system. Together this theory seems to be able to predict how we deliberate and solve problems fairly well, also when the nature of the problems are *moral*.

I have attempted to generalize five ways how we potentially can derive moral knowledge from the literature and examples discussed in this paper. From these I see a potential explanation of what kind of mistake or fault it can be claimed is performed when deriving explicit rules from implicit principles. I suggest that at least one aspect of the mistake, as with the *Naturalistic Fallacy*, is that the implicit principle is an *approximation*, but the explicit rule is assumed to be an *exact* and *definite*. The mistake is then when one would try to use this false assumption of an exact rule, to cases which was not covered by the original approximation. I do not believe it can be solved by claiming we implicitly have the exact distribution, as for example "70/30". We humans are notoriously bad at thinking in statistical probabilities. As the tribe in the Amazon mention in the beginning, we do not appear to have any kind of "*natural sense of numbers",* unless we explicitly are taught.

This sheds light on why the more flexible explicit functions are important, to include correct responses to rare or unique situations. That either involves limited amount of chunks that could be aggregated, or when the *ready-made* implicit principles is lacking or conflicting. It is however a very limited and effortful resource to use, and it would be challenging to use in all situations. This then in turn sheds light on why the implicit system is important. With explicit monitoring control, even if very effortful, we have the ability to re-program or re-shape our Elephant. To me, this seem rather hopeful and liberating, as it entails that we as humans are not left without control of our actions, to evolutionary and "socially exposed" implicit conditions.

So what kind of fault can Kant really be accused of to have committed? He didn't seem to disagree on what grounds moral rules should be decided, it can not be observations or facts alone. But perhaps it was when attempting to apply the *definite explicit rule* of the Categorical

Imperative to real situations. No such rational and unconditional rule has yet been found, arguably neither in physics or any other scientific field, and the world remains a mysterious guessing game. Hence, this would be why we still would need to calculate new principles, to new situations.

As far as an ethical theory would go an important account in *moral psychology* appears to be the "Dual-process theory" of mind. In an introspective feat, Richard Hare could be claimed to have identified and taken the two processes of mind in account when proposing his "Two-level utilitarianism". I believe an important addition would be a bigger emphasizes on virtue ethics, if it at all could be claimed to be included in the mentioned framework. Together with the explicit "Elephant training" as mentioned above, I believe it could show *some* support for the three major disciplines in ethical theory, they do not appear to be mutually exclusive. As we learn more about to what extent these two processes interact, cooperate and what it is they actually produce, I personally believe that an even bigger emphasis has to be put somewhere in the vague middle part of the two extremes, between deontology and consequentialism.

Perhaps more importantly, the implicit and the explicit system seems to be what is given to humans and animals alike, to solve problems in the world. Equally as important as to decide what is morally good and bad, seems then to decide what we can praise or blame an agent for performing.

References

Bartels, D. M. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. *Cognition*, 108: 381-417.

Baumeister, R., Bratslavsky, E., Finkenauer, C., & Vohs, K. (2001). "Bad is Stronger than Good". *Review of General Psychology* 5 (4): 323–370.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? *American Economic Review*, 94, 991–1013

Borg, J. S., Hynes, C., Van Horn, J., Grafton, S. and Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: an fMRI investigation. *Journal of Cognitive Neuroscience*, 18, 803–17.

Clark , A. (2008). *Supersizing the Brain*. Oxford University Press.

Cowan, N. (2001). "The magical number 4 in short-term memory: A reconsideration of mental storage capacity". *Behavioral and Brain Sciences* 24 (1): 87–114; discussion 114–85.

Dalla Barba, G. (1993). Confabulation: knowledge and recollective experience. *Cognitive Neuropsychology*, 10(1), 1-20

Dawkins, R. (1976). *The Selfish Gene.* New York City: Oxford University Press. ISBN 0-19-286092-5.

Edwards, W. (1954). The theory of decision making. *Psychological Bulletin*, 51, 380–417.

Franklin, B. (1840). *Memoirs of Benjamin Franklin*, McCarty & Davis, Philadelphia.

Fredrick, S. (2005). Cognitive Reflection and Decision Making, *Journal of Economic Perspectives*, Volume 19, 4, Fall 2005, pp. 25-42).

Gazzaniga, M. S., Ivry, R. B., Mangun, G. R. (2008). *Cognitive Neuroscience: The Biology of the Mind (3rd ed.)*, W.W. Norton & Company.

Gazzaniga, M.S., & LeDoux, J.E. (1976). *The Integrated Mind.* New York, Plenum Press.

Gigerenzer, G. (2004). Dread risk, Septemper 11, and fatal traffic accidents. *Psychological Science,* 15, 286-287.

Gommer, H. (2011). *A Biological Theory of Law: Natural Law Theory Revisited.* Amazon Publishing, 2011. Available at SSRN: http://ssrn.com/abstract=191568

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

Greene, J. (2003). From neural 'is' to moral 'ought': What are themoral implications of neuroscientific moral psychology? *Nat. Rev.Neurosci*. 4, 846–849.

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

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

Greene, J. D. (2008). The Secret Joke of Kant's Soul. In W. Sinnott-Armstrong (Ed.), *Moral Psychology* (Vol. 3). Cambridge: MIT Press.

Haidt, J., Bjorklund, F., & Murphy, S. (2000). *Moral Dumbfounding: When intuition find no reason.* Unpublished manuscript, University of Virginia.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108, 814–834.

Haidt, J. (2006). *The happiness hypothesis: Finding modern truth in ancient wisdom*. Basic Books.

Hamilton, W. (1964). "The genetical evolution of social behaviour. II". *Journal of Theoretical Biology* **7** (1): 17–52.

Hardman, D. (2008). Moral dilemmas: Who makes utilitarian choices. In Hare (ed.), Hare Psychopathy Checklist--Revised (PCL-R): 2nd Edition. Multi-Health Systems, Inc.

Hare, R. M. (1981). *Moral Thinking*. Oxford Univ. Press. ISBN 0198246609.

Hauser M., 2006. *Moral Minds: How Nature Designed our Universal Sense of Right and Wrong*, HarperCollins, New York

Hooker, B. (2004) 'Rule Consequentialism', *The Stanford Encyclopedia of Philosophy* (Spring 2004 Edition), Edward N. Zalta (ed.), Accessed 24-7-0.

Howard-Snyder, F. (2002). Doing vs. allowing harm. In E. N. Zalta (Ed.), *TheStanford encyclopedia of philosophy.* Available at http://plato.stanford.edu/entries/doing-allowing/.

Hume, D. A (1739/1978) *Treatise of Human Nature* (eds Selby-Bigge, L. A. & Nidditch, P. H.

Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition*, 11, 123–141.

Kahneman D. (2003) A perspective on judgement and choice. *American Psychologist*. 58, 697-720 & West 2000, pp.658-659.

Kahane, G., Wiech, K., Shackel, N., Farias, M., Savulescu, J., Tracey, I. (2011) The neural basis of intuitive and counterintuitivemoral judgment. *Social Cognitive and Affective Neuroscience Advance Access*, published March 18.

Kahane, G., Shackel, N. (2010). Methodological issues in the neuroscienceof moral judgment. *Mind and Language*, 25(5), 561–82.

Kant, I. (1785/1983). *On a supposed right to lie because of philanthropic concerns.* Indianapolis: Hackett.

Kant, I. (1785/1959). *Foundation of the metaphysics of morals* (L. W. Beck, Trans.). Indianapolis: Bobbs-Merrill.

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(7138), 908–911.

Leonard, Thomas C. (2009) Origins of the Myth of Social Darwinism: The Ambiguous Legacy of Richard Hofstadter's Social Darwinism in American Thought. *Journal of Economic Behavior & Organization* 71, p.37–51

McIntyre, A. (2004). Doctrine of Double effect. *The Stanford Encyclopedia of Philosophy,* publiched 2004, rev. 2011. Available at http://plato.stanford.edu/entries/double-effect/.

McNaughton, David A. (1988). *Moral Vision*. Blackwell Publishing, ISBN 0631159452.)

Mikhail J., 2008. "Moral cognition and computational theory", in W. Sinnott-Armstrong (ed.), *Moral Psychology*, 81-91

Miller, G. A. (1956). "The magical number seven, plus or minus two: Some limits on our capacity for processing information". *Psychological Review* 63 (2): 81–97.

Moore, Clark, & Kane (2008). Who shalt not kill?: Individual differences in working memory capacity, executive control, and moral judgment. *Psychological Science*, 19: 549-557)

Moore, G. E. (1903). *Principia Ethica.* Cambridge: Cambridge University Press

Moscovitch M. (1995). Confabulation. In (Eds. Schacter D.L., Coyle J.T., Fischbach G.D., Mesulum M.M. & Sullivan L.G.), *Memory Distortion* (pp. 226-251). Cambridge, MA: Harvard University Press.

Nisbett, Richard, & Wilson, Timothy. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review,* 84, 231-259

Pica, P., Lemer, C., Izard, V., Dehaene, S. (2004). Exact and Approximate Arithmetic in an Amazonian Indigene Group, *Science* 15 October 2004: Vol. 306 no. 5695 pp. 499-503 DOI: 10.1126/science.1102085.

Rooth, D. (2007). Implicit discrimination in hiring: Real world evidence (IZA Discussion Paper No. 2764). Bonn, Germany: Forschungsinstitut zur Zukunft der Arbeit (Institute for the Study of Labor).

Sahlin, N-E., Wallin, A., Persson, J. (2010). Decision science: from Ramsey to dual process theories. Synthese (2010) 172, 129–143.

Saul, J. (2011, work in progress). *Implicit Bias and Women in Philosophy,* University of Sheffield, http://www.shef.ac.uk/philosophy/research/publications/saulj

Shettleworth, J. (2010). *Cognition, Evolution and Behaviour.* Oxford University Press.

Singer, P. (2005). "Ethics and Intuitions.", *The Journal of Ethics*. October, 2005.

Spelke, E.S., Hirst, W., & Neisser, U. (1976). Skills of divided attention. *Cognition*, 4, 215–230.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? Behavioral and Brain Sciences, 23(5), 645–664.

Sun, R. (2001). *Duality of the Mind.* Mahwah, NJ: Lawrence Erlbaum Associates.

Thomson, J.J. (1986). *Rights, Restitution, and Risk: Essays, in Moral Theory,* Cambridge, MA: Harvard University Press.

Wason, P. C., & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem? *Quarterly Journal of Experimental Psychology*, 23, 63–71.