How can we know what is “moral”?

Philosophical commitments in empirical research on moral judgment

Dissertation at the

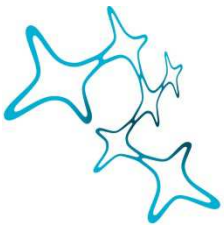
Graduate School of Systemic Neurosciences
at the Ludwig-Maximilians-Universität München

submitted by

David Kaufmann
Research Center for Neurophilosophy and Ethics of Neurosciences
Ludwig-Maximilians-Universität München

Munich,

10.09.2015



Graduate School of
Systemic Neurosciences
LMU Munich

Supervisor: Prof. Dr. Stephan Sellmaier
Second expert appraiser: Prof. Dr. Stefan Glasauer
Third expert appraiser: Prof. Dr. Stephan Schleim
Date of oral examination: 12.06.2015

Contents

1. Introduction

> God appears to man and says: "Believe in me, for without belief, I'm nothing"
>
> Man replies: "Ha! That was empirical evidence of your existence! Therefore, I know you exist. Evidence defeats belief, so, I don't believe you exist. So, you don't exist. Q.E.D."
>
> "Oh, how come I didn't think of this before!" says god and disappears in a puff of logic.
>
> "Wow, that was easy." says man, proves that black equals white and is run over by a truck at the next zebra crossing.
>
> *Very loosely cited form Douglas Adams: "The Hitchiker's Guide to the Galaxy"*

This little episode from Douglas Adams's novel "The Hitchhiker's Guide to the Galaxy" that I modified slightly to fit my needs teaches us several lessons that are of importance for this thesis. It shows us how easily misunderstandings can occur and that these misunderstandings can have truly disastrous consequences if overlooked. The misunderstanding at hand in this situation is rooted in the different meanings that god and man attribute to the term *belief*[1]: While *belief* on the one hand can be taken to mean something synonymous with *conviction,* on the other it can be taken to mean something synonymous with *assumption*. Belief in the sense of *conviction* can be described as a state of mind in which a person would agree to a certain sentence once it is stated – *belief* in the sense of *assumption* as a state of mind in which a person would agree to that sentence once it is stated, but only *provisionally*. While the former is a *condition* for someone to have knowledge (I cannot know sentence X if I am not convinced of X), the latter is something that can be *juxtaposed* to knowledge and assumed to be less certain (I will not assume sentence X if I know better).

While it makes sense to issue the statement "believe in me for without belief I am nothing" with an understanding of *being convinced of something,* it does not seem sensible to understand the term as *assuming until further notice*. In man's answer, things are the other way around: knowledge does defeat belief only if the latter is understood as *assuming until further notice* – but not if *to believe*

[1] In order to make the text more readable, I decided to put concepts in italic when talking about a concept or word and not about the thing the concept or word refers to. For example: "*Paris* consists of nothing but five letters. There is no Eiffel Tower in *Paris*." Note however that not everything written in italic will refer to concepts or words – sometimes it will just be to highlight a certain aspect within a sentence. I trust the judgment of the reader to recognize which use I have in mind in the respective situation.

means *being convinced*. God and man are accidentally speaking about two similar, but different things - with consequences that are as cruel as they are humorous.

This thesis deals with situations of a similar kind, even though not as cruel and not as humorous. It deals with similar, yet diverging interpretations of *moral judgment*. I argue that *moral judgment*, a term referring to a matter which has been and is still being investigated from a number of different perspectives in interdisciplinary research, is used in a wide variety of ways – depending on the perspective. With *moral judgments* I refer to beliefs like:

a) "It was wrong of Peter to lie to his mother."
b) "It is wrong to commit tax fraud."
c) "Killing is worse than beating up, but better than torturing."
d) "A good person will always help those in need."
e) "For the ancient Greeks, it was ok to own and sell human beings like cattle, but nowadays it would be very wrong."

Intuitively, affirmations of these sentences can all be considered moral judgments; in contrast to for example "bunnies are cute". But just like *belief*, the abstract term *moral judgment* can be interpreted in different ways: it can be said to be a judgment about the *character* of a person (like example d) or a judgment about an *action* (like example a); it can be said to be about ordering different options for action according the criterion "*morally better*" (like in example c) or just to be about *right or wrong* and nothing in between (like example b); its truth can be said to be *objective* (like in example d) or *relative* (for example relative to time, place, or culture, like in example e). These are just some possibilities for ways in which misunderstandings like the one about *belief* might originate. Note that misunderstandings about the normally rather fuzzy term *moral judgment* can be expected to arise mainly from the need to *specify* what *moral judgment* is – e.g. in order to make the indistinct common language concept of moral judgment more concise and apt for research and/or argument. The misunderstandings about *moral judgment* which I describe in this thesis are not an issue within our everyday language – they are however crucial in academic research about the subject matter. In this work, I demonstrate that misunderstandings along the lines of the just presented ambiguities are in fact not only common in interdisciplinary research practice but also frequently overlooked. I will therefore try to help advance interdisciplinary research on moral judgment by making these conceptual danger zones explicit. As a first step, I establish a theoretical framework that sets the stage for my investigation. This framework comprises a specific idea of what a scientific theory is and what it is that I understand as the *meaning of a term*. As a next step, I introduce the notion of "philosophical commitments" in psychology or neuroscience of moral judgment, to formulate my hypotheses more precisely and to make them easier accessible to the reader.

1.1. Interactions between psychology and philosophy of moral judgment

The research field of psychology of moral judgment was established in the late sixties of the twentieth century, when Kohlberg came up with his stepwise development model of moral judgment in infanthood (Kohlberg, 1969), a reason based model of moral justification and moral judgment that has been –with minor qualifications – part of the psychological orthodoxy until recently. Beginning in the nineties, with a boom of studies and experiments that showed the importance of intuitive and automatic decision making for many situations in our daily lives[2], this picture started to crumble. Based on empirical foundations established in that time, a "new age" of psychological research about moral judgment began in the first decade of the twenty-first century, with a couple of new rivalling theories of moral judgment appearing on the scientific stage[3]. Some were drawn from linguistics (Mikhail, 2007), others from neuroscience (Greene, 2004), social psychology (Haidt, 2001), and more "classical" cognitive psychology (Sinnott-Armstrong et al., 2010). What they all had in common was how they did *not* focus on moral justification as the primary object of research, but instead on the often unconscious psychological processes underlying moral judgments. The fact that these processes were unconscious and not "steered by consciousness" led to a lively discussion around these approaches and the supporting empirical studies in public as well as in moral philosophy, where the philosophical relevance of these findings is still hotly debated: while several authors (among them Jesse Prinz, Joshua Greene and Jonathan Haidt) suggest the inference of philosophical conclusions from or at least with the help of empirical findings, there has been a substantial body of criticism mainly about the methodology of single prominent empirical studies (for example Kahane (2012), Sauer (2012)) as well as of the idea of empirical or *experimental philosophy* as a new kind of approach to philosophical problems in general. (Sosa, 2006, for an overview of the debate see: Kaufmann & Kleinknecht, 2013)

This general criticism of empirical methods for making philosophical points has a history that goes further back than the recent attempts to construct philosophical arguments that rely on findings from cognitive science. The idea of challenging philosophical hypotheses with empirical data has had very illustrious proponents over the course of the 20th century, one of the most important W.V.O. Quine, who suggested to "naturalize epistemology": the basic idea, propagated in the essay *Episte-*

[2] For an overview see Bargh, (1994)

[3] For a brief overview, see Waldmann et al.(2012)

mology Naturalized (Quine, 1969), was that as the question of the nature of knowledge and understanding cannot be answered from a purely conceptual (a.k.a. philosophical) stance and that one should better leave that question to cognitive science. This essay can be considered as the birth of the movement of modern naturalism in philosophy, a philosophical movement convinced that the rigor and the exactness of science would in the end suffice to solve all (or at least many) philosophical problems – or that at least philosophy could not play the role of the foundation of science but vice versa (for an overview, see Maddy, 2007). Many good arguments have been exchanged for and against this idea, until the already mentioned developments in moral psychology led to the first sincere attempts to actually explain what moral judgment is with the help of scientific method. This work will make a contribution to the question of whether naturalism makes sense by investigating the chances for success of this enterprise by looking at what has been achieved so far. But the contribution will not take the form of a clear argument for or against naturalism. The contribution will consist in insights of what the process of scientific explanation of moral judgment could look like. And the answer will be that philosophically uninformed science on its own is not enough. The reason for this is that as soon as we want to investigate rather hard-to-define matters like moral judgment, we will still need a rather precise idea of what we are looking for. And just as I suggested before, the concept of *moral judgment* can be interpreted in a variety of ways. For specifying moral judgment in a sensible and (just as importantly) explicit way, it would seem foolish not to take into account the loads of diverging approaches to specify *moral judgment* that have been undertaken and defended by various philosophers over the centuries.

1.1. Explication and scientific fruitfulness

The problem that our common language concepts (and not only our common language concept of *moral judgment*) are often not as precise as science would like to have them, and that making concepts more precise poses the question which of the more precise concepts one should choose has already been known for quite some time. Rudolf Carnap describes the issues that we face in this respect very precisely in the first chapter of his “Logical Foundations of Probability” (Carnap, 1967): Certain concepts are in their everyday usage too vague to be properly used for scientific purposes. That is why Carnap argues that even when making *philosophical* enquiries about them (let alone scientific ones), one ought to specify very precisely what one means by them. Carnap calls this refinement of concepts *explication*. But of course there is more to explicating a concept than just assigning a more precise meaning to a term that is *somehow* similar to “the old one”. He formulates several

conditions for a term to be the explication of another: the conditions of s*cientific fruitfulness*, *simplicity*, *exactness*, and *similarity*.

The most basic condition would be similarity: In most situations in which we have used the old term, we should still be able to use the new one. It should not be a coincidence that most entities referred to by the unrefined concept (for example the everyday notion of *utility)* should still be referred to by the refined concept (e.g. the economist concept of *utility* that plays a major role in microeconomics*)*.

While I come back to the points of *simplicity* and *exactness* later, the most important benefit from explicating a concept (and hence the most important criterion for its goodness) is increased **scientific fruitfulness** of the explicatum. *Fruitfulness* can be defined along the lines of practicality for theory building:

> "A scientific concept is more fruitful the more it can be brought into connection with other concepts on the basis of observed facts; in other words, the more it can be used for the formulation of laws." (Carnap, 1967, p. 6)

This dependence of *scientific fruitfulness* and hence *explication* on relatability to observed facts is crucial for my argument, as it serves to illustrate a key challenge of interdisciplinary research: different theoretic backgrounds of researchers entail different observed facts *of interest*. Different observed facts of interest will in turn result *in different ideas of what is a good explication*.

Carnap was highlighting the importance of a scientific concept being as close as possible to the original meaning of the term while pointing out the promises of improving scientific fruitfulness of a concept through explication. I am making a similar, yet different point: I am highlighting the discrepancies between different explications of the same prescientific concept made in respect to different scientific disciplines.

Let me give an example referring to research on moral judgment: An explication of *moral judgment* that puts great emphasis on linguistic behaviour might be regarded as particularly scientifically fruitful by a linguist – however as not at all scientifically fruitful by a behavioural biologist. Remember that *scientific fruitfulness* was determined by how easily the explicatum was to be put in relationship to other entities the respective research field would be about. While *moral* understood in terms of linguistic behaviour can be related easily to linguistics, it is not at all relatable to findings in laboratory mice or apes. The biologist will therefore rely on a different concept of *moral* in order to investigate moral judgment, probably one that is rather behaviour-driven. The problem is however not limited to interdisciplinary discourse: with different ways of operationalizing *moral judgment*, different

ways of *measuring* moral judgment, there are bound to be different ideas of moral judgment even *within* a given discipline.

If the reader had been wondering about how different, possibly conflicting meanings of *moral judgment* might come into existence, this section should have provided the answer: with varying research interests of researchers, different explications of *moral judgment* tend to be the most appealing.

With reference to psychology of moral judgment compared to moral philosophy, an important point in this direction was made by Jeanette Kennett in her paper "Would the real moral judgment please stand up?" (Kennett & Fine, 2009): she criticizes psychologists for regarding moral judgment as an instantaneous judgment (like judgment "a" on page 6) instead of a long-term disposition to behave in a certain way (like judgment "b" on page 6). She defends reason based concepts of morality on the grounds that the *real* moral judgment is actually a long term disposition to behave and not short-term reactions to certain situations. While I find the validity of her arguments regarding this specific point debatable, the idea of there being a) several meanings of moral *judgment* and b) these meanings not being made explicit at all struck me as just as right as important. If the last pages show how different, rivalling meanings of *moral judgment* could coexist in research, the next pages offer an explanation why they are hardly made explicit and therefore almost never acknowledged.

1.1. From conceptual holism to implicit philosophical commitments

Being educated in the tradition of analytical philosophy of language, I learned in my undergraduate studies how the meaning of a term is dependent on its role in the language or theory of the world in which it appears. One important point about this view is that a theory **is** a language and a language **is** a theory about the world for the following reasons: just like "mass" as used in Newtonian Mechanics receives its meaning from the role it plays in in this theory relating weight, length and density, the term "red" receives its meaning from the role it plays in relating different entities to each other (the red ones, to be more specific) in given situations: under certain conditions, which are mainly about proper lighting, objects with the quality "red" seem to have a similar hue. Note however that as soon as our idea of these conditions changes, our idea of "red" changes. Take for example that I find out that my normally white shirt seems to have a slightly red hue once I enter the hallway of my apartment. Knowing that the shirt which now seems red is under usual circumstances white, I will conclude that the lighting in my hallway is rather distorting and that I should not rely on my colour perception under the lighting conditions in my hallway. My adherence to standards of how to use the

term “red” will force me to reconsider judgments of colour that I make in my hallway. If however I decided not to adhere to these standards and decided to call everything “red” that seems to be red in my hallway, I would use the word “red” in a significantly different way than most people. In my idiolect, my shirt that is white under normal circumstances but would then literally turn red in my hallway. “Being red” in my idiolect would mean something close to the commonly used term “looking red”.

A similar story can be told about the term “mass” in physics when relativity theory succeeded Newtonian mechanics. This idea of *meaning* that is generally called *conceptual holism* has been famously proposed among others by Sellars (1951), Quine (1961), and Wittgenstein (2003).

From this idea of language and meaning one can infer that a term has different meanings if it plays different roles in different parts of a language: The term *red* when applied to red wine will usually correspond to another wavelength (reaching from dark rose to dark purple) than the term *red* when applied to the colour of the soil (sometimes almost orange). The term *mass* will mean something different when I am talking about a big gathering of people (“a mass of people”) and when describing a certain material (“a greyish, sticky, mass”). This leads me directly to the problem of the concept *moral judgment* that this dissertation is about: there are a lot of differing explanations and theories of what moral judgment is and which role it plays in our lives. And in each of these explanations, the term *moral judgment* plays a different role and has different features – just like *red* and *mass* play different roles in different contexts. According to the considerations from philosophy of language above, different philosophical theories about morality do in fact instantiate different ideas of the meaning of the term *moral judgment*. Note that according to this notion of *meaning*, the meaning of a term is not just set by its explicit definition, but also, probably even primarily, by its actual application guided by implicit rules[4]. If we find out something new about a previously defined entity, let’s say the *planet Mars* defined as the planet which is the planet orbiting the sun between Earth and Jupiter, our conception of *Mars* changes – even though the explicit definition stays the same. Let’s say for example that it is discovered that Mars is actually inhabited by a species of blue humanoid aliens. This does not change the definition of *Mars*. It would however dramatically change the role

[4] Important exceptions from this rule are of course highly axiomatic theories like one can find them in mathematics and physics. Here, the way a term is understood within a theoretical framework is almost exclusively dependent on its formal definition. However, these theories have a long tradition of philosophical debate about what they are actually about: Philosophy of mathematics deals with the question whether for example numbers are entities or abstractions, and also the ontological status of unobservable entities has a long tradition of philosophical debate. (Duhem 1954/1906, Carnap 1950, Quine 1953, Van Fraasen 1980, Putnam 1982, Ben-Menahem 2006) Axiomatization seems to shift unclarity rather than to eliminate it.

Mars plays in our world view and the way we think about Mars. *Mars* before the discovery of life will, in my sense, not have the same meaning as after. The meaning of a term is dependent on more than its definition. It is dependent on the role the term plays in our view of the world.

This view of *meaning* has an important implication: it becomes much more difficult to call something the *actual*, or *correct* meaning of a term. If a term is used differently by two people, there is no straightforward way of finding out who is using the term correctly and who is not. As I point out in the concluding remarks of this thesis, the question which is the *actual* meaning of a term is mostly a pragmatic question. This has a very comfortable consequence: I do not need to take a stance about who is going to have the last word in defining what moral judgment is, whether it is philosophers, psychologists or neuroscientists. If there is no *actual* meaning of *moral judgment*, there is no designated method to declare what this actual meaning is.

This leads me to a first important milestone for the hypotheses of this thesis: the meaning of a term is dependent on the role it plays in describing the world. Different descriptions of the world may attribute different meanings to the same term.

Let me now introduce the next step in my initial argument with a little story about communication between theorists of different subject matters: biology and gastronomy. This story highlights how one can validate hypotheses about one's own research interest by taking findings from another research discipline into account. It furthermore illustrates what I mean with inter-theoretical commitment (in this example a gastronomic commitment) and to explain the consequences of such commitments.

While biology's business is explaining and exploring the ways and workings of living things, gastronomy's job is the refinement of cuisine and how to combine eatable things in order to achieve a maximal satisfaction in humans. As almost everything that gastronomy deals with is subject to biology (at least as long as it is alive),it can be expected that there should be biological findings that can serve to argue for certain positions in gastronomy. Imagine for this purpose a biologist and two gastronomes.

Say the biologist defines *sea fish* as "and live in ocean waters". This biologist points out that there is plenty of evidence that all sea fish show high concentrations of Trimethylamine oxide (TMO), a compound that serves for stabilizing the concentration of salt in the sea fish' cells. After the death of the organism, TMO is gradually decomposed to Trimethylamine, the substance responsible for "fishy" smell.

The gastronomes have learnt that fishy smelling, yet edible food is best served with garlic. But while gastronome A understands *sea fish* to consist in "everything you take out of the sea", including cut-

tlefish and lobster, gastronome B understands *sea fish* to be "cranial animal with gills and fins without limbs with digits" that happen to live in the ocean. According to the considerations above, we can call gastronomist A's understanding of gastronomy *gastronomical theory A* and gastronomist B's understanding *gastronomical theory B*.

One can make two observations: The first observation (see figure 1a) is about the biologist: he will have to regard Gastronomist A's definition of *sea fish* as wrong. The biologist will say that (at least from a biological perspective), gastronomist B is right about what *sea fish* means. This is a gastronomical commitment of the biologist: he regards one gastronomical theory B as a better description of the world.

The second observation (see figure 1b) is that gastronomist B (but not gastronomist A) will be able to draw on the biologist's results to argue for the truth of the sentence: "Sea fish should always be served with garlic after a storing time of one day". The biologist had shown that sea fish contain large amounts of TMO that causes fishy smell after a short period of storage. Both gastronomists know that garlic goes well with fishy smelling food. But only one gastronomist shares the biologist's definition of *sea fish*. Therefore only one of them can argue that biology has proven that all sea fish should be served with garlic after the first day.[5] This is a constraint for interdisciplinary arguments between biologists and gastronomists: if their concepts of *sea fish* are incompatible, they cannot argue for hypotheses about fish employing results from the other.

[5] Actually, most sea animals contain some TMO. However, to my knowledge the concentration in sea fish is several times higher than say, lobster. (International Commission on Microbiological Specifications for Foods 1998)

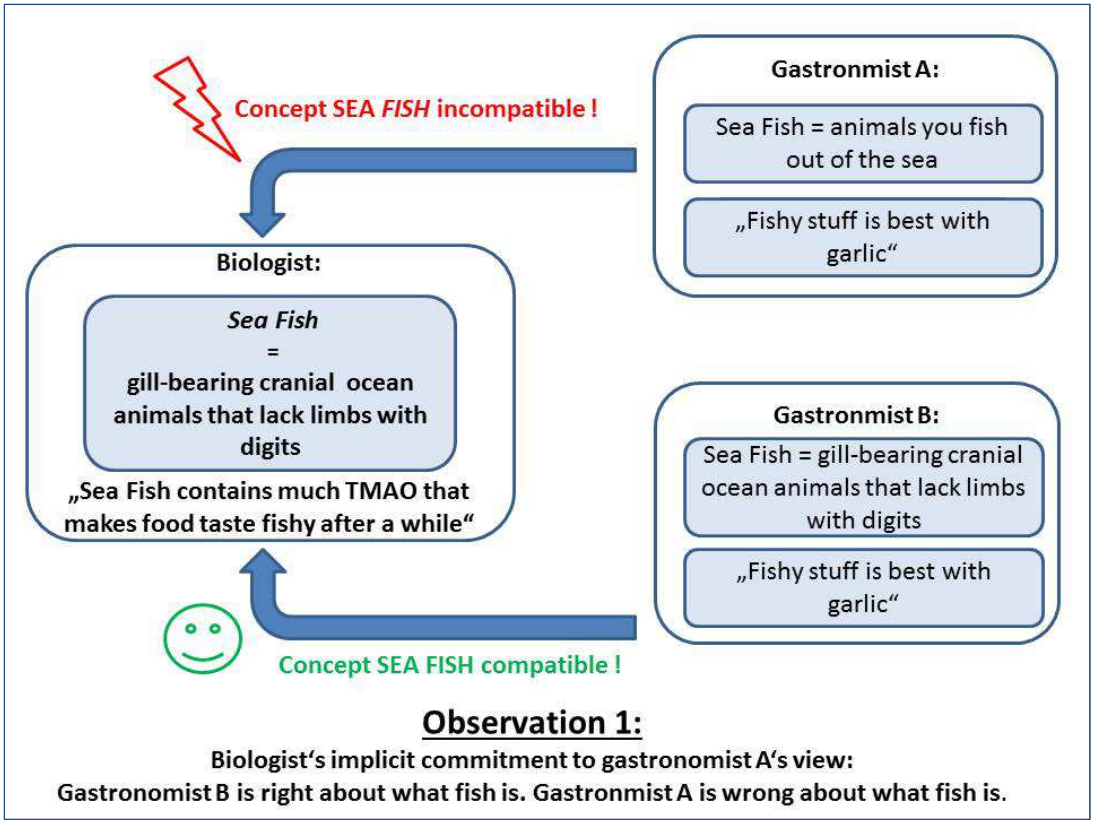


Figure 1a): Biologist's definition of *sea fish* implies a gastronomical commitment

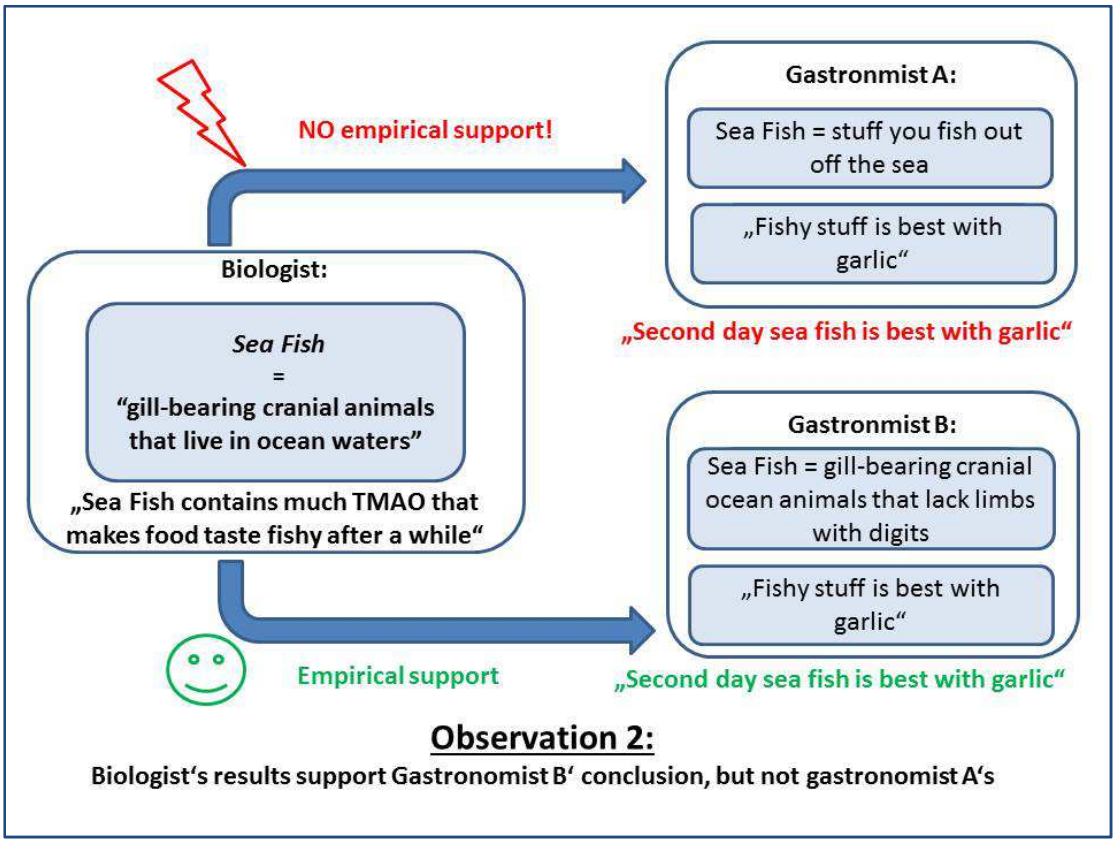


Figure 1b): Biologist's result supports Gastronomist B'S conclusion, but not Gastronomist A's.

The relationship between moral philosophy and psychology of moral judgment can be described in the same way as employing observations about philosophy and observations about psychology: as soon as a psychologist is *using* the term *moral judgment* in a certain way, and there is a philosophical theory that uses the term in a different way, the psychologist is actually talking about something different than this philosophical theory is. If for example a philosophical theory "Kantianism" assumes that moral judgments take the form "X is morally wrong" (like judgment "a" from page 6) and a psychological theory A assumes that moral judgments take the form "X is morally better than Y" (like judgment "c" from page 6), the phenomena they are referring to while talking about *moral judgment* are two different phenomena. A and its adherents are making what I call a *philosophical commitment*: "Kantianism is not about moral judgment in the way our theory uses the term" and therefore: "Kantianism is wrong because it uses 'moral judgment' differently than this theory does – it says that moral judgments are something different than what this theory says they are".[6]

A *philosophical commitment* as I am using the term is the refutation of certain philosophical theories by adaption of a certain understanding of a given concept.

Of course, this pretty narrow minded position about meaning as a theory-dependent property of terms is normally not how one is thinking about fuzzy terms like *moral judgment*. In our everyday language usage, we accept a certain vagueness when using certain terms: Take for example colour terms like "red" – a speaker will normally accept a pretty large variety of hue to count as "red". This however changes when one is doing research – philosophical or empirical. Then one will be trying to become more precise about what one means with a term. In the case of "red", one would define a range of wavelengths that can be measured and which would make certain light beams count as "red". And in the case of psychology of moral judgment, one will have to think about how exactly to ask people for moral judgment. This is the moment when one should be aware of philosophical

[6] For this conclusion, I have to assume that people who are talking about moral judgment have the opinion that they know what the term means and would therefore affirm the position "my concept of moral judgment is correct". I am willing to do so, and I would ask readers with legitimate criticism against this point to consider the arguments given in the philosophical debate about the nature of truth in favor of deflationary theories of truth (for example Frege 1918, Ramsay 1927, Quine 1970). However, even though it is clear *within* the theories that give meaning to different concepts of *moral judgment* what is the *correct* meaning of *moral judgment* - from a meta-perspective, which is the perspective that I take in this thesis, the question what is the *correct* meaning of *moral judgment* becomes less trivial, as I indicated before.

commitments in the sense described – the greater need for exact definition and explicitness that empirical research brings along transforms *moral judgment* from a vague concept to an ambiguous concept with several different, theory-dependent meanings. The problem with this kind of commitment is that these commitments are normally not very evident, as we as common language speakers are not used to think about moral judgment with such conceptual scrutiny. But as soon as one leaves the fuzziness of common language and enters the exactness of science, subtle differences in concept use can lead to talking about quite different phenomena. As one can see in the little story of god and man from the first page of the introduction, even in non-scientific contexts subtle conceptual differences can have manifest consequences. In science of moral judgment, these subtleties will in the end decide what science is looking for: in behaviour, in the brain, in society. This leads to an absurd situation: scientists want to find out how morality works. But unfortunately, it is not that clear what morality is actually supposed to be to begin with.

This lack of awareness of conceptual subtleties becomes especially problematic when empirical results are featured in philosophical arguments: if a study or theory commits itself to philosophical standpoints that either already assume the conclusion of the argument they are supposed to support or if it stands in open conflict with it then the argument is not feasible. Understanding philosophical commitments of empirical research about morality therefore becomes indispensable for empirical philosophy. Let's imagine for example a philosopher who reads that psychologist Peter found out that emotions drive moral judgment. This philosopher knows that Kantianism postulates that reason drives moral judgment. If the philosopher will now write a paper about psychologist Peter falsifying Kantianism, he will make the same mistake as gastronomist A might have. Just like gastronomist A would be wrong to assume that science has shown that garlic goes particularly well with overstored *sea fish* (as he understands it), the philosopher would be wrong to assume that Peter's result falsifies Kantianism. Kantianism and Peter are talking about different phenomena to begin with, just like the biologist and gastronomist A.

As a last step, I want to combine this line of thought with the idea developed earlier about the meaning of a term often being determined not only explicitly but implicitly. The gastronomist and the biologist had a fairly easy guess spotting their conceptual incongruence, as their definitions were extremely explicit. However, the considerations about conceptual holism suggest that there might be many constituents of meaning that are *implicit*, especially for the term *moral judgment*. This can lead to situations like the one in which god and man are not aware of using the same word to refer to different things. It is important that this is not only a matter that affects inter-theoretical communication between philosophers and empirical researchers. As highlighted earlier in the passage on *explication*, we can expect substantial disagreement of working concepts of *moral judgment* even

among empirical researchers. And if the largest part of the meaning of *moral judgment* in science remains determined by implicit rules, interdisciplinary discourse between empirical researches is bound to encounter the same difficulties as the one between scientists and philosophers.

1.1. My central hypotheses

The line of thought that I presented until here suggests three hypotheses:

Hypothesis 1: There are philosophical commitments at work in empirical research on moral judgment

Hypothesis 2: These commitments have strong effects on empirical philosophy

Hypothesis 3: These commitments can sabotage empirical research because they are overlooked easily

This dissertation is about actually finding and describing philosophical commitments in psychology, their effects on empirical philosophy and their effects on psychological practice. This is an uncommon approach to philosophical work. Usually, one would expect a complicated and lengthy argument in favour of a certain position. I, however, do present evidence - accompanied by many small arguments for why the evidence in question should be considered as evidence for my hypotheses.[7] As a result of investigating the overlap between psychology, neuroscience and philosophy, I am forced to jump back and forth between two different philosophical disciplines:

- *metascience or philosophy of science*, that tries to establish how science does its job (in my case by highlighting in how far certain implicit commitments about the nature of moral judgments are made by science), and
- *metaphilosophy* that tries to establish how philosophy does its job (in my case by looking into how these implicit commitments affect philosophical arguments that try to employ scientific findings).

[7] This approach led to some bit of confusion in my thesis advisory committee, whose members would constantly challenge me: "what is your argument?" I would like to take the opportunity to thank them for giving me such a hard time with their questions, as since it forced me to rethink about what I was actually doing – over and over again, until I came up with this concept. This is what I came up with.

1.2. What I am <u>not</u> arguing for: ethics, psychology and what some call *experimental philosophy*

Let me issue a disclaimer at this point. This is a text that goes back and forth between metascience and metaphilosophy and no matter what, it stays *meta*. Accordingly, I deal with a number of hypotheses and how they are derived. My primary interest concerns *how* these hypotheses are derived and defended. I criticize philosophical arguments. I criticize definitions. I criticize that certain small but important conceptual distinctions are overlooked or ignored. This critique does however come from a metaperspective. I am not arguing against particular results, I am arguing that certain lines of reasoning fall victim to a methodological flaw.

In order to avoid misunderstandings, let me repeat myself: my focus is on **method**. This dissertation is about arguments for normative positions without making normative claims. This dissertation is about metaethical arguments without making metaethical claims. This dissertation is about psychological arguments without making psychological claims. I do not intend to make a point ***in*** normative ethics, metaethics or psychology. I intend to make a point ***about*** them.

One last disclaimer: my considerations so far might have reminded the reader of the debate about *experimental philosophy* mentioned earlier. This philosophical brand name is attributed to a group of philosophers inspired by naturalism that try to solve philosophical questions with the help of surveys that investigate laymen's intuitions about philosophical thought experiments. The underlying idea of experimental philosophy is that many philosophical arguments rely on an appeal to intuition, for example in the form of thought experiments like the Gettier-Case (Gettier, 1963) or the Chinese Room (Searle, 1984). These intuitions however do not always seem as waterproof as some philosophical authors might think: in cases like the Gettier cases, laypeople could be demonstrated to have differing intuitions than philosophical authors on hypothetical scenarios (for example about Gettier: Weinberg et al., 2001). This led to the idea of experimental philosophy: philosophers should investigate whether the intuitions that they (or others) rely on in philosophical arguments are indeed commonly shared – and in the optimal case, they should be even able to explain these intuitions psychologically (Knobe et al., 2010). Experimental psychology is objected by its opponents for its characterization of the role of intuition in philosophy on the one hand (Williamson, 2004) and on the other for its claim that surveys are apt to overrule philosophers' intuitions in general (Sosa, 2006).

The reader most probably has already spotted an important difference between experimental philosophy and the kinds of proceedings that I describe in this thesis: Most importantly, I am not interested in survey results trying to overrule anyone's intuition, but rather in psychological research and the application of its results in philosophy. I am interested in an exchange between two differing ways of understanding moral judgment with mutual benefit. That is why neither the literature nor the criticism of experimental philosophy figure in this thesis: I am describing a different phenomenon.

1.3. How I proceed from here

As my claim is about both empirical philosophy *and* psychology of moral judgment, and space is limited, I restrict myself to stick to just one paradigmatic psychological theory of moral judgment in order to highlight how the role that a concept plays in this specific psychological theory does commit this theory to philosophical standpoints. I furthermore restrict myself to highlight the same kind of phenomenon on only one typical empirical study that is taken to support the theory – even though I also mention other studies in due progress to contrast their approach with the paradigmatic one. This theory and this specific piece of confirmatory evidence is the main object of the part of this thesis that can be fully attributed to philosophy of science. As my hypothesis is about the existence of philosophical commitments in principle, this restriction to one paradigmatic example does no harm to the validation of my hypotheses.

After laying out the paradigmatic theory and the paradigmatic study in necessary detail in the second chapter of this book, I look into the way a limited number of philosophically loaded concepts are used in the third chapter. A section of this third chapter is dedicated to each of these concepts. In these sections I present paradigmatic philosophical approaches with which the theory or the study agrees and philosophical approaches that do not agree with the way the sample theory understands the concept. I then point out the changes to the theory that a diverging understanding of the concept would entail. All of this can be said to belong to the discipline of philosophy of science.

In every section of the third chapter, there is however also a subsection dedicated to metaphilosophy: after the assessment the philosophical commitment for the respective concept, the reader is presented one or several philosophical arguments in which the example study of the example theory play a decisive role. It is then assessed whether the use of the empirical data in this particular example of empirical philosophy is feasible or not. This is done by investigating whether the philosophical

commitments of the study do either conflict with the philosophical position they are supposed to support or whether they already presume the conclusion of the argument they are supposed to play a role in.

Subsequent to the rather destructive third chapter, the thesis closes with two suggestions on how to adapt research methodology to a pluralistic application of the term *moral judgment*. Firstly, I highlight ways to devise empirically minded philosophical arguments that are not only consistent in their concept application but that even profit from acknowledging the ambiguous way *moral judgment* is used in research. Secondly, I develop the idea that pluralism of meaning of *moral judgment* makes it necessary to develop tools for inter-theoretical *translation* of findings. I argue that these kinds of tools can be developed with the help of empirical research and exemplify this point with a study that I conducted in collaboration with Professor Stephan Glasauer from the Center for Sensorimotor Research at the Department of Neurology and Gloria Benson from the Neuro-Cognitive Psychology master program of the Ludwig-Maximilians-University Munich.

I finish with the conclusion that accounting for the plurality of meanings of *moral judgment* is not only necessary to keep research findings consistent, but also offers great opportunities for new methodologies and findings.

2. An Example to work with

In German football, there is a saying about the relationship between theorizing about tactics, which players should play on which positions and the practical application of these thoughts: “Was zählt, ist aufm Platz”. This sentence can be translated as “What counts is what actually happens on the playing field”. It is a pragmatic statement that relativizes the importance and validity of theoretical considerations to their practical utility in application. I regard this as a very healthy kind of procedure: something that sounds pretty sensible in theory but does not work out in reality is not helping anyone. It is in this spirit that I apply the ideas that I developed and explained in the introduction to an exemplary paradigmatic psychological approach to moral judgment. As a first step, I introduce two psychological models of moral judgment that an exemplary study is taken to be evidence of. I establish the point that the two models can be understood as one theory since they provide mutual support to each other. After introducing the theory itself, I present the method and set up of a study that supports it. I make explicit the rationale that derives the particular predictions validated by that study from the two models. Besides the exemplary theory of morality itself, it is this rationale and the included definitions and syllogisms that set the stage for the search for open or hidden philosophical premises in the following third chapter.

The present chapter takes the shape of an analysis rather than of a repetition of the paradigmatic approach. It deals with some important nuances about this approach that happen to be misunderstood, misinterpreted or overlooked very easily – often because important information is hidden in the literature that is cited along the way. If therefore the reader should already feel familiar with the theory and the study that I present and analyse on the following pages, I would still recommend taking a glimpse at this passage since its contents go beyond what can be found in the papers in which the respective approach is originally presented.

As I mentioned, the truth must be found on the playing field. Let me now begin to introduce this playing field.

2.1. The Socio-Intuitive Model (SIM) of moral judgment

The first of the two intertwined models of moral judgment that I am to present is Jonathan Haidt's *Socio-Intuitive-Model of Moral Judgment* (SIM) (Haidt, 2001). The key ideas of the model are that

a) moral judgments are not the result of a conscious reasoning process but instead of an unconscious, intuitive process and that

b) the function of reasoning in moral judgment is not the derivation of a judgment from a given situation and a set of moral norms, but rather the mere justification of the already intuitively derived moral judgment.

In other words the reasoning process does not begin with the question "What should I do?" but with the question "How can I justify my judgment that X is right?". The SIM has been one of the most influential psychological models of moral judgment in the last two decades and has been subject to a lot of psychological and philosophical discussion (Waldman et al., 2012; Kennet & Fine 2009; Saltzstein & Kasachkoff 2004;). It is relatively extensively debated and employed in philosophical "secondary literature" which allows me to illustrate the consequences of the many philosophical commitments implicitly made by this model – in the "secondary literature", there is a lot of misunderstandings that can be tracked back to the ignorance of these philosophical premises. On the following pages I therefore first introduce the most important fundamental concepts of this model and then explain the actual content of the model in terms of these concepts. This detailed analysis pays off in the later chapters.

2.1.1. The notion of *moral judgment* in the SIM

A key challenge for any psychological theory of moral judgment consists in formulating a working concept *moral judgment* that is on the one hand loose enough to capture the at times vague meaning of the term, on the other hand narrow enough to make it operationalizeable. Many might consider this task a lost cause and just deny that any scientific research on moral judgment is possible. P. Churchland's position in "Eliminative Materialism" (Churchland, 1981) would for example suggest

such a position[8]. If you do not want to admit defeat so easily however, you will need to make the compromise just pointed out. Haidt is doing this in the SIM:

> "Moral Judgments are therefore defined as evaluations (good vs. bad) of the actions or character of a person that are made with respect to a set of virtues held to be obligatory by a culture or subculture"
> (Haidt, 2001, p. 817)

According to this definition, there are three criteria that have to be fulfilled by an entity X to qualify as a moral judgment:

- X must be an *evaluation*
- X must be *about an action or a character*
- X must be "made *with respect to a set of virtues held obligatory by a culture* or subculture"

What does Haidt have in mind when he uses these terms? Haidt adopts what he calls an *empirical stance* about what moral judgment is, claiming that "in every society, people talk about and evaluate actions of other people, and these evaluations have consequences for future interactions" (Haidt, 2001, p.817). He is distinguishing between evaluations that are about virtues concerning the individual role of persons, like for example being a good cook on the one hand, and evaluations that are about virtues that are applied to everyone in the society or a certain group of people within that society on the other, like being brave (for everyone) or charitable (everyone who is sufficiently wealthy). If somebody fails (or succeeds) to live in accordance to this kind of virtue he or she will be object to criticism (or praise) and social sanctions (or rewards). Evaluations of this later kind are what Haidt has in mind when he speaks of "moral judgments". He explicitly admits that this allows for a large grey area of judgments that are "marginally moral" giving the example of health-conscious subcultures whose members have been shown by Stein & Nemeroff (1995) to regard people who eat cheeseburgers and milkshakes as morally inferior to those who eat salad and chicken (Haidt, 2001, p. 817).

This leads us to a more thorough understanding of "virtue being held obligatory by a culture or subculture": a virtue would have to be

a) universal, that means relating to everybody or at least everybody in a certain group of people, and

[8] If that was ever Churchland's position, it has changed over the years: 20 years after, he presents in "Toward a Cognitive Neurobiology of the Moral Virtues" (Churchland, 2007) some (very sketchy) ideas how moral judgment might work.

b) sanctioned if transgressed.

According to this definition of moral judgment, the domain of moral judgment varies across cultures: what is ok in some cultures (showing one's hair in public, for example) can be considered highly immoral in others. A judgment about a woman showing her hair in public would be exactly then a moral issue for a culture or subculture if in that certain culture or subculture there were a norm relating to that kind of behaviour and this norm were taken seriously enough – which means a group of people (women) were universally regarded as obliged to display the behaviour in question and that non-compliance were sanctioned in a certain way.

2.1.2. The concepts of *reason* and *intuition* in the SIM

Now that the way *moral judgment* is used within the SIM is nailed down, the next step is to clarify the terms referring to the mental inventory that Haidt relies on to explain how moral judgment works. As the model is about the relation between *moral reasoning, moral intuition* and *moral judgment*, all of these terms have to be understood thoroughly to grasp its scope and main ideas. Accordingly the focus of the following pages is on *moral reasoning* and *moral intuition*. Both concepts share a section of this chapter because explaining one of the terms necessitates an understanding of the other and vice versa, as both terms are defined as being mutually exclusive. I start off this section by working my way through the details of a working definition of *moral reasoning*. After that, I shortly explain Haidt's definition of *moral intuition* which is basically the inverse of *moral reasoning*.

Moral reasoning is defined by Haidt in the following way:

> "Building on Galotti (1989), moral reasoning can now be defined as conscious mental activity that consists of transforming given information about people in order to reach a moral judgment. To say that moral reasoning is a conscious process means that the process is intentional, effortful, and controllable and that the reasoner is aware that it is going on (Bargh, 1994)" (Haidt, 2001, p. 818)

This definition gives us two properties that are decisive features of reasoning:

- Reasoning consists in *stepwise* transformation of information.
- These steps are *conscious,* which involves the process of reasoning being
    - *intentional*,

- *effortful*,
- *controllable* and
- that the person who is reasoning is *aware* of this process.

Like in the case of the definition of *moral judgment*, some qualifications should prove useful to ease the understanding of *reasoning*. For this purpose, the subsequent passage runs through the single features of *reasoning* mentioned in the above explication and explains in a nutshell what is meant with these terms.

2.1.2.1.Steps of reasoning

The first important feature about reasoning in Haidt's definition is that it is a stepwise process, in which the single steps consist in activities like weighing evidence and reaching conclusions. But what about these steps and why is there no clear definition of which steps ought to be included and which ones ought not to? In his definition, Haidt refers to Kathleen Galotti's approach to reasoning who in Galotti (1989) defines reasoning as:

> "Mental activity that consists in transforming given information (called the set of premises) in order to reach conclusions. This activity must be focussed on at least one goal (but may be focussed on more than one). The activity must not be inconsistent with systems of logic when all its premises are specified, although there may not always be an applicable system of logic to govern specific instances of reasoning. The activity may or may not be self-contained; that is, people may implicitly add to, subtract from, or otherwise modify any or all of the premises supplied. When original premises are modified, the final conclusion must be consistent with the modified premises. (…) The conclusions may, but need not, be deductively valid." (Galotti, 1989, p. 333)

This is what Galotti tells us about the nature of steps of reasoning: steps of reasoning are directed towards an aim, so there is a direction of reasoning. The steps are supposed to be not incompatible with systems of logic, keeping in mind that additional premises to the given ones would be allowed for. This is a very loose description of what is considered a step of reasoning, but we are already better off than we were before: goal-directedness and some kind of right/wrong restrictions in the form of a "system of logic" give us an a bit more detailed idea of what is meant with "steps of reasoning". The reason for Galotti being so inexplicit about which kinds of transformation steps could be used to

identify reasoning is the problems that one faces when focussing on one specific kind of information transformation step. Galotti gives the example of Evans (1982) who in her opinion leans too much towards deductive syllogisms paying the price of losing descriptive force by losing sight of the everyday practice of reasoning, while Holyoak and Bisbett (1988) would lean too much towards induction, failing to account for the rigor of stepwise deductive reasoning. Galotti argues that in both cases, the definition given does a good job in the laboratory, but does neither really fit our everyday concept of *reasoning* nor our everyday practice of reasoning. Note that her definition is constructed mainly in a way that allows keeping certain processes out of the scope of *reasoning*: memories, intuition, or "gut feelings". This is the main idea behind introducing the vague term of "stepwise transformation": it ought to exclude one-step processes without making too many further constraints. (Galotti, 1989, p. 334)

It is exactly this purpose of the definition that makes it so attractive to Haidt. The importance of "steps of reasoning" for keeping reasoning and intuition apart, even though "steps of reasoning" is only very loosely defined becomes evident in the following passage:

> (…) this definition excludes any one-step mental process as an instance of reasoning. Sudden flashes of insight, if indeed they are instantaneous (…) are thus excluded. So-called gut reactions or evaluations when left unanalysed (…), fail to count as reasoning. Responses not involving the transformation of information (e.g. simple memory retrievals) also do not count, nor does daydreaming or other forms of free association, because these activities lack a goal or focus.
> (ibid. p. 333)

Under the label of *one-step processes,* all kinds of cognitive processes are taken out of the scope of reasoning: sudden instantaneous ideas, gut feelings, memory retrievals, free associations – basically every representation that one is not consciously working towards but which just "falls into one's leap". Later, it should become clear how Haidt defines *intuition* exactly as a one-step process as understood in the quote above. "Steps of reasoning" in his definition serves mainly the purpose of differentiating reasoning from intuition, not the purpose of increasing our understanding of reasoning.

After this very detailed detour about the "stepwise" part of the definition of *moral reasoning*, let me get back again to Haidt's definition: he does not rely on the "stepwise"-characteristic as a single criterion – instead, he adds the condition that the steps undertaken need to be conscious in some way. This part of the definition relies on the work of John A. Bargh which the next section is dedicated to.

2.1.2.2.The concept "conscious" in the SIM

The second property that a mental process ought to display in order to count as an instance of reasoning is being conscious. The term "conscious" is in this case defined as "*intentional*, *effortful*, *controllable* and the person who is reasoning being *aware* of this process". On first sight, these terms do not really appear more illuminating than *conscious*, which is why the next few pages are dedicated to give some short explanations of these terms. With these four criteria for being conscious, Haidt relies on Bargh (1994). This paper however deals with the exact contraries of these properties which are introduced as attributes of automatic processes. The term "conscious" as defined by Haidt can therefore alternatively be spelled out as non-automatic in Bargh's sense. This should not bother us too much, but it should be useful to keep this little twist in mind for the next few pages.

Bargh argues that a process can be considered automatic if it is unintentional, uncontrollable, effortless or the cognizer is not aware of the process. Note that only one of these properties is necessary for a process to count as unconscious. Haidt now turns the whole thing upside-down by making the *conjunction* of the inverses of these properties the condition for a process to be a (conscious) reasoning process. Keep in mind that Haidt's second condition for a process to be a reasoning process was that the steps within are intentional *and* effortful *and* controllable *and* that the person who is reasoning is aware of this process. Having issued this qualification, I now proceed with the single properties of automatic processes as described by Bargh in order to describe Haidt's criteria for non-automatic processes.

Awareness

The most evident criterion for a process to be called *automatic* should be that the person undergoing that process is not fully aware of it. This can happen in several ways: If you are aware that the piece of chocolate in front of you looks appealing, you are not only being accountable for the fact that there is something lying on the table, but also that this something has certain qualities, like being edible and of sweet taste, and why it is that you attribute these properties to it, which is former experiences as well as its sweet smell. If you were not aware of anything lying on the table, you would not be aware of a sweet piece of chocolate lying on the table. If you were not aware of its being actually chocolate, but something nonedible, you would not be aware of a sweet piece of chocolate. And even if you were - if you attributed the sweetness unaware of why you do so, you would not be

fully aware of the attribution of sweetness to the piece of chocolate. Normally we are not aware of the latter two circumstances when attributing sweetness. In these cases, the judgment that a piece of chocolate on the table is sweet is therefore *unware* and therefore *automatic*.

Accordingly, Bargh distinguishes three ways in which a person can be unaware of a mental process:

> "1. A Person may be unaware of the stimulus itself, as in subliminal perception.
>
> 2. A person may be unaware of the way in which that stimulus event is interpreted or categorized, as stereotyping and construct accessibility research have demonstrated.
>
> 3. The Person may be unaware of the determining influences on his or her judgments or subjective feeling states (e.g. the use of felt ease of perceptual categorization or of retrieval from memory as a cue to the validity of perception or the frequency of a stored event) and thus may misattribute the reason to a plausible and salient possible cause of which he or she is aware."
> (Bargh, 1994, p.7)

The most important thing to conclude from this definition is that there are not only different types of unawareness, but that for a person to be aware of a process the person will have to be aware in all three respects: she will need to be aware of the stimulus itself, its categorization and the determinants of this categorization. Only if I know on the basis on which stimulus and due to which of the properties I assign to it and due to which causal reasons I judge something to be the case, I am aware of the process leading up to that judgment. If awareness of any of these components is missing, the process is not aware and according to Haidt and Bargh therefore *not conscious*.

Intentionality and control

The next two criteria for being a conscious mental activity can best be explained with an appeal to an important role that automaticity of actions plays in our daily life: "I am sorry, I did this automatically" is often taken as an excuse for having done in a wrongly manner. The term seems to play this role mainly because automatic actions are beyond the limits of our cognitive control. To say "I did this automatically" means "I couldn't help doing that". There are two situations in which this excuse is valid: if an action was unintended, like punching somebody in the face who was sneaking up behind you yelling "booh!" in order to scare you – and if something is out of your cognitive control and you cannot consciously steer or stop the process, like when you cannot help to find a person smart even

though nothing that person says seems to make any sense[9]. It is this similarity that leads Bargh to take intention and cognitive control to be two criteria of a similar kind:

> "Intentionality has to do with whether one is in control over the instigation or "start up" of processes, whereas controllability has to do with one's ability to stifle or stop a process once started, or at least to override its influence if so desired. To the extent that perceptual, judgemental, and behavioural processes are triggered by the environment and start up without intention, the environment is more in control (…). To the extent that these processes, once started, can be stopped by an act of will, they are controllable by the individual (…)." (Bargh, 1994, p.16)

Both the *intention* condition as well as the *control* condition are describing how "one just cannot help doing something". A process is intentional in the case that the agent is able to willingly decide whether to start it or not – intentionality of a process does not however entail anything about one's capacity to stop the process or to willingly override its consequences. The later capacity is called *cognitive control* and does not entail anything about one's capacity to deliberately start or stop the process in question.

According to the definition above, a conscious process is therefore a process that is started intentionally and can be stopped or at least rendered ineffective by overriding its result at will.

Efficiency and effort

The last condition for an action being conscious is its efficiency or the amount of cognitive effort that this action necessitates. Certain actions can normally be performed without the agent paying attention. A great example for this is driving a car. Under normal conditions, an experienced driver is able to have a conversation with the person in the passenger's seat while driving the car. In tricky situations like when something unexpected is happening on the road or under unusual conditions like in a snow storm, it normally becomes much harder to participate in a conversation. The driver becomes silent and focusses on driving. Whether a car ride is experienced as exhausting or not is normally dependent on how much of the ride was spent in this concentrated fashion and how much driving could be done "on the side". It is this kind of mental effort that makes difficult car rides exhausting that Bargh and Haidt are referring to when seeing cognitive effort as a criterion for whether an act

[9] I know the latter is something you rather do not apologize for, but I like to think of humans as rather regarding people as too smart as as too stupid.

was automatic or not: Bargh is using the term *efficiency* to describe the property of automatic acts that only a minimum of attention and effort are needed to successfully perform a cognitive act:

> "The efficiency aspect of automaticity refers to the extent to which the perceptual or judgmental process demands attentional resources. To the degree that it does, it may not occur when the attentional demands of the situation are high. Such conditions of overload are not unusual."
> (Bargh, 1994, p.24)

The *effort* or *efficiency* condition can be regarded as a condition quite similar to the *control* condition: while the *control* condition is about the capacity to stop a process, the *effort* condition is about the capacity to sustain it. The idea behind it is that non-automatic processes like reasoning will necessitate a certain level of attention, while automatic processes do not – they are automatic and independent of our attentional capacities.

A reasoning process is therefore one that can be influenced or even stopped by diverting the attention of the respective person.

2.1.2.3.Putting the pieces together

Putting the pieces from the previous sections together, I conclude that according to Haidt, reasoning consists of transformation of information that occurs stepwise. The steps are to bear at least some kind of resemblance to logical deduction or induction steps – and at least one of these steps must display the properties that

- one knows on the basis of which premises and what kinds of the properties involved which conclusions are taken for what reason (awareness);
- one is able to make a decision whether to undertake this reasoning step or not (intention);
- one is able to abort this reasoning step or at least to decide not to let one's actions be guided by its result (control);
- one is not able to perform this step if one is not able to display a minimum degree of attention (effort).

Having gone through this lengthy process of understanding what Haidt's idea of moral reasoning is about, we are now able to grasp his idea of intuition much more quickly. Indeed, he describes intuition more or less via the inverse properties of reasoning:

> "The most important distinctions are that intuition occurs quickly, effortlessly, and automatically, such that the outcome but not the process is accessible to consciousness, whereas reasoning occurs more slowly, requires some effort, and involves at least some steps that are accessible to consciousness."
> (Haidt, 2001, p. 818)

Intuition in Haidt's sense is pretty much understood as the inverse of *reasoning*: if a process has no step that displays one of the properties that are typical for non-reasoning processes according to Bargh, we have an intuitive process at hand – we will not be able to control it, or to start it intentionally, or we will not be aware of its premises or causal steps involved, or we will not require any attentional resources to sustain this process.

2.1.3. The central claims of the SIM

Having defined the basic concepts I am now in a position to explain the central claims of the SIM. The easiest way to introduce the SIM is to state what it refutes - very often it helps to understand the point of a model, theory, or argument to know what this position is against. Haidt's model is against what he calls the *rationalist tradition in moral psychology*. This tradition sees moral reasoning as the core mechanism behind moral judgments (Haidt, 2001, pp. 814). The SIM is intended to be an alternative to this rationalist picture of moral judgment.

The SIM postulates a number of relationships between moral judgment, moral intuition and moral reasoning. In "The Emotional Dog and its Rational Tail", Haidt chooses to call these postulated relationships *links* between certain processes (moral reasoning, moral intuition) and their end product (moral judgment). He distinguishes five links which can be illustrated in a graphical way as pictured in figure 2 below. The two most important links are the links number 1 and number 2:

> "1. The intuitive judgment link. The model proposes that moral judgments appear in consciousness automatically and effortlessly as the result of moral intuitions. (…)

2. The post hoc reasoning link. The model proposes that moral reasoning is an effortful process, engaged in after a moral judgment is made, in which a person searches for arguments that will support an already-made judgment. "

(Haidt 2001, p. 818)

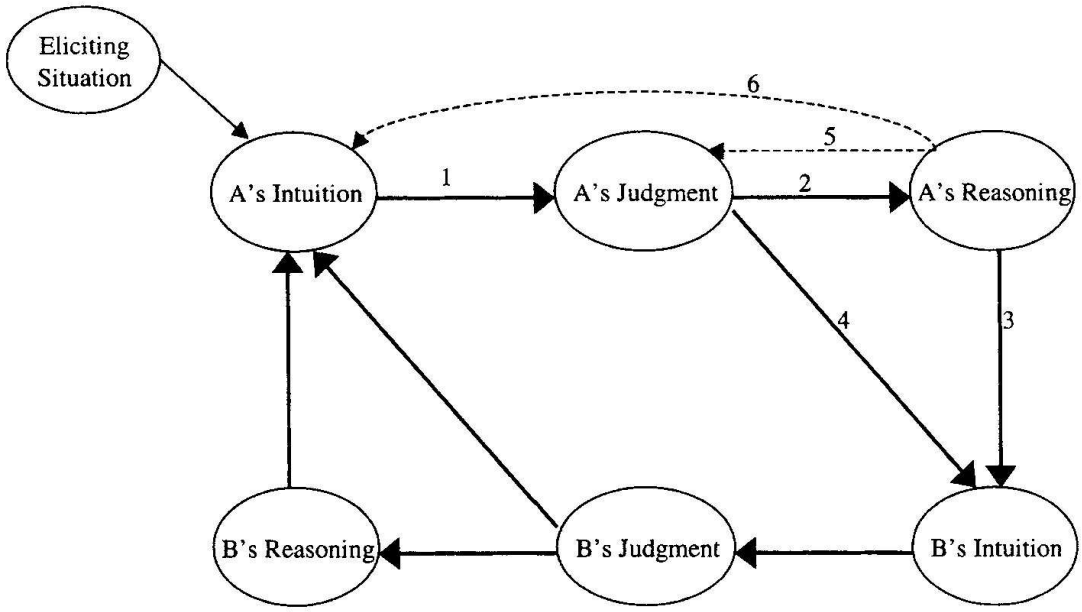


Figure 2. The social intuitionist model of moral judgment. The numbered links, drawn for Person A only, are (1) the intuitive judgment link, (2) the post hoc reasoning link, (3) the reasoned persuasion link, and (4) the social persuasion link. Two additional links are hypothesized to occur less frequently: (5) the reasoned judgment link and (6) the private reflection link.

***Figure 2**: A schematic depiction of the SIM from Haidt (2001)*

The decisive point of the SIM is that moral judgments are (normally) automatically computed and not the result of a conscious deduction process. Conscious moral reasoning exists – but it happens only post hoc, **not affecting the judgment in any way**.

In sum, this is how moral judgment generally works: **An eliciting situation triggers an unconscious cognitive process (moral intuition) that leads to an evaluative propositional attitude (moral judgment) that is later on argued for through conscious inference (moral reasoning).**

If one sees somebody needlessly hurt another person, according to the SIM, *one cannot help to judge* the displayed behaviour as wrong. The mental process that leads to this condemnation involves no steps that are wilfully stoppable (no control), can be wilfully omitted (no intention), necessitate attention (effortless), or that one is aware of causal reasons for one's judgment, its logical premises or

the features attributed to these logical premises (no awareness). The mental processes that involve awareness of premises and causal reasons of results, that at the same time are stoppable, that can be started or omitted at will, and that require some attention and focus set in after the actual moral judgment is already made. As Haidt puts it, reason behaves less like a scientist that *finds out* what is right or wrong, but rather like a lawyer that *defends the position that it is handed* by the intuitive system (Haidt 2001, pp.120).

Note that the model is not intended to exclude the possibility of every now and then there being moral judgments that are derived through a process of moral reasoning or to exclude that our moral reasoning can affect our intuition. It is intended to describe how moral judgment *normally* works. That is why the SIM includes links 5 and 6 that explicitly describe these possibilities. Link 5 accounts for the possibility of there being moral judgments that are caused by moral reasoning: "People may at times reason their way to a judgment by the sheer force of logic, overriding their initial intuition. In such cases reasoning truly is causal, and cannot be said to be the 'slave of the passions.' However such reasoning is hypothesized to be rare, occurring primarily in cases in which the initial intuition is weak and processing capacity is high." (Haidt 2001, p. 819)

Link 6 accounts for the theoretical possibility that even moral intuition can be influenced by moral reasoning: "In the course of thinking about a situation a person may spontaneously activate a new intuition that contradicts the initial intuitive judgment. The most widely discussed method of triggering new intuitions is role taking (Selman, 1971). Simply by putting oneself into the shoes of another person one may instantly feel pain, sympathy, or other vicarious emotional responses." (Haidt 2001, p. 819). Even though he admits that causal effects of moral reasoning on moral judgment are in principle possible, Haidt makes it very clear and explicit that he considers these cases to be very rare and far from being the standard.

There are two further links in the model that describe interpersonal interaction. With morality being a sociological phenomenon by Haidt's definition, society ought to have some involvement into moral judgment. And this is how it works: through communication of judgments and their rational explanation, persons influence each other's moral intuition. This can happen through actual argumentation and therefore reasoning, as described in Link3: "The model proposes that moral reasoning is produced and sent forth verbally in order to justify one's already-made moral judgment to others. Such reasoning can sometimes affect other people, although moral discussions and arguments are notorious for the rarity with which persuasion takes place". Note that this kind of causal effect of moral reasoning is considered by Haidt to be much more effective and common than intrapersonal moral reasoning. But also in interpersonal links, the importance of reasoning stays rather limited: the causal effect of the mere expression of the moral judgment without giving a clear cut reason (Link 4) for it is

regarded by Haidt as at least just as efficient: "Because people are highly attuned to the emergence of group norms, the model proposes that the mere fact that friends, allies, and acquaintances have made a moral judgment exerts a direct influence on others, even if no reasoned persuasion is used."

These links describe how witnessing people from your group making moral judgments will influence you in your own moral judgment. Take for example many people's first experiences with alcohol: witnessing your peers consuming it kind of softens the imperative not to consume alcohol as a minor until you do not feel it to be a real issue anymore. Similar processes concerning peer-group morality can be witnessed in vegan circles or among consumers of outlawed drugs. This would be a paradigmatic example of the link number 4 in which reasoning does not really play a role. Reasoning does however play a role in many situations in which a deeper understanding of the situation can lead to different judgments, like when someone is telling you to not cross red traffic lights while a child is watching, as that would be a bad example to the child, potentially causing it to do something stupid one day or another. This is a classical consequentialist argument regarding a certain kind of behaviour, and it often leads people to change their behaviour accordingly. This would be a paradigmatic example for link number 3.

2.1.4. Conclusion

With all this in our hands, we are now able to sum up the most important points of the SIM:

Moral judgments are defined as evaluations (good vs. bad) of the actions or character of a person that are made with respect to a set of virtues held to be obligatory by a culture or subculture, the latter meaning that a) the respective virtue is expected to be displayed by all people or a certain group of people and b) non-accordance to the norms pertaining to that virtue is generally sanctioned.

Moral reasoning is defined as stepwise transformation of given information about people and situations in order to reach a *moral judgment, which includes steps that are a) directed towards a certain aim and that are b) not incompatible with systems of logic leading to a result compatible with the premises* c) *with at least one of the steps being* conscious, which means intentional (the cognizer is able to start the process at will), controllable (the cognizer is able to stop the process or override its effects), effortful (a minimum of attentional resources will be required for the process to be entertained), and that the cognizer is aware of the premises, the way these premises are categorized and

of the causal chain leading from the premises to the result. Note that for a cognitive process to be conscious it is a conjunction of necessary properties that has to be displayed.

Moral intuition is defined as the sudden appearance in consciousness of a *moral judgment*, including an affective valence (good-bad, like-dislike) without any conscious awareness of having gone through steps of searching, weighing evidence, or inferring a conclusion. Not that in order to be unconscious, it is enough for a cognitive process to lack any of the properties that are necessary for a process to be conscious.

Based on these conceptual considerations, the main points of the SIM are that Moral judgments appear in consciousness automatically and effortlessly *as the result of moral intuitions* and that moral reasoning is an effortful process, engaged in after a moral judgment is made, in which a person searches for arguments that will *support an already-made judgment*.

The SIM furthermore accounts for interpersonal relations between moral judgments: arguing for a certain moral judgment or just enforcing it in conversation can affect the moral intuition of another person. What is considered right and what is considered wrong in a given society is therefore shaped by social interaction to at least some extent.

The way these social interactions shape moral judgment is the topic of the Moral Foundation Theory (MFT), the second of the two models about morality that I consider in this thesis.

The MFT tells us about what kinds of transgressions moral judgments actually are – something that the SIM tells us very little about. The MFT furthermore sheds some additional light upon the notion of "virtues held obligatory", that I have found to play such a vital role in the SIM in section 2.1.1.

Both models, MFT and SIM, can be seen as separate models designed to explain separate aspects of morality, but evidence for the one can often be considered evidence for the other, as there clearly is a relation of mutual support between both theories.

2.2. The Moral Foundations Theory (MFT)

There was a myth at my school about a student who had answered to the test question "Where is vine cultivated?" with an enumeration of the areas where vine is cultivated: "Italy, Germany, France, U.S., Australia, etc.": He was said to have received zero points for this answer, as the requested an-

swer was “on south-eastern slopes of hills as well as in valley areas if a warm and sunny climate allows for it.”. Similar to the question where vine is cultivated, the question what moral judgment is allows for different kinds of answer: one that accounts for what kind of process leads up to a moral judgment – and one that accounts for the types of situations in which moral judgment is bound to occur. The SIM was an answer of the first kind – the Moral Foundations Theory (MFT) is one of the second.

As we have seen before, the term *moral* is defined rather loosely in the context of the SIM: Even though in the context of the SIM importance within a culture plays a vital role for what is considered moral, very little is said about *what it actually is* that is held important enough to count as moral. The Moral Foundations Theory is an empirical theory based on the SIM’s definition of morality that sharpens our idea of what moral judgments in Haidt’s sense actually are by applying the definition from the SIM to empirical anthropological research. The MFT uses anthropological data to answer the questions: Which situations are *moral situations*? Is the domain of moral judgments, the *moral domain*, one where – in principle - “anything goes”? Or are there certain kinds of situations in which a moral judgment is *bound* to occur? And how does morality get *into* the guts from where it is then supposed to come out of in the form of automatic, intuitive moral judgments?

As an answer to this question, the MFT postulates the existence of differentiable innate capacities to form automatic reaction patterns to certain social stimuli. These reaction patterns are what the SIM calls *moral intuitions*. The innate capacities that allow for moral intuition to develop are tied to different kinds of social situations and can be nurtured or suppressed by society, leading to cultural variations in what is considered moral and what is not. What is meant by *moral judgment* in the context of this theory, the nature of the mentioned capacities and social situations, as well as their role in the mentioned theory is discussed in the following section.

2.2.1. The concept of *moral* within the framework of the MFT

As I already mentioned above it is often the case that theorizers, when sketching a theory, have another theory in mind that they regard as blatantly wrong. Very often they explicitly refer to this theory when explaining their own position and they argue just as explicitly why their own theory is better than the one they want to “shoot down”. Considering “against whom” a certain theory is formulated therefore very often helps to understand the very point of that theory. The position that Haidt keeps

referring to as his personal punching bag when defending the MFT is the understanding of *moral judgment* formulated by Elliot Turiel. The following passage about "the enemy" is therefore very illustrative.

> (…) Elliot Turiel, a former student of Kohlberg and a major figure in moral psychology, codified this individual-centered view of morality in his influential definition of the moral domain as
>
> > prescriptive judgments of justice, rights, and welfare pertaining to how people ought to relate to each other. Moral prescriptions are not relative to the social context, nor are they defined by it. Correspondingly, children's moral judgments are not derived directly from social institutional systems but from features inherent to social relationships -- including experiences involving harm to persons, violations of rights, and conflicts of competing claims. (Turiel, 1983, p.3)
>
> Turiel's delimiting of the moral domain seems obviously valid to many people in modern Western cultures. However, for people in more traditional cultures, the definition does not capture all that they see as falling within the moral domain. In other words, Turiel's definition (we are asserting) is inadequate as an inductive generalization. It is a stipulative definition which does not match the empirical facts. When the moral domain is defined as "justice, rights, and welfare," then the psychology that emerges cannot be a true psychology of morality; it can only be a psychology of judgments about justice, rights, and welfare. And when the domain of morality is narrowed in this way, then overly parsimonious theories of moral psychology flourish.
> (Haidt & Joseph, 2007, pp. 370)

This piece of text gives several insights on what the Moral Foundation Theory says and what it explicitly refutes. Let me sum up Turiel's idea of moral judgments:

- They embody prescriptive judgments of justice, rights, and welfare.
- They are not relative or defined by social context.
- They relate to features inherent to social relationships.

What is meant with these terms? In order to ease the understanding of this approach, let me add an additional puzzle piece: the kind of judgment that Turiel is contrasting moral judgment with is judgment of *social convention*. A paradigmatic case to illustrate the difference that Turiel has in mind would be the difference of the situations of a boy wearing a skirt to school (transgression of a conventional norm), and one kid hitting another kid (transgression of a moral norm) (Turiel, 1983). The distinguishing criterion of morality in value judgments is in Turiel's case its independence of cultural or societal norms. After all, morality is generally considered to be something that everybody is bound to. Haidt rejects this universalistic idea of morality – even though universality is also a criterion which Haidt considers necessary for a judgment to be a moral judgment. The difference between the two

kinds of universality is that while Haidt's universality is a subjective universality, a universality attributed by the judging, in Turiel's case it's an objective universality, a universality attributed by the onlooker. Haidt argues that the assumption of this objective universality contradicts empirical evidence: in many non-western cultures, people typically attribute "moral" properties to certain "conventional" rules, taking these rules just as seriously as norms of justice, welfare and harm. Especially "traditional" cultures tend to have a much broader spectrum of morality, including lots of sexual or religious taboos. The moralizing stance that non-westerners take on these subjects has been demonstrated among others by Paul Rozin, who formulated a predecessor of the MFT in collaboration with Haidt (Rozin et al., 1999). Haidt regards Turiel's formulation as eurocentristic and ignorant of non-harmbased moralities that can be found for example in India (Shweder et al, 1997) or parts of Brazil (Haidt, Koller & Dias, 1993) and discards Turiel's formulation therefore as empirically invalid.

In his criticism of Turiel's definition, Haidt adds a remark that is particularly interesting for this work: If our understanding of morality were a different one, the story we would tell about its psychology would have to become a completely different one, too. It is a nice example for how much diversity there is in the psychology and neuroscience of morality that Turiel's concept of *morality* that Haidt outlines as a "beware"-scenario is a fairly good description of the approach adopted by Patricia Churchland in her book *Braintrust* (Churchland, 2011) that has been published several years later – a neuropsychological approach to explain morality understood as "caring for others".

The core idea behind Haidt's criticism of Turiel should become clearer now. If we want to understand what *moral judgment* means, we have to look out for what people actually regard as moral judgment. But how could this work? Here an outline:

> The first step in mapping the moral domain of any culture, we believe, should therefore be to list and count the norms that get the most attention. What norms and norm violations do people gossip about? What norms are broken and punished in myths and folk tales? When people reject or criticize other members of their community, or when they express shock at the practices of another community, which norms are involved?
> (Haidt & Joseph 2007, p. 372)

The relative importance that people attribute to norms is much more important for Haidt than their intercultural validity or even conceptual considerations about the "nature" of moral judgment. The moral domain is deliberately held very variable on the conceptual side and determined to be something that can only be assessed empirically.

Let me sum up the most important points so far: according to the SIM, moral judgments are intuitive judgments. Furthermore, the SIM claims that moral intuition is influenced by interpersonal relation-

ships and society. We were confronted with Haidt's approach to morality from the MFT, where moral judgments are defined as judgments concerning norms that are considered the most important by a certain culture. If one connects both stories with each other, one sees that intuitions of people within a society will stabilize each other on the long run and establish a certain set of norms that people will share joint intuitions about.

With this, I have laid down the conceptual foundations necessary for a thorough understanding of the MFT. I am now ready to get to its actual claims.

2.2.2. The origin of moral intuition

In the beginning of this chapter it has already been mentioned that the MFT would be about certain innate dispositions to develop moral intuitions. This kind of claim can come in various strengths: One extreme would be that moral intuitions are innate in a very detailed fashion, with for example "abortion is wrong" being hardwired into the human mind by evolution (Tooby & Cosmides, 1994) or god. The other extreme would be that moral intuition can only develop to be about certain situations, but is quite adaptable within this range. The later one is the one that Haidt would lean towards: According to the MFT, there is a set of hardwired innate capacities to form judgments of special rigor and seriousness about certain kinds of behaviour much easier than about other kinds of behaviour – just like it is easier to develop a fear of spiders and snakes than fear of cars or stock markets crashes. Exactly which kind of social situation will in the end actually trigger one of these judgments is however not pre-set by nature but culture dependent:

> (…) our more complex abilities are often better described as a 'preparedness' to learn something. For example, humans are born with few hardwired fears, but we come prepared to acquire certain fears easily (e.g., of snakes, spiders, mice, open spaces), and cultures vary in the degree to which they reinforce or oppose such fears. On the other hand, it is very difficult to create a fear of flowers, or even of such dangerous things as knives and fire, because evolution did not 'prepare' our minds to learn such associations.
> (Haidt & Joseph 2004, p. 58)

So this is how moral intuition develops: there are certain dispositions to automatically form evaluative judgments in given social situations. If these judgments are not enforced by society, they will not

develop and “remain silent”, if they are enforced, they will develop to mental mechanisms that implement moral intuition. This is the basic claim of the MFT:

> All we insist upon is that the moral mind is partially structured in advance of experience so that five (or more) classes of social concerns are likely to become moralized during development. Social issues that cannot be related to one of the foundations are much harder to teach, or to inspire people to care about.
> (Haidt & Joseph, 2007, p. 381)

In summary, our morality is derived from a set of innate dispositions to show affective reactions towards certain kinds of stimuli. These dispositions can be nurtured and developed but also be silenced and rendered ineffective by the cultural environment in which we grow up. This nurturing and silencing is already part of the SIM and described by the links 3 and 4 (chapter 2.1).

This adaptive picture of morality accounts for example for the extreme cultural differences when it comes to moral judgments like those concerning sexual taboos: while there is a universal tendency to moralize norms about sexuality, the exact nature of these norms remains extremely culture dependent. There are cultures in which polygamy is wrong and cultures in which it is not, there are cultures in which adultery is intrinsically wrong and those in which it is not, there are cultures in which homosexual relationships are wrong and those in which they are not, there are varying definitions of what counts as incest, what counts as a homosexual act, etc.

Another part of the passage cited above relates directly to the question that I want to deal with next: the passage speaks of “five (or more) classes of social concerns”. These classes of social concerns are in a certain way the answer to the second question asked before, namely the question of what kinds of situations moral judgment can possibly be about.

2.2.3. The different domains

So there is only one part missing in the story of how our moral judgment works, which is the situations that are typically evoking moral judgments. According to Haidt (2004), humans have an innate disposition to make evaluative judgments concerning situations which display suffering, hierarchy, reciprocity, or purity. The number of the moral domains in Haidt’s moral theory varies between publications, due to classificatory difficulties. This is the reason why the quote from last page mentioned five moral domains, not four. In this particular publication (Haidt, 2007), the domains were

Harm/Care, Fairness/Reciprocity, Ingroup/ Loyalty, Authority/ Respect, and Purity/ Sanctity. The difference to the four domains is that a distinction between two subdomains of the "hierarchy" domain is introduced: Ingroup/Loyalty describes a hierarchy between members of one's social environment and more distant persons, while Authority/Respect describes hierarchy *within* one's cultural environment.

These situations can be associated with certain emotions that people tend to display when these situations occur. These emotions are normally associated with a clear evaluation of the eliciting situation. They are also normally associated with different but to some respect very similar kinds of situations and Haidt hypothesizes that originally (evolutionarily speaking) the domain of these emotions was not necessarily a moral one – that they however became culturally associated with "new" situations and that often the "moral" function of the emotion was taken over later. Two example domains with their respective emotions should help to illustrate this point.

- The *Suffering Domain* is associated by Haidt with the virtues of kindness and compassion and the emotion of compassion. The original domain of the emotion is the suffering and the vulnerability of one's offspring. We can very easily pinpoint an evolutionary benefit to this social emotion. However, through cultural nourishment of the emotional judgments involved, the domain of compassion came to include also baby seals and cartoon characters, group members who are not related to oneself, even strangers. Compassion broadened its domain. Societies *learnt* to feel compassion for non-related persons, foreigners, animals. Patricia Churchland (2011) offers some fascinating empirical data about this process by taking a closer look at research about the evolutionary development of compassion in mammals and especially humans and the correlating change in brain systems employing the neurotransmitters vasopressin and oxytocin.

- The *Purity Domain* is associated with virtues like (spiritual) cleanliness, (spiritual) purity and chastity, and consequently often the adherence to the "right" religious commandments - taking off one's shoes at the entrance of a Hindu temple, ritual washing in front of a mosque, dipping one's fingers in holy water and making a cross gesture when entering a catholic church. The emotion associated with transgression of rules pertaining to this domain is disgust, the original domain of which is taken to be waste products, as well as people and animals with diseases or parasites. The evolutionary benefit from feeling disgust in these situations entailing a clear avoidance bias is the reduced risk of contamination with potentially life threatening germs. Haidt (or rather Paul Rozin (Rozin et al., 2009), with whom he had developed the mentionned forerunner of the MFT (Rozin et al., 1999)) theorizes that disgust be-

came a social emotion through association with taboo ideas, religiously or sexually deviant people, or "traitors" in general. Societies learnt to find sexually or religiously deviant behaviour disgusting (or at least tasteless).

So Haidt's answer to the question in which situations moral judgments are formed consists of two parts: on the one hand, moral judgments can in principle occur in situations that display suffering, reciprocity, purity, or hierarchy. On the other hand, whether and how moral intuition about this domain is actually formed depends on whether and how emotional reactions are culturally nurtured.

Now that I have mentioned emotional reactions in the moral domain and moral intuition in the same sentence, I cannot hide from the question that the keen observer has probably already in her mind: how exactly does the relationship between emotion and intuition work? In the following chapter, I attempt to give some answers.

Four moral modules and the emotions and virtues associated with them

	Suffering	Hierarchy	Reciprocity	Purity
Proper domain (original triggers)	*Suffering and vulnerability of one's children*	*Physical size and strength, domination, and protection*	*Cheating vs. cooperation in joint ventures, food sharing*	*People with diseases or parasites, waste products*
Actual domain (modern examples)	*Baby seals, cartoon characters*	*Bosses, gods*	*Marital fidelity, broken vending machines*	*Taboo ideas (communism, racism)*
Characteristic emotions	*Compassion*	*Resentment vs. respect/awe*	*Anger/guilt vs. gratitude*	*Disgust*
Relevant virtues	*Kindness, compassion*	*Obedience, deference, loyalty*	*Fairness, justice, trustworthiness*	*Cleanliness, purity, chastity*

Figure 3: *Table mapping moral domains and moral emotions from Haidt & Joseph (2004)*

2.2.4. Conclusion

In conclusion, it can be said that the MFT fleshes out many aspect of morality left vague or open by the SIM. The scope of morality was introduced as culture dependent, in contrast to the objectivist idea of morality defended by rationalist theorists. This makes the scope of morality an empirical rather than a conceptual question, leading the way for an empirical theory of morality understood as "the set of norms estimated most important by a certain culture or subculture": the MFT.

The Moral Foundations Theory suggests that humans are predisposed to form moral judgments only in regard to behaviour associated with a certain class of social concerns: the moral domains of suffering, hierarchy, purity and reciprocity. The predisposition to moralize judgments related to these concerns is instantiated through an innate tendency to show certain emotions towards transgressions within these domains. This tendency can be enforced or suppressed through social feedback, as described in the SIM links 3 and 4 (see section 2.1). The purity domain for example that concerns transgressions of norms associated with (spiritual) purity is associated with the emotion of disgust.

We can see that while the SIM explains what kinds of mental processes are involved in moral judgments and how these processes affect each other intra- and interpersonally, the MFT introduces further empirical findings and provides us with an idea what moral judgments are about and why. In the MFT, *emotions* play a central role, while the SIM focusses on the role of *intuition* in moral judgment. I already gave a detailed account of what is to count as an *intuition* (see section 2.1). I did however not yet give an account of how the term *emotion* is understood by Haidt.

So what is an emotion? What is the relationship between emotion and intuition? These are the last two issues which have to be addressed to fully grasp both models and their relationship to each other. I address them in this upcoming short section.

2.2.5. The concept of *emotion* in MFT and SIM

So there is a last issue that one faces when bridging the gap between MFT and SIM: while the MFT is about moral emotions that mediate moral judgments, the SIM is about moral intuition. In order to get the relationship between both models right, one has to come to understand how the relationship between *emotion* and *intuition* works.

The result of this section is that in Haidt's terminology, emotions can be regarded as a kind of intuition, and can be used almost synonymously in moral contexts. The question how an emotion can change its domain as postulated by the MFT is also addressed.

2.2.6. What is an emotion? – Damasio's Somatic Marker Hypothesis

Like in the case of moral reasoning and moral intuition, it should be helpful to have a closer look at one of the authors that Haidt refers to when arguing about the role of emotions. In his explanation of the role of emotions in the SIM (Haidt, 2001, p.824), Haidt refers to Antonio Damasio's Somatic Marker Hypothesis - a neuropsychological approach to the nature of emotions. This hypothesis about emotion is immensely helpful for understanding the role of emotion in the MFT and the SIM.

The basic idea behind the Somatic Marker Hypothesis is that emotions are representations of somatic (a.k.a. bodily) states (like nausea, stress, aggressiveness) associated with (and also triggered by) representations of certain entities or matters of fact about the world (like a mouldy sandwich, a dangerous situation, a person insulting you). These somatic states normally cause or imply a certain action and/or attention bias (like reluctance to touch or go near the object, increased alertness, increased readiness for offensive/defensive action). Our action guiding somatic reaction *marks* the originally action-neutral trigger object or situation with a certain "feel" and more importantly, with a bias to act in a certain way. The somatic marker can be understood as the representation of the "digustingness" of rotten food, as the "alertingness" of a dangerous situation, as the "provocativeness" of another person. Emotions are therefore considered as a *coupling or association of two different kinds of representation.*

Consequently, emotions have a *cognitive* component (*recognizing* the stimulus) that is coupled with a somatosensory component (the entailed bodily changes). The evolutionary benefit lies within the action and attention guiding nature of emotion: Once a certain representation (let's say an attacking elephant) is evoked, a representation of a certain somatic state (stress, excitement, tension, in one word: fear) is triggered. This representation then guides our actions ("Get ready to run!") and our attention ("Where can I hide? How close is the elephant? How much time until it reaches me?"). This automatic bias towards certain actions reduces reaction time in critical situations by cutting off big branches from the decision tree (branches like "Wouldn't a cup of tea be a lovely idea right now?") and ensures that especially important information (which is information that coincides with remem-

bered changes of our somatic state from the past, say for example nausea after eating old bread with a fluffy green coat) is action guiding (Damasio, 1996, pp. 1414). Forming somatic markers therefore leads to significant evolutionary and everyday advantages to be found in reaction time and reduction of cognitive load: immediate readiness for action and focussing of attention as a reaction to an unforeseen marked event, as well as the capacity to handle huge loads of new information in a very efficient way are vital features for survival in nature as well as in human cultural environment. A somatic marker will restrict the alternatives for action in an automatic fashion that speeds up and facilitates decision making considerably.

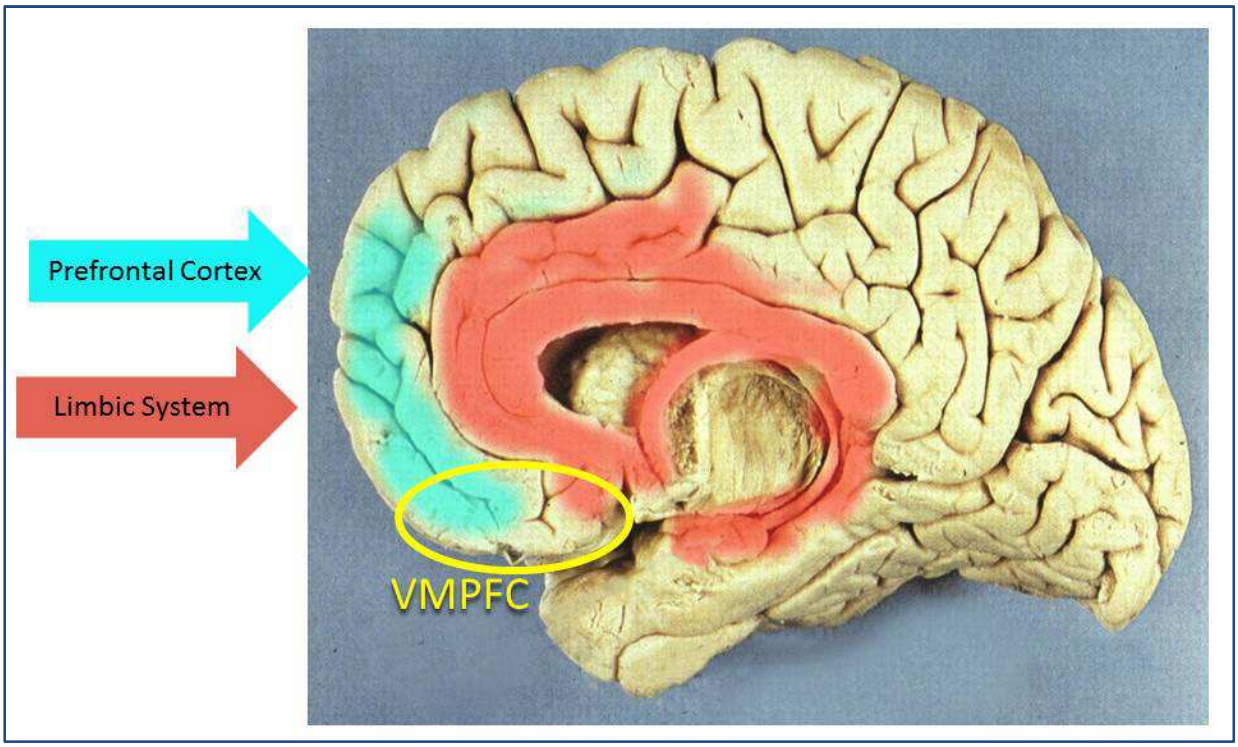


Figure 4: *The ventromedial cortex (VMPFC) is the anatomic connection between the frontal lobe (blue, associated with representations of external events) and the limbic system (red, associated with representations of body states). Picture based on picture from http://homepage.smc.edu/wissmann_paul/physnet/anatomynet/anatomy/image33.jpg*

The Somatic Marker Hypothesis gets its empirical support mainly due to the fact that there is only one area in the brain where there could plausibly be a neuronal connection between somatosensory areas (limbic system and attached parts of the somatosensory cortex) and areas where representations about external situations take place (frontal and prefrontal cortex): the ventromedial ("the middle of the downside") part of the prefrontal cortex, that forms the only direct physiological connection between these areas (see figure 4). People who lack this structure (for example because it

had to be removed surgically due to a cancer or epilepsy condition) display certain kinds of cognitive malfunctioning in regard to emotion that equip the Somatic Marker Hypothesis with substantial neuropsychological support (Damasio, 1996, pp. 1417).

2.2.7. The concepts of *emotion* and *intuition*

In order to understand the relationship between *emotion* and *intuition* in Haidt's theory, it is useful to understand how he connects the importance of emotions in the MFT with the importance of intuition in the SIM. He does this in the course of presenting the empirical foundations[10] of the SIM by matching his intuitive-rational distinction with the idea of there being an emotional and a rational system:

> Metcalfe and Mischel (1999) proposed a dual process model of willpower in which two separate but interacting systems govern human behaviour in the face of temptation. The "hot" system is specialized for quick emotional processing, and it makes heavy use of amygdala-based memory. The "cool" system is specialized for complex spatiotemporal and episodic representation and thought. It relies on hippocampal memory and frontal lobe planning and inhibition areas. It can block the impulses of the hot system, but it develops later in life, making childhood and adolescence seem like a long struggle to overcome impulsiveness and gain self-control.
> (Haidt, 2001, p. 823)

The parallel between this "hot system – cold system" distinction and the "intuition – reason" distinction is already figuring and becomes truly evident once we have a look at the descriptions that Metcalfe and Mischel give of the systems: "The cool system is narrative, weaving knowledge about sensations and emotions, thought, actions, and context into an ongoing narrative that is coherent, goal-sensitive, and strategic." (Metcalfe & MIschel, 1999, p.6) contains two of the reasoning typical features given in Galotti's (1989) definition: a transformation of given information in a goal-oriented and coherent way. The characterization of "the hot emotional system" (p4) leads to similar conclusions, it is described as "largely under 'stimulus control', characterized by rapid automatic triggering, conditioned responding, inflexibility, stereotyping and affective primacy"(p.6) . The intention and control condition of intuitiveness are already contained in the description, which also calls the pro-

[10] Reasoning-models of moral judgment have difficulties to explain how moral action comes about. (Haidt, 2001, pp. 823)

cess “automatic”, a term used synonymously with *intuitive* by Haidt. The description of the “hot System” as “under stimulus control”, combined with the learnability and automaticity of this response make it a perfect match with the SMH and therefore with Haidt’s corresponding concept of emotion.

However, even though Haidt ties intuition and emotion very closely together (on many occasions he uses the concepts more or less like synonyms, for example in the title of his most influential paper “The emotional Dog and its Rational Tail”) he keeps on distinguishing both concepts. So, even though moral emotions and moral intuitions are extremely close that they almost become indistinguishable they are not the same. So, what is the difference? Looking at the definitions of emotion and intuition, we can spot the difference that intuition is not necessarily action or attention driving while emotions are always connected with a tendency to show a certain kind of behaviour (that can of course be overridden) (Haidt, 2003). This is however only a minor difference – one of the reasons why Haidt uses the concepts almost interchangeably.

The closer look at the Somatic Marker Hypothesis allows also for another explanation not yet given on behalf of the MFT: the MFT postulates the change of domains of certain emotions over time. With the SMH, one is able to explain how this can work: as emotions are basically learnt associations, associating a situation A with a situation B that is somatically marked will over time lead to an extension of the somatic marker to situation A – extension of the domain of an emotion becomes a matter of classical conditioning.

2.2.8. Conclusion

So what can we say about emotion in the MFT and SIM? Mainly that **emotions** are a) pairings of representations of somatic states and other entities and b) action and attention guiding through activation of these somatic states (like for example nausea).

Haidt uses the terms *emotion* and *intuition* similarly, almost interchangeably, which can be attributed to the automatic character of emotions. This helps to bridge the conceptual gap between the SIM and the MFT: *moral emotions* are *moral intuitions* and *moral intuitions* (normally) are *moral emotions*. The MFT can therefore be seen as a more fleshed out account of how the processes postulated by the SIM work – the SIM can be regarded as a conceptual foundation of the MFT. Both models give each other mutual support and together they form a much stronger, more holistic picture of morality

than each of them on its own. I chose to introduce both models instead of only one of them exactly for this reason – together they form a joint theory of morality that has the explanatory power required to draw strong philosophical conclusions.

Having elucidated this part of Haidt's approach I turn now to presenting one selected study that supports both models. It is only one of the many empirical studies that support the MFT and the SIM. This is why I hereby point out once again that the study I present next, just like the MFT and the SIM, is just an exemplification chosen for illustrative purposes. I do not intend to attack or support Haidt's moral theory.

2.3. An exemplary study: why it is wrong to eat your dog

In the German movie „Wer früher stirbt ist länger tot", a coming of age story set in rural Bavaria, a little boy that is obsessed with the afterlife asks the priest of the village what he can do to get to heaven. The priest answers "Very simple: believe and act according to the messages of Jesus Christ". The boy gives him back a puzzled look and answers timidly: "…and what does this mean precisely?".

This little episode shows very explicitly a common problem with moral judgment. It is very easy to say what moral judgment is, as long as one is not confronted with manifest situations. Difficulties arise once one is asked to lay down concretely what moral judgment is. If one is to measure moral judgment, one will have to do so. And this, the operationalization of moral judgment, is where I expect to find the most implicit philosophical commitments – here it is impossible to crouch behind the vagueness of our common language term *moral judgment*, as one is simply forced to present stimuli and ask for judgment. After going through basic concepts and statements of an exemplary psychological theory of moral judgment, I will therefore now introduce a paradigmatic study which is referred to as a validation of the exact same theory. It is a rather old and by now already "classical" study conducted by Haidt et. al. in the early nineties (Haidt et al., 1993). [11]

[11] Indeed, this is not the kind of validation that one has normally in mind: normally one would expect that predictions are made on the basis of a theory that are then validated by an experiment or study. In fact, the predictions made were derived from *precursors* of the MFT and the SIM, not from SIM and MFT themselves. Here, I treat the study as evidence though, as the predictions made are actually direct consequences of the theory. In Chapter 2.4.2 I will deduce the predictions made for the study from the theory presented above.

2.3.1. Method

The idea of the study described in the paper with the brilliant name „Affect, Culture, and Morality, or Is It Wrong to Eat Your Dog“ was to confront people from different cultural environments, of different socioeconomic standard (SES), and of different age with moral and nonmoral situations that would elicit emotional reactions in the participants. The participants were to judge the situations in a number of ways. The results were supposed to detect correlations between the three mentioned variables and variations in moral judgment and related psychological mechanisms. For keeping things short, I focus only on predictions about the study results that can be derived from the MFT and the SIM. Therefore, I ignore the variable *age* and predictions made independently from the MFT and the SIM.

Controlled variables

Westernization

The study participants were recruited in three different cities that were supposed to display a different degree of “westernization” understood as the “degree to which each of three cities has a cultural and symbolic life based on European traditions, including a democratic political structure and an industrialized economy“(Haidt et al., 1993, p. 615)[12]. The selected cities were Philadelphia (high westernization) in the U.S., Porto Alegre (medium westernization) in southern Brazil, and Recife (low westernization) in north-eastern Brazil. While Porto Alegre was significantly above national average on most indicators of industrial development at the time[13], Recife was significantly below. (ibid., p. 616)

SES

The third controlled variable (as I mentioned earlier, *age* was controlled as a second variable but does not figure in this investigation) was the socioeconomic standard of the participants. This variable was introduced to prevent the influence of SES to be confounded with the effects of westernization. For each city and age group, a high SES and a low SES group were tested. The different groups were recruited at different locations that indicated their SES - public schools vs. expensive private

[12] What is understood as “western” in the context of this work is solely based on Haidt’s use of this concept. Personally, I regard the concept of “western” culture, “western” values and “western” traditions as quite problematic for a variety of reasons.

[13] “e.g. economic activity, income, health, education, and suicide” (Haidt et al., 1993, p.616)

schools, College vs. MacDonald's restaurant in an economically weak neighbourhood, direct approach to university students vs. manual labourers or evening schoolers. (ibid, pp. 616)

Stimuli

The procedure was a structured interview. Participants were presented with different situations displaying persons showing different kinds of behavior:

- Some involved harmful behavior, some involved harmless norm violations.
- Among the harmless violations there were some that were displaying "unconventional food and sexual practices" and some that were not. According to the predecessor of the MFT at the time[14], these acts were to be perceived as disgusting.
- Among those harmless stories that were not intended to invoke disgust, there were two scenarios that involved disrespect or disobedience.

Note the close match between the moral domains of harm, purity and authority postulated above and the test stimuli. Examples can be found in figure 5 below.

[14] The already hinted-at CAD-Triad-Hypothesis (Rozin et. al., 1999)

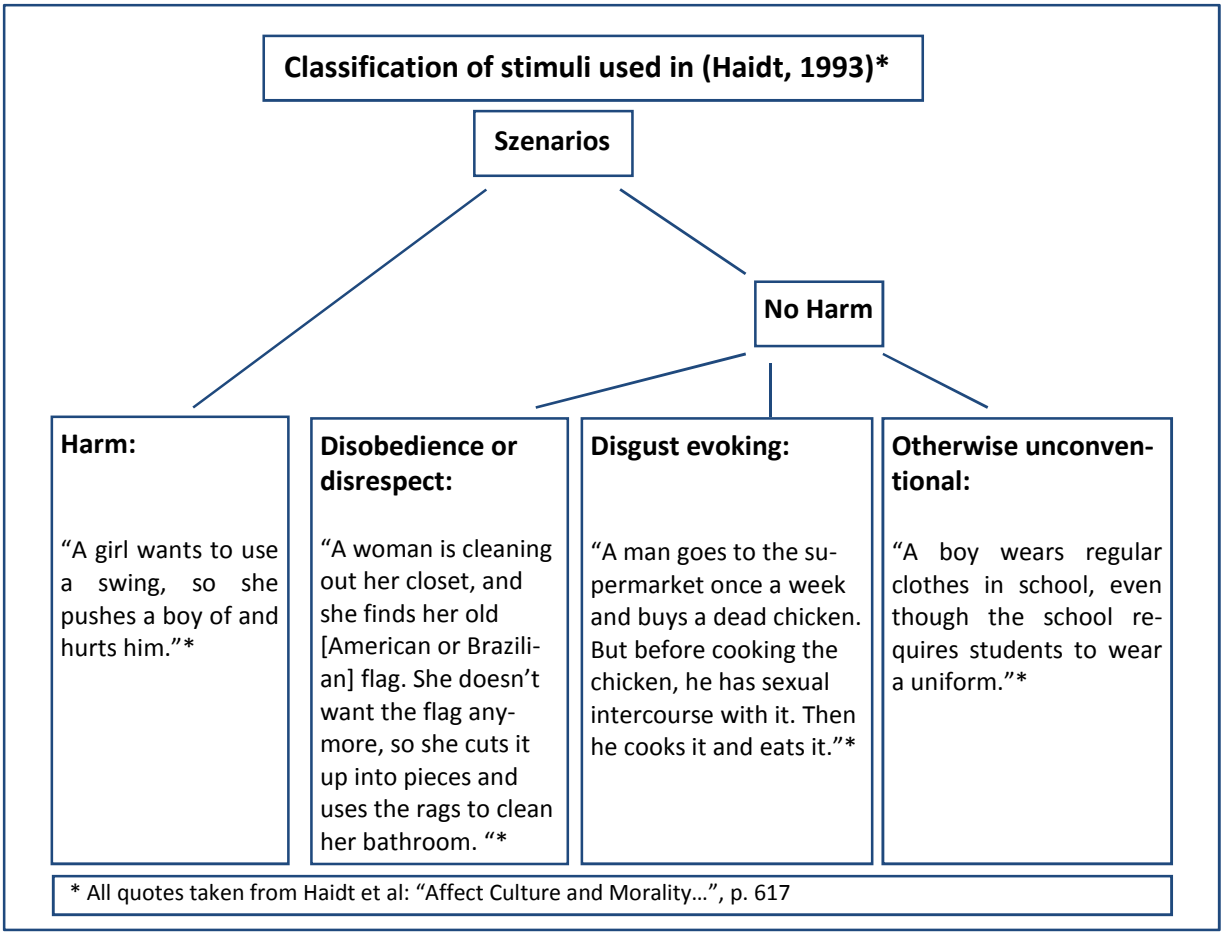


Figure 5: *Classification of stimuli used in Haidt (1993)*

Questions

After presenting each stimulus, the following questions were asked:

- (a) *Evaluation*: "What do you think about this? Is it very wrong, a little wrong, or is it perfectly OK to [act specified]?"
- (b) *Justification*: "Can you tell me why?"
- (c) *Harm*: "Is anyone hurt by what [the actor] did? Who? How?"
- (d) *Bother*: "Imagine that you actually saw someone [performing the act]. Would it bother you, or would you not care?"
- (e) *Interference*: "Should the actor be stopped or punished in any way?"
- (f) *Universal*: "Suppose you learn about two different foreign countries. In country A, people [do that act] very often, and in country B, they never [do that act]. Are both of these customs OK, or is one of them bad or wrong?"

  (Haidt et al., 1993)

The questions that determine whether the judgment given in *Evaluation* is a moral judgment are the *Interference* and the *Universal* question.

- *Interference* "establishes whether the action is seen as the actor's own business or whether outside interference would be legitimate and appropriate." (ibid. p. 617)
- *Universal* "establishes whether the action is seen as universally wrong, regardless of local customs and consensus, or whether it is seen as a social convention that can be different in different places." (ibid., p. 617)

The questions *Harm* and *Bother* are introduced to test the predictive power of harm vs. affective valence concerning moral judgment.

The open question *Justification* is asked to categorize the kind of morality that is at play: The answers are categorized into the three domains of *autonomy*, *community*, *divinity,* three domains postulated by the CAD-Triad-hypothesis, a predecessor of the MFT. In the MFT, these domains are the domains of harm, authority and purity. Answers that did not refer to any of the domains but just affirmed the norm that was transgressed (e.g. "You're not supposed to have sex with a chicken!") were coded as norm statement; answers that were not sortable into these four categories were scored as un-codable.

2.3.2. Predictions

For this study, Haidt and colleagues make the predictions, that

(1) "A majority of the high SES Philadelphia subjects will take a permissive stance, because this group has a harm based morality.
(2) There will be a main effect of city, or westernization, such that the harmless-offensive stories will be moralized most in Recife and least in Philadelphia.
(3) There will be a main effect of SES, such that within each city the harmless-offensive stories will be moralized more by low-SES subjects than by high-SES subjects.
(4) A majority of the low-SES Recife subjects will moralize the harmless-offensive stories, because this group is likely to have a broader, non-harm-based morality.

(…)

No age effects are predicted."

(Haidt et al., 1993, p. 616)

Before coming to the results of the study, let me demonstrate how the predictions (2) – (4) and an additional prediction (5) that validates the SIM can be derived from the MFT and the SIM. It is this derivation which allows one to regard the study as validating both models. Prediction (1) is formed on the basis of earlier results and is needed for deriving the others: Unrelated studies had shown that college students from the US tend to make moral judgments based mainly on the reciprocity/suffering domain, providing support for the additional premise that Americans with high SES base their moral judgments on suffering and reciprocity. A terminological point is that "prototypical moral-conventional distinctions" as referred to in the experimental design are distinctions according to Turiel's harm criterion that Haidt rejects for reasons given in section 2.2.1 and against which this study is supposed to provide support.

Besides these additional presumptions, the following elements of the models are involved in the derivation of the predictions:

(1) **A moral judgment** is a judgment that is made according to a moral norm *(see section 2.2)*
(2) **The moral norms** of a culture are
    a. the set of norms that are seen as most important by its representatives *(see section 2.2)*
    b. dependent on which moral domains are nourished in the cultural environment of an individual. *(see section 2.2)*
(3) **Moral judgments** appear in consciousness *as the result of moral intuitions (see section 2.1)*
(4) **Moral intuition** is defined as the sudden appearance in consciousness of a *moral judgment* without any conscious awareness of having gone through steps of searching, weighing evidence, or inferring a conclusion. *(see section 2.1)*
(5) **Moral reasoning** is a mental activity that consists of conscious transforming of given information in order to reach a *moral judgment*. *(see section 2.1)*
(6) **The role of moral reasoning** is not generation of judgments but only their post-hoc rationalisation *(see section 2.1)*

The key concepts within these premises have been operationalized in the following way:

- **Cultural differences** were operationalized through the constructs of *westernization* and *SES*.
- **Relative Importance** of judgments was operationalized through the *interference* and *universalization* question during the interview.
- A judgment is mediated by a certain **moral domain** if the associated emotion is displayed by the judging person and the norm given in justification question is associated with that domain.
- **Emotional arousal** is operationalized through the *Bother* question.

Prediction (1): A majority of the high SES Philadelphia subjects will take a permissive stance towards the harmless-offensive stories, because this group has a harm based morality.

This first prediction is actually just a reformulation of the additional premise mentioned above. It does not really support the MFT, but is needed to derive the other predictions.

Prediction (2): There will be a main effect of city, or westernization, in such a way that the harmless-offensive stories will be moralized most in Recife and least in Philadelphia.

&

Prediction (3): There will be a main effect of SES, such that within each city the harmless-offensive stories will be moralized more by low-SES subjects than by high-SES subjects.

These predictions are derived from the MFT's postulate of culture dependence of the moral domain on the one side, and on the other side from the operationalization of culture differences as differences in westernization and SES. The culture differences implied by differences between the cities and socioeconomic standards of the groups are seen as big enough to ensure a significant change in morality.

Prediction (4): A majority of the low-SES Recife subjects will moralize the harmless-offensive stories, because this group is likely to have a broader, non-harm-based morality.

As westerners moralize the suffering and the reciprocity domain but as far as predicted not the other domains, the place to look for cultural influences on morality is the non-harm-based domains. The MFT postulates these domains to pertain to transgressions against norms of hierarchy and purity. Accordingly, the deeper the cultural differences, the higher the probability that subjects moralize harmless-offensive stories, not however the nonoffensive stories.

Prediction (5) In populations with broader morality*, Bothering* will have a greater predictive power than *Harm* about the universalization and enforcement of a judgment.

This prediction can be derived from the SIM claim that moral judgment is automatic, intuitive, and emotional combined with the operationalization of moral judgments as judgments that are universal-

ized and enforced (being about norms that are rated as very important). As emotional arousal is measured through responses to the *Bothering* question, the SIM predicts *Bothering* to have greater predictive power towards moralization than *Harm* in populations whose morality extends the harm-related domains.

2.3.3. Results

All hypotheses were confirmed by the results. The scenarios were generally perceived as the experimenters did hope: Harmless scenarios were generally perceived as harmless while harmful scenarios were generally perceived as harmful[15]. Among the harmless offensive acts, those pertaining to transgressions of authority or community values proved to be considered only mildly offensive, while the ones displaying disgusting behaviour proved to be more effective. *Bothering* had a greater predictive power than harm about the universalization and enforcement of a judgment. The authors themselves conclude

> Both of the Philadelphia high-SES groups made large distinctions between the harmful story (Swings) and the harmless-offensive stories (Prediction 1); the low-SES Recife subjects made small or nonsignificant distinctions (Prediction 4); and the moral-harmless distinction was affected by city (Prediction 2) and SES (Prediction 3) in the predicted ways.
>
> (Haidt et al., 1993, p. 623)

A detailed account of the results can be found in Appendix 1. As the focus of this work is on operationalization and concept use, I leave it to the interested reader to assess whether the data really confirms the predictions or not.

[15] Exception kissing siblings

2.4. The next steps

I told three stories in the course of this introduction of my exemplary specimen of empirical work on moral judgment. I told the story that in football, theoretical (for example tactical, psychological and physiological) considerations are considered only as valuable as the practical results they help to achieve. I told the story of a fellow student who gave the "wrong" correct answer in a test and received zero points for misunderstanding the question. I told the story of the little boy who was not at all satisfied with a priest's advice because it failed to provide clear instructions on what to do. All these stories relate to ways in which unclearity can thrive and ultimately lead to failure: it can thrive by not testing how theories and concepts do when confronted with real world situations, as in the case of football. It can thrive if questions are not formulated clearly enough like in the case of my fellow student. And it can thrive as long as we do not go all the way into the details of application like in the advice of the priest.

I have tackled the first way with the mere introduction of the theory as an object to test my hypotheses on. I have tackled the second way by presenting the MFT and the SIM together and making clear which questions about morality are actually answered by them. I have tackled the third by thoroughly analyzing the concepts within Haidt's moral theory and explaining how this understanding can lead to the manifest operationalization of these concepts and consequently to the validation of the theory.

I hope that my analysis of Haidt's moral theory and its key concepts helped to clarify that some of these key concepts fit as perfectly and smoothly into Haidt's moral theory as if they were tailor-made for it– and that other interpretations of these concepts might just not work out. Take for example the exclusiveness of the intuition vs. reason distinction; the very neat overlap between emotion and intuition; the learned character of emotion and the culture dependence of moral domain formation.

These concepts and their mutual dependencies carry Haidt's theory. It is these neat conceptual interdependencies that are the target of the next section. Among these concepts I look for philosophical commitments - and I find them. It is these philosophical commitments that render several philosophical conclusions drawn from Haidt's theory illegitimate.

3. **Philosophical commitments in MFT and SIM**

After introducing the two exemplary models and one exemplary study about moral judgment, I am now finally in a position to look at where exactly philosophical assumptions make their way into the cognitive science[16] of morality. In each part of this chapter, I deal with one philosophically tricky aspect of the mentioned models and operationalization. It is my strategy to individuate certain philosophically challenging aspects of the term *moral judgment* that play an important role for defining what moral judgment is - and consequently what the mentioned psychological models are actually about. These aspects include on the one hand the meaning of terms which are employed for explaining moral judgment like *intuition*, *reason*, *emotion*, and *intention*, and on the other hand rather typical metaethical matters like "what is judged in moral judgment", whether or how moral judgments can be true or false and what the syntactical structure of *moral judgment* is.

I dedicate one section of this chapter to each of the matters just addressed. In these sections, I normally proceed in the following way:

Each section has a rather metascientific part which contains roughly the following steps:

- The introduction of the standpoint that Haidt's theory has about this particular aspect
- A comparison of this standpoint with significant philosophical points about the aspect under discussion – in general one that is in accordance with Haidt and one that is not
- An assessment of what would happen if one substituted the alternative philosophical conceptions for the understanding of the aspect shown in the model. Generally, the substitution would have the consequence that the empirical predictions or the interpretation of the models would change or the rationale for the model validation would not work out anymore. I regard this as an easy way to establish that the MFT or the SIM in effect depend on a certain philosophical understanding of *moral judgment*.

After this analysis of the implicit assumptions made in psychological research, each section concludes with a metaphilosophical part:

- I generally finish each section by introducing an example of a philosophical argument that is trying to argue along the lines of Haidt's theory of moral judgment and fails to do so due to unawareness of the substantial or terminological issues that have been discussed before in the metascientific part of the section.

[16] I will use *cognitive science* in this thesis to refer to the conglomerate of sciences dealing with cognitive processes.

The sections in the beginning of this chapter have a rather metascientific focus, the metaphilosophical share of the single sections gradually increases from section to section. This order is interrupted only at times when I need to introduce an idea in one section that play an important role in another, subsequent section. This chapter contains the following section dedicated to the following concepts:

The segment about **intuition and reason** shows that the dichotomy between intuition and reason as upheld in psychology does not hold in philosophy as here, intuition and reason have a slightly different explanatory function. The philosophical commitment of the SIM consists in a use of these concepts that is distinct from philosophical and probably also from common language use. *Intuition* in the SIM is something slightly different from *intuition* in philosophy and common language. In the metaphilosophical part I demonstrate that due to this difference in concept use, it is highly dangerous to draw *philosophical* conclusions about reason and intuition on the basis of *psychological* results about reason and intuition.

Subsequently, it is confirmed that Haidt's concept of **emotion** does in fact imply a philosophical position about what emotions are – adopting certain philosophical standpoints about emotion would render the example evidence inconclusive. In the metaphilosophical part, I highlight that the philosophical approach to emotion of the philosophical theorizer determines which philosophical conclusion can be drawn from Haidt's results.

The section about **syntactical structure** of moral judgment reveals that in the formulation of what actual moral judgment is, a choice is necessary between graded moral judgments ("patricide is worse than homicide, but better than infanticide") and binary moral judgments ("patricide, infanticide and homicide are all morally wrong and not morally good"). This choice can be mapped to different philosophical positions. If that choice is not made or neglected, the risk of mixing up possibly distinct psychological effects arises. It is exactly this what happens to Haidt who assumes that his theory is supported by a certain experiment that I introduce at this point. In the metaphilosophical part, I attest that this risk also exists for philosophical arguments.

The segment about the **object** of moral judgment shows that depending on whether persons or their actions are the carrier of moral properties, evidence has to be interpreted differently. Furthermore it is established that Haidt's own philosophical interpretation of his own empirical results undermines the conclusiveness of the evidence of his own theory.

The role of **intention assessment** in moral judgment is investigated in the first of two sections that are mainly of metaphilosophical interest: I argue that the indifference that Haidt shows about inten-

tion-sensitivity of moral judgment (whether something was done on purpose or not) allows for different philosophical interpretations of his theory. Certain interpretations of Haidt's approach would not be legitimate for certain philosophical positions and would lead to new perspectives to use empirical methods to make philosophical points.

Lastly, the section about **objective or relative truth** of moral judgment deals with a mainly philosophical point, too: What does it mean that a moral judgment is true? Does the moral foundations theory make a philosophical commitment that moral judgments are only and just true relative to culture? I argue that it does not and that it can also teach valuable lessons to adherents of objective moral truth.

I decided to start this analysis with the two most prominent terms used in the SIM and MFT to explain moral judgment. The explanation of these terms in the last chapter asked for a lot of concentrated attention from the reader and they are so central to Haidt's moral theory that any investigation about philosophical commitments will have to start from here. As mentioned above, I come to the rather surprising result that Haidt uses the terms slightly, and really just slightly different than philosophers normally do. This difference has no effects for scientific practice, but for the philosophical interpretation of Haidt's moral theory, these differences are crucial and do, in fact, make a difference.

3.1. Primitives I: Intuition and reason

During my first semesters as a philosophy student, I learnt a great practice how to say to somebody that what she or he says sounds terribly stupid: You go along, let the person finish and then say something like „this all sounds pretty nice, but this-and-that really strikes me as quite *counterintuitive*." The reason why this is more polite than saying "this sounds stupid" is that it implies that you have not spent much thought on the matter and would be rather curious which arguments can be given in favor of that point. In saying "this is counterintuitive" there is always a hidden "convince me", while in "this sounds stupid" there is always a "you are stupid".

This way of dealing with intuition as a challenge to argue is somehow typical for philosophy. If there is no reasonable argument to be found, our intuition will have to do – but can be challenged by people with opposite intuitions. As soon as you find a convincing argument, intuition however is always

the loser – and you are the winner. Intuition can basically be viewed as a type of placeholder waiting to be replaced by a proper argument. Or as Timothy Williamson puts it: "when contemporary analytic philosophers run out of arguments, they appeal to intuition" (Williamson, 2004, p.109).

Especially in moral philosophy, where discussions tend to be rather impassioned and knock-down arguments tend to be rare, the relationship between reason and intuition plays a far reaching role. Alongside the concept *moral* itself and its different philosophically challenging aspects, *intuition* and *reason* are therefore obvious candidates for hidden philosophical assumptions in the presented study and models. In what follows, I demonstrate that even though an important component of the typical philosophical understanding of *intuition* is pretty compatible with Haidt's approach, there is one key difference in how this idea is generally looked upon in philosophy. While the main interest in empirical science is about how beliefs are generated, in philosophy the *truth* of these beliefs is taken to be at least as important a matter. And while science employs intuition and reason as ways to *create* beliefs, philosophy employs them also as ways to *justify* beliefs. In this section, I establish the point that this difference becomes a problem for empirically minded *philosophers* the moment they are trying to draw philosophical conclusions about the relationship between intuition and reason on grounds of the SIM. This is exemplified by an argument that relies on the identity of the philosophical and the psychological terms *intuition* and *reason* is invalidated by taking their disparate use into account.

But back to the start: In order to contrast the way Haidt is using the terms with the way it is used in philosophy, I now introduce two typical examples for the role that *intuition* plays in moral philosophy. The aim of this step is to enable me to draw the conclusion that at least with his usage of the terms *intuition* and *reason*, he does not commit himself to a philosophical position – he rather shuts himself off from philosophical discourse by using the term differently in general.

3.1.1. Intuition in philosophy: a defeasible indicator of truth

The way that intuition becomes interesting for philosophers can be assessed very easily by looking at the role the appeal to intuition plays in our everyday practice. Intuition is appealed to:

- in order to justify a foundation, premise or starting point of an argument
- in order to highlight that we have not yet fully understood a matter but will until further notice endorse a certain attitude about it

- in order to say that we "just know" something – that there is no point in arguing as this thing is self-evident to us.

In sum, wherever there is arguing, there seems to lure some kind of appeal to intuition. Due to this constant reappearance, intuition has come to appear in arguments for and against certain philosophical positions again and again. And very importantly, intuition gives not only a way of generating beliefs, but also a way of justifying beliefs. Let me introduce two examples of classic works from moral philosophy to show that this is not a mere claim: G.E. Moore's account of *intuition* in his consequentialist, intuitionist metaethics and the role of *intuition* in John Rawls's deontologist[17], rationalist theory of justice.

Example #1: Moore, the good and intuition

In his classic *Principia Ethica* (Moore, 1903), George Edward Moore gives an account of a consequentialist[18] ethics that is decisively non-utilitarian[19]. He argues against Utilitarianism on the grounds that he regards the term *good* as not definable and therefore not possibly equivalent to "maximizing utility". *Good* is, just like *yellow*, a primitive (not analysable) term – our understanding of what *good* means is purely intuitive.

So how does Moore use the term *intuition*? First of all, intuitive beliefs, or *intuitions* are *self-evident*[20]. What it means for a belief to be *self-evident* is however a tricky business.

[17] An account of morality is called *deontologist* if an action's rightness is an intrinsic feature of that action, that means this action is morally good in itself and not because it brings along something else that is the actual morally good thing.

[18] An account of morality is called *consequentialist* if an action's rightness depends on the goodness of its consequences

[19] A utilitarian account of morality is an account in which an action's rightness depends on the utility of its consequences – however there is room for substantial disagreement how utility is to be defined. (for example Williams 1973, Parfit 1984, Rawls 1971)

[20] As can be seen for example in §36 (Moore, 1903): "(…) of all hedonistic writers, Prof. Sidgwick alone has clearly recognized that by good we do mean something unanalysable, and has alone been led thereby to emphasise the fact that, if Hedonism be true, its claims to be so must be rested solely on its self-evidence—that we must maintain Pleasure is the sole good to be mere intuition. "

> "By saying that a proposition is self-evident, we mean emphatically that its appearing so to us, is *not* the reason why it is true: for we mean that it has absolutely no reason. (…) That it appears true to us may indeed be the *cause* of our asserting it, or the reason why we think and say that it is true: but a reason in this sense is something utterly different from a logical reason, or reason why something is true. (…) Again that a proposition is evident to us may not only be the reason why we do think or affirm it, it may even be a *reason* why we ought to think it or affirm it. But a reason, in this sense too, is not a logical reason for the truth of the proposition, though it is a logical reason for the rightness of holding the proposition. In our common language, however, these three meanings of 'reason' are constantly confused, whenever we say 'I have a reason for thinking that true'. But it is absolutely essential, if we are to get clear notions about Ethics or, indeed, about any other, especially any philosophical, study, that we should distinguish them."
>
> G.E. Moore: *Principia Ethica,* CUP, §86

Moore's point here is that causal reasons and logical reasons for a belief have to be carefully distinguished. While the fact that something *appears* to us is the *causal* reason why we hold a belief, it is not a *logical* reason for the truth of a belief. It can however be a good indicator that there *could* be some good logical reasons and a good indicator that we *should* hold the belief for true. The role of intuition concerning self-evident beliefs is threefold: it is the causal root of self-evident beliefs. It is an indicator for their truth. It is no logical reason for their truth. Concerning the relationship between reason and intuition, Moore adds the following point: "We must not therefore look on Intuition, as if it were an alternative to reasoning. Nothing whatever can take the place of reasons for the truth of any proposition: intuition can only furnish a reason for holding any proposition to be true: this however it must do when any proposition is self-evident, when, in fact, there are no reasons which prove its truth." (Moore, 1903; §86)

This point is crucial for the rest of this chapter: while intuition and reason are alternative ways to *create* beliefs, they are not in the same way comparable as means to *justify* beliefs. A belief that once was a spontaneous idea, for example that the square of the hypotenuse of a square-angled triangle equals the sum of the squares of its catheti ("$a^2 + b^2 = c^2$"), can be a causal product of intuition but justified by mathematical proof (a paradigmatic form of reasoning). Actually, this is how proving something works quite often: one has a more or less spontaneous idea, then a reasoning process sets in and a proof for the idea is derived. Once proven, the sentence remains intuitive in the sense that it is still a causal product of intuition, but our reliance on its truth is not dependent on its intuitiveness anymore. If the sentence were not intuitive, we would not have had the idea that there is something to prove.

We can conclude that intuition is described by Moore as a) a way to derive beliefs and b) a good reason to entertain a belief if no reason against its truth can be found. C) Beliefs generated by intuition have the property that we are not aware of any logical reason to hold them at the time of their generation.

If we hold a belief because it is self-evident, no inference steps have been taken to derive its truth from somewhere else. By calling intuitions self-evident, Moore has to be seen as describing intuition as a one-step-process that generates a judgment about the world without the cognizer being aware of why she holds the belief. This fits very neatly with Haidt's definition. But Moore also emphasizes another role that intuition and reason play that figures in the SIM only very peripherally: justification of beliefs, in the sense of giving an account for their truth. Of course the search for reasons for one's moral judgment figures in the SIM as a peripheral process. But this view of justification as a *process* fails to account for justification as a *normative reason to regard a belief as true*. Moore's point here is that truth is (at least for philosophy) just as important as causation of moral judgments – if not even more so. If intuition and reason are understood this way, a belief can be intuitive **and** reason-based at the same time if it is generated by intuition and justified by reason. Indeed the main job of reasoning is to find out about *truth* of beliefs, not to generate beliefs. The important part here is the double role that intuition and reasoning play: on the one hand, they are processes in a causal chain, on the other they are normative reasons for regarding beliefs as true.

This line of thought should help us understand that the SIM and the MFT are concerned with how moral judgment is generated and why (causally) it is *regarded* as true. Moral reasoning can however also play a role that has nothing to do with this: it can be about why certain moral judgments are *in fact* true. A nice example of this distinction would be John Rawls's theory of political justice that I am to describe next.

Example #2: intuition in Rawls's reflective equilibrium

In his opus magnum "A Theory of Justice" (Rawls, 1971), John Rawls lays down certain principles that are to govern an institution in order for it to be *just*. Whether an institution (like for example a state) is just can be assessed by checking how far its structure is in agreement with certain principles of justice. These principles are 1) the liberty principle which says that "each person is to have an equal right to the most extensive basic liberty compatible with a similar liberty for others." (Rawls, 1971, p.60) and 2) the equality principle which says that "Social and economic inequalities are to be ar-

ranged so that they are both (a) to the greatest benefit of the least advantaged and (b) attached to offices and positions open to all under conditions of fair equality of opportunity." (ibid. p.83).

In the course of deriving these principles, Rawls devotes a lot of attention to the methodological aspect of the challenge how to distinguish good or true candidates for principles of a just society from bad or wrong principles. In the end, whether a set of principles is acceptable will be dependent on its match with our moral intuition:

> "We can note whether applying these principles [of justice] would lead us to make the same judgments about the basic structure of society which we now make intuitively and in which we have the greatest confidence; or whether, in cases where our present judgments are in doubt and given in hesitation offer a resolution which we can affirm on reflection. There are questions which we feel sure must be answered in a certain way. For example, we are confident that religious intolerance and racial discrimination are unjust. We think that we have examined things with care and have reached what we believe is an impartial judgment not likely to be distorted by an excessive attention to our own interests. These convictions are provisional fixed points which we presume any conception of justice must fit."
> (Rawls, 1971, pp. 19)

Our intuition therefore is the final litmus test and foundation for whether the sentence "The principles for a just society are X,Y,Z" can be *true*: we are able to derive in a reasoning process that X,Y,Z entail certain sentences concerning situations about which we have very clear intuitions and of the truth of which we are firmly convinced. These intuitive convictions are what Rawls calls "fixed points" of morality. If the principles X,Y,Z in question would entail a contradiction to these fixed point sentences, one would have to regard X,Y,Z as very implausible candidates for principles of a just society. Only principles that are in agreement with these intuitive fixed points of morality can be considered eligible as principles of justice. As soon as this kind of justification of X,Y,Z is established, we are able to derive other moral judgments from X,Y,Z about situations about which we have no particular intuition. This use of reason in order to guide our judgment in difficult situations is why Rawls's theory is regarded as a rationalist approach to moral judgment. Principles of justice can only be justified by reason, but they are justified by reason according to their fit with our moral intuition. Our benefit in having these principles lies in being able to have a way of distinguishing right from wrong in questions where intuition is either lacking or not giving us a clear result.

What it comes down to is that just like in Moore's *Principia Ethica*, in the context of Rawls's theory of justice intuition is not exclusively treated as a *mechanism* that leads us to entertain certain beliefs– it is treated as a vehicle of justification. It is seen as a basic kind of judgment that is not further analys-

able – the kind that Moore calls self-evident. Otherwise, it would not give us fixed points about morality. We can therefore assume that intuition understood in the way that Rawls is using the term will be a) a mechanism that provides us with beliefs that are noninferential and b) a way to justify a moral belief. And just like Moore, Rawls seems to regard arguments against it as an important reason to drop a certain intuitive belief: otherwise he would not have included the qualifier to the "fixed points" of morality that "We think that we have examined thing with care". This examination of self-evident beliefs will just have to include the search for arguments pointing in the opposite direction – for example towards a conclusion that the intuition in question is not a moral intuition but just a selfish act of self-deception. Therefore, I will add the condition that c) intuition is a reason to uphold a belief that can be invalidated by acts of reasoning. c) is furthermore supported by a widely accepted argument of Richard Dworkin's in his discussion of Rawls's reflective equilibrium (Dworkin, 1977, pp. 159) that shows that the reason part of the reflective equilibrium ought to be understood as being able to determine whether certain of our intuitions conflict even under the assumption of our best principles of morality and therefore have to be reconsidered.

I therefore conclude that just like in the *Principia Ethica* the role of intuition in *A Theory of Justice* is not only that of a one-step mechanism to create beliefs like understood by Haidt but also that of a *justification* for beliefs. Again, justification should not be understood as a process, but as a *normative reason to regard a belief as true*.

3.1.2. *Intuition* and *reason* in psychology and philosophy – a minor disagreement

Let me repeat the main points that have been made so far about *intuition*: Haidt agrees with Moore (a consequentialist, intuitionist philosopher) and Rawls (a deontologist, rationalist philosopher) that intuition is a mechanism to derive beliefs. They agree that it is a one-step mechanism - that we are aware of the result of the mechanism but not of the process (which means *how exactly* we arrived at a certain idea). We have no control over the mechanism – we cannot decide what our intuition will be (hence it is stable over time and can be relied on to indicate the truth of a belief) and we cannot just switch it off (with which I do not mean one cannot *ignore* one's intuition – I actually mean wilfully not having a certain intuition).

But there is an important disagreement that Moore has already warned us of: intuition is no logical reason for a belief to be *true*. In fact, it is merely what helps us out once there is no logical reason for a belief. This disagreement is not really substantial at first sight. It seems rather like a qualification or

minor caveat that is added to the mass of joint properties that intuition in Haidt's sense and intuition in the philosophers' sense share. And indeed, as long as we just regard *intuition* as a concept on its own, this is the case. But there is an important piece missing from the picture: while Haidt' definition of *moral intuition* allows him to regard *moral reasoning* as its counterpart, Moore and Rawls have to paint a more fine-grained picture to describe the relationship between *moral reason* and *moral intuition* – according to this fine-grained relationship, a belief can be intuitive in the causal sense while in the justificatory sense it is *justified and shown to be true* by reason and therefore not a pure product of intuition. In philosophy, reason and intuition are not exclusive, due to their ambiguous role in philosophical thinking. A belief derived from an automatic process is a product of intuition; the contrary would be a belief that is originally derived through an act of reasoning. However, the *truth* of an intuitive belief can be assessed much better through reasoning than through intuition. An intuitive belief gains strength from reasoning – it can therefore be both a product of reasoning (concerning its truth) and a product of intuition (concerning its generation). I referred to intuitions before as ideas. Spontaneous ideas are by definition intuitive, but whether they are good ideas or bad ideas will be often determined through acts of reason.

The disagreement between Haidt's approach and the philosophers' approach is therefore less a disagreement about the actual procedural role of reasoning and intuition in moral judgment, it is rather a disagreeement about what is the intresting part of moral judgment: how it is derived or how it can be understood to be true.

Let me recapitulate what was shown so far about the SIM and the MFT: the way intuitive and reason-based judgments are juxtaposed in the SIM only works as long as one sticks to Haidt's definition of the terms *intuition* and *reason* that are strictly about causation of beliefs and not about justification. As soon as justification as a normative function of intuition is brought into the game, intuition and reason tend to play slightly different roles that do not allow for the dichotomy that Haidt suggests. Even though this peculiarity about concept use does not directly affect scientific practice negatively per se, it is an aspect of moral psychology that one should always keep in mind, especially when communicating results to laymen who might share the justificatory conception of intuition and reason of philosophy. Not making explicit this subtle but important difference could lead to serious misunderstandings of what Haidt's theory actually says.

This concludes the metascientific part of this section. In the metaphilosophical part of this section I now demonstrate how the implicit character of the specific meaning of *intuition* in the SIM and MFT can lead philosophers astray. If the SIM is combined with philosophical argument to confirm a philosophical conclusion, one should be very aware about the difference between the SIM-understanding of *intuition* and the philosophical understanding of *intuition*.

3.1.3. An example: Joshua Greene's argument for wrongness of deontology

In his essay "The Secret Joke of Kant's soul" (Greene, 2007), the neuroscientist and philosopher Joshua Greene aims to argue against deontologist positions in ethics on empirical grounds[21]. In due process, he even cites the example study as evidence for his position. This section shows that the problem with one of the several arguments that he suggests in his essay is that he regards deontological moral philosophy as a theory about how moral judgments are generated. I argue that this view about deontological moral philosophy is wrong and that it is rather a theory about why certain moral judgments are *true*.

For the argument in question, Greene divides moral judgments into two groups:

- Characteristically consequentialist judgments – judgments that are justifiable with regard to the positive or negative consequences of an action
- Characteristically deontologist judgments –judgments that are justifiable by appealing to rights or moral laws [22] (Greene, 2007,pp.38)

You might already have noted that these two groups of judgment are not at all exclusive of each other. In fact, in most cases a consequentialist judgment will be a deontologist judgment: "Murder is wrong", "Theft is wrong", et cetera. However, the suggested distinction allows Greene to evaluate a certain type of response to a certain type of stimulus in a special way: over the years, a lot of research has been done about people's responses to moral dilemmas (see Christensen & Gomilla, 2012 for a review) in which study participants were confronted with stimulus situations that featured a trade-off in human life and moral obligations, for example:

[21] This essay has sparked a lot of criticism, methodological, psychological and philosophical. (for example in the very same collection of essays: Mikhail 2007, Timmons 2007) For me, it will purely have the purpose of demonstrating the point I just made and the damage it can do. So please, please, stay with me.

[22] This conception of deontology and consequentialism is not shared by many ethicists and there are reasons for and against accepting it. As Greene's argument is introduced here for mere demonstrative purposes, I feel free to use the term in the context of Greene's argument in the way Greene suggests it.

> Enemy soldiers have taken over your village. They have orders to kill all remaining civilians. You and some of your townspeople have sought refuge in the cellar of a large house. Outside you hear the voices of soldiers who have come to search the house for valuables.
>
> Your baby begins to cry loudly. You cover his mouth to block the sound. If you remove your hand from his mouth his crying will summon the attention of the soldiers who will kill you, your child, and the others hiding out in the cellar. To save yourself and the others you must smother your child to death.
>
> Is it appropriate for you to smother your child in order to save yourself and the other townspeople?
> (Greene 2008, pp.1147)

His distinction between typically deontologist and typically consequentialist moral judgments allows Greene to regard this type of stimulus as a test to which kind of moral philosophy a person would stick: not smothering the baby would imply regarding the right of the baby to live as untouchable – regardless of that this right means killing everyone else including the baby. Smothering the baby would imply that the wellbeing of everyone else would overrule the right of the baby to live. Greene's research about this kind of moral dilemmas led him to a dual process theory of moral judgment that has to be distinguished from Haidt's theory, even though it takes Haidt's study into account: It claims that characteristically deontologist judgments are driven by emotion while characteristically consequentialist moral judgments are driven by what he calls *cognition*. Let me explain what he means by that – and in how far his take on *emotion* and *cognition* is very close to Haidt's take on *intuition* and *reason*.

Greene describes emotions as fixed bits of input-output programming that are essentially automatic and entail an action bias. The action bias part reminds very strongly of Haidt's concept of emotion implying that emotions as understood by Greene are intuitions in Haidt's sense. The part about the "fixed bits of input-output programming that are essentially automatic" supports this: it entails that emotions are uncontrolled and unintentional. The unawareness and efficiency condition are also met by Greene's understanding of emotions: in one study, he concludes deontological judgments to be emotion driven based on their corresponding reaction time being shorter than the one of consequentialist judgments (higher efficiency) (Greene, 2008). What Greene calls *emotion* can therefore be understood without any problem as belonging to the category of *intuition* in Haidt's terminology.

Cognitive processes as Greene uses the term in his essay are defined to be processes that we can actively steer and that cannot be described in terms of stimulus-response patterns (ibid. p.40). In other words, "cognitive" processes require the control, intention and awareness of the actor, which implies that they are not intuitive processes and therefore will incorporate reasoning processes according to Haidt's definition above as the conscious stepwise transformation of sentences definitely falls into the group of processes we can actively steer. They "do not automatically trigger particular behavioral responses or dispositions, while "emotional" representations do have such automatic effects, and are therefore behaviorally valenced"(ibid. p.40).

Due to these considerations, I regard Greene's distinction between *emotional* and *cognitive* judgments as close enough to Haidt's distinction between *intuitive* and *reason based* judgments to regard mistakes made on behalf of the former as mistakes that could have also been made on behalf of the latter.

Let me now turn to Greene's actual argument: first, Greene establishes the idea that characteristically deontologist judgments are emotional and not cognitive. This is the empirical input for his argument, and it is an input that could be derived from the SIM, which insists that *all* moral judgments are intuitive. This point is important and it is why I made such effort to establish that Greene's concepts of *emotional* and *cognitive* can be mapped on Haidt's concepts of *intuition* and *reason*. The SIM's claim that *all* moral judgment is intuitive can be translated into all moral judgment being emotional in Greene's sense. The moralization of non-harm violations (that are per definition non-consequentialist) demonstrated in the example study combined with the higher predictive power of emotion compared to rational justification can just as well be seen as supporting Greene's claim. Greene concludes therefore from his empirical point -*that is derivable from the SIM*- that (deontologist) reasoning has no causal effect on the formation of deontologist moral judgment. Instead, he claims that "our moral judgments are driven by a hodgepodge of emotional dispositions, which themselves were shaped by a hodgepodge of evolutionary forces, both biological and cultural. Because of this, it is exceedingly unlikely that there is any rationally coherent normative moral theory that can accommodate our moral intuitions. Moreover, anyone who claims to have such a theory at hand and actively endorses it is almost certainly wrong. Instead, what that person probably has at hand is a moral rationalization."
(Greene, 2007, p.72).

The term *rationalization* is not explicitly defined by him, but he gives a variety of examples the "most striking" (Greene, 2007, p.62) being the following:

> "Perhaps the most striking example of this kind of post hoc rationalization comes from studies of split-brain patients, people in whom there is no direct neuronal communication between the cerebral hemispheres. In one study, a patient's right hemisphere was shown a snow scene and instructed to select a matching picture. Using his left hand, the hand controlled by the right hemisphere, he selected a picture of a shovel. At the same time, the patient's left hemisphere, the hemisphere that is dominant for language, was shown a picture of a chicken claw. The patient was asked verbally why he chose the shovel with his left hand. He answered, "I saw a claw and picked a chicken, and you have to clean out the chicken shed with a shovel" (Gazzaniga & Le Doux, 1978; Wilson, 2002). (…) This widespread tendency for rationalization is only revealed in carefully controlled experiments in which the psychological inputs and behavioral outputs can be carefully monitored, or in studies of abnormal individuals who are forced to construct a plausible narrative out of meager raw material."
> (Greene, 2007, p.62)

According to this example, I take *rationalization* to be "making up a wrong explanation for one's own behavior in the absence of knowledge of the true psychological causes for it". The reader is probably already able to guess the point this passage is to establish: deontology is described as *rationalization*, *rationalization* being a wrong account of *causal reasons*. Greene sums it up in the following way:

> "What should we expect from creatures who exhibit social and moral behavior that is driven largely by intuitive emotional responses and who are prone to rationalization of their behaviors? The answer, I believe, is deontological moral philosophy. (…)
> Deontology, then, is a kind of moral confabulation. We have strong feelings that tell us in clear and uncertain terms that some things simply cannot be done and that other things simply must be done. But it is not obvious how to make sense of these feelings, and so we, with the help of some especially creative philosophers, make up a rationally appealing story: There are these things called "rights" which people have, and when someone has a right you can't do anything that would take it away. "
> (Greene 2007, p. 63)

Let me summarize Greene's line of thought, keeping in mind that Greene is not making the distinction between the justifying and causal roles of intuition and reason that I just introduced:

Deontology is wrong. That is because it is a rationalization - which means a made-up wrong explanation of the causal reasons of our behavior. This is to be the case because rationalization is typical

behavior in situations in which one is not able to explain one's own behavior by means of introspection and because morality is just too complicated to be expressed through a set of norms. Our moral emotions are so diffuse and manifold (sometimes they even lead to contradicting biases) that it seems hardly conceivable that it should be possible to find a set of norms that could possibly account for our decisions. Deontology therefore cannot account for the "real" (a.k.a. causal) reason for our moral judgment which are "just" emotionally triggered intuitive judgments. Deontology is therefore a kind of confabulation or myth, comparable with the ancient idea that in every river, there is a river god who decides when there will be floods and draughts. Today, as we know about the actual causes for floods and draughts, we regard this once widely held belief as clearly mistaken.

3.1.4. The problem with Greene's argument

Now let me put the pieces together: Greene mistakes deontological justification for a) an *account of the generation* of moral judgment and b) for an *axiomatic account of our moral beliefs* while in fact it is an *attempt to justify* certain moral standpoints. This is why he takes the intuitive character of deontological moral judgment to indicate the wrongness of deontological moral theories. And that is where the twofold role of reason and intuition comes to bear: he may be right by stating that deontology would do a bad job in telling us about the generation of moral judgment, given the validity of the evidence for deontological judgments being purely emotional[23]. But that is not the job of deontological thought after all. Its job is normative justification of beliefs - not explanation of their causal origin. If for example we regard the example of Rawls's account of justice, we can see very easily that it does not try to explain judgments and decisions in a causal way. The purpose of this theory is finding out principles that would entail the truth of our moral fixed points and on which we could trust to guide us in situations where we have no clear intuitive response. Moral intuition is seen as an important test to assess whether a theory is acceptable or not. Deontology is therefore not a myth or a story we tell about the causation of moral judgments. It is the justification of why we should consider them **true, or at least justified**.

Note that the inference of the wrongness of deontology is dependent on the reason-intuition dichotomy which can only be upheld in the original SIM-meaning of the term, as I established above. If the philosophical approach is adopted, intuition and reason play the double role of belief generation *and*

[23] Even this point remains under debate. But this is not the issue under discussion.

justification. In this case, it would not be possible to conclude from the (causal) intuitiveness of deontological moral judgments that moral judgment cannot be the (logical) result moral reasoning. So, regarding deontology as wrong is a result of mistaking the intuitive nature of deontological moral judgments as proof that they cannot be a product of reasoning and therefore reason.

Greene's second criticism against deontology is that he regards moral emotions as so complex and diffuse that it seems hardly possible to create a logically consistent set of principles that is in agreement with them. However, this is not the point about deontological thought either. The contrary is the case: only through generating principles of morality it becomes possible to spot incoherencies among our moral intuitions in the first place – and then to decide which intuition should be considered moral error and which intuition should be considered moral truth. It is not the main job of principles of morality to adequately *describe* all incoherencies of moral emotion or even to explain them away –its main job is to help us *deal* with these incoherencies.

This indicates that it would be a wrong conclusion to infer the wrongness of a normative moral theory from the intuitiveness of moral judgment as understood and postulated by the SIM. The reason for this lies in the way that *intuition* is used in the SIM, which is not entirely compatible with the way *intuition* is used in the context of philosophical moral theory.

3.1.5. Conclusion

In this chapter, I explained that an important part of the understanding of *intuition* in moral philosophy matches pretty well with its understanding by Haidt – but that in philosophy, intuition is not only a factor for the generation of beliefs but also a factor concerning their justification, and therefore plays a twofold role. Whether a belief is intuitive in the one sense or in the other can be assessed independently, meaning that a belief can be at the same time a (causal) product of intuition without being mere intuition in the justificatory sense, as good reasons can be found to regard it as true. The meaning of *intuition* and *reason* in Haidt's moral theory is therefore, from the philosopher's and probably also from the layman's perspective, a special case. Haidt can be regarded as committing himself to a certain, very narrow meaning of *intuition* and therefore to a certain very restrictive philosophical standpoint about what moral intuition and moral reason do.

This becomes especially tricky once one is to employ Haidt's theory in philosophical argumentation about reason and intuition: ignorance of the disparity of the common language term *intuition* and

Haidt's conception can lead to wrong philosophical conclusions from empirical accounts similar to the example theory and study. That confirms my second hypothesis that as soon as one makes a commitment about theoretical terms, these commitments have grave effects on empirical philosophy. In the philosophical argument at hand, the original SIM-understanding of intuition was used to make inferences about moral intuition in the philosophical sense of the term. The SIM was taken to be committed to a certain meaning of *intuition*, and it was overlooked that this meaning is is not the same as the one in deontological moral philosophy. It might come as quite a surprise that the SIM has actually relatively little to offer about the relationship between reason and intuition in moral judgment as understood in the philosophical sense.

In the beginning I told an anecdote about the word "counterintuitive" and that it entails the invitation to give and explain reasons for a belief. And indeed the justifying roles of intuition and reason can be nicely illustrated with that situation – the question for reasons can be analyzed as making a twofold statement: In the case that the opposite has no reasons for her or his claim, it entails the repudiation of the point of the other side – the other side might of course still regard the claim as justified by her particular intuition. The two of you are in a kind of impasse. However in the case that the other person *is* able to give good reasons, you promise to accept the conclusion, as the best thing *you* are able to offer is just an intuition about it.

This concludes the part of this investigation that deals with the concept *intuition*. In a way however, it sticks around for the rest of the thesis: the next chapter is dedicated to something very close to intuitions, namely emotions. And indeed, the term *emotion* taken for itself seems much less philosophically loaded than the terms *intuition* and *reasoning* – which is why it has been used much more eagerly in order to explain moral judgment in moral psychology. In the next chapter I demonstrate that it is a mistake to assume that *emotion* has less philosophical baggage than *intuitio*n. There are a lot of philosophical points to be made about emotions and therefore a lot of philosophical commitments to be bought in the psychology of moral emotion. I present three concurring philosophical accounts of emotion and find that two of them are compatible with Haidt's theory – while one of them neither fits with explicit identity criteria that Haidt assumes for emotions, nor with the operationalization of emotional valence in the example study. Haidt can therefore be assumed to exclude this particular philosophical standpoint and make a clear commitment about the meaning of the term *emotion*. In the more metaphilosophical part, I demonstrate how different understandings of *emotion* can lead us to differing interpretations of the very same empirical data.

3.2. **Primitives II: emotion**

If psychology's and philosophy's key concepts to describe the inventory of the mind are in the end incompatible in the way shown above, one might question the very use of experiments on moral psychology for philosophy . But I want to assure the reader that the interests, and accordingly the concepts of philosophy and psychology are not always this disparate. In fact, if one looks at how experiments and models like Haidt's are employed in philosopical practice, a much more promising concept emerges that might build bridges from moral intuition in psychology to morality in philosophy: the concept of **emotion.**

And indeed, when it comes to how moral judgment works, emotion, passion, desire and other mental inventory different from "pure reason" is appealed to much more often than cold intuition. One of the most famous sentences in philosophy in this regard would probably be David Hume's statement that "reason is, and ought only to be the slave of the passions" (Hume, 1975, p. 415). As early as in Aristotelean ethics we can find the virtue of οργη (*orgé*), the capacity to feel the right amount of anger at the right time, as a constituent of how a human being should behave.

When it comes to crossdisciplinary arguments, especially sentimentalist theories of moral judgment seem apt to be supported by empirical research about emotion[24] and Haidt's theory. However, there is some considerate amount of discussion in philosophy about what emotions actually are. On the following pages, I present different philosophical accounts of emotion and check whether they are compatible with Haidt's approach – I conclude that one of the accounts presented is indeed not compatible with Haidt's definition. I continue with demonstrating that if we assume this philosophical stance about emotion, the example study does no longer support the MFT and the SIM – therefore Haidt' theory can be regarded as committed to a certain philosophical understanding of emotion. I further establish that even in neuropsychology, there are differing ideas about emotion which in the end affect the philosophical conclusions one can take from the example study. In the end, I demonstrate how a more restrictive concept of emotion than Damasio's allows to draw additional philosophical conclusions from the example study.

[24] And they are, as I am to demonstrate in due course.

3.2.1. Philosophical theories of emotion

If one looks at how philosophy deals with emotion, one can very often find a general distinction between *cognitive* theories of emotion and *perceptive* theories of emotion.

Even though Damasio's theory of emotion is sometimes considered as belonging to the perceptive side of the spectrum (Deonna & Teroni, 2012), I highlight one philosophical account from each side (Jesse Prinz's from the perceptive side and Robert Roberts's from the cognitive side) that fits very nicely to those points of Damasio's account that are important to Haidt's models. However, I also present a cognitive account (Martha Nussbaum's) that decisively does not fit Damasio's and therefore Haidt's account.

Let me begin by presenting these philosophical conceptions of emotion and then explain in how far I see them in agreement with or in opposition to the most important points about emotion in the SIM and the MFT.

Example #1: Jesse Prinz's perceptivist account

According to perceptive theories of the mind, nothing needs to be understood, judged or appraised in order for an emotion to occur. There are two ways in which emotion can be understood to be perceptive: the emotion is either a directly perceived state of the soma (for example according to the James-Lange theory, see James 1884) or the emotion is considered to be a vehicle for perceiving qualities of our environment (for example in Prinz, 2003). Note that these ways that an emotion can be perceptive are not exclusive of each other. As Deonna and Teroni put it, this concept of emotion postulates that "emotions are essentially felt" (Deonna & Teroni, 2012, p.74). The emotion of fear for example can be explained as the *perception* of a bodily state (increased heart rate, disposition to sweat and other effects of stress-hormones) and/or as a vehicle for *perceiving* something as dangerous. It would not be necessary to *understand* what the situation or object perceived as dangerous *is* (for example, whether it is a leopard or a jaguar) - apart from being dangerous. There would be no "judgment" involved; just like there is no "judgment" to whether an object looks yellow (or, in the case of the jaguar and the leopard, yellow with black dots).

A typical example for an up-to-date perceptive theory of emotion would be Jesse Prinz's account of emotions: he understands emotions to be perceptions of qualities outside of the body (for example

dangerous) *via* perceptions of somatic states (fear-related body changes). Emotions therefore are a vehicle for perceiving qualities of our environment. Note that no *concepts* of external qualities need to be mastered for undergoing an emotion: perceiving a situation as dangerous via perception of bodily changes does neither require a concept of *danger* nor a concept of the *kind of thing/situation* that is in fact perceived as dangerous. The contrary would be the case:

> The perception of that bodily state represents danger, because it is under the reliable causal control of dangerousness. Danger is the property in virtue of which these highly desperate eliciting conditions have come to perturb our bodies. If loud noises and looming objects were not dangerous, they would not have their characteristic effects.
> (Prinz, 2003, p.55)

Concepts of danger, obscenity or sadness are based on our practice of perceiving these qualities directly via our emotions. *Dangerous* could therefore be analyzed to mean something like "fear eliciting under given circumstances". Of course, a lot will depend on these given circumstances – but perception is always dependent on certain circumstances[25], which makes this additional condition rather unproblematic. In summary: according to Prinz's perceptionist theory, emotions are a) essentially felt, b) require no knowledge of the world or concept mastery, and c) can be described as a way of perceiving qualities of our environment. Accordingly, d) emotions can be regarded as the perceptual bases for certain descriptive terms referring to objective qualities of our environment, like "dangerous". The properties of independence of concepts and foundationality for descriptions of the world, and in a certain way also the perceptiveness of emotions are refuted by the approach that I am to present next: Robert Roberts cognitive theory of emotion.

Example #2: Robert Roberts's cognitive theory

Cognitive theories of emotion regard acts of cognition or judgments as necessary elements of emotions. Typical examples given by cognitive theorists to emphasize their point are grief or jealousy (Nussbaum, 2004). Grief is thereby not considered as a *result* of my being aware of the loss of a dear person, it is a) in itself a decisive part of *understanding* the loss, in that an evaluative judgment about a given situation is made and furthermore b) grief necessitates an understanding of the situation. Note that in this interpretation, but not necessarily within a perceptive framework of emotion, the

[25] For example sufficient light in the case of vision, not having a cold in the case of olfaction

statement “you are wrong to be angry/sad/jealous” is a valid statement. Judgments can be wrong. Perceptions cannot[26].

An up-to-date cognitive theory of emotion is Robert Roberts’s (Roberts 2013). According to his approach, emotions are a vehicle for slicing up our perception of the world in a way that is beneficial for our well-being. If for example we are confronted with a raw sensory stimulus, say, a farmer’s market, there are two ways in which we can use the term “perceive”:

- We can understand *perceive* as experiencing an uninterpreted stream of data: sound-waves, brightness and color hue values that are attached to a place within our visual field et cetera. Understood this way, the object of perception is something like “green” or “loud”.
- We can also understand *perceive* as experiencing the presence or absence of a particular object or quality: whether an apple is there or not and whether it is crunchy and delicious or not. Understood this way, the object of perception is something like “apple” or “delicious”.

The relation between perception in the first and in the second sense has an interesting feature: For one and the same perception of the farmer’s market in the first understanding of the term we can have many completely different ways of perception in the second: We can perceive the same farmer’s market as an assembly of market stands, or of persons, or of goods. We can perceive it as the ongoing transaction between farmers and customers. And so on. Each time, the perceived objects are different ones, depending on the stance of the observer – whether she is interested in microeconomics, anthropology, architecture, or just shopping for groceries. Perception in this more abstract understanding seems to be dependent on a number of properties of the perceiver: which beliefs she has about the world, which concepts she has mastered, which desires she has. This is why Roberts calls a perception of an entity of the latter sort a *construal*.

Let me get back to emotions. Emotions are, according to Roberts, concern-based construals. This means that emotions are (like in Prinz’ case) a vehicle for perceiving qualities of objects or situations and that they (*unlike* in Prinz’s case) depend on the person’s concerns, concepts and desires. Fear is the construal of a situation or object as danger to my concerns (the wellbeing of my family, the absence of pain in my body). The constructor will need to be aware of her concerns in a certain way and have a certain understanding of the world in order to undergo an emotion. Undergoing the emotion will further increase her understanding of the situation: if I watch a sad movie, I will have a better understanding of its sadness than when I read a critics article describing its sadness in every detail.

[26] Perceptions can be misleading in that they make us believe wrong propositions. But there is no truth in perception as it has no propositional content. A philosophical classic that helped establish this now generally held belief is W.Sellar’s *Empiricism and the philosophy of mind* (Sellars, 1997)

Example #3: Martha Nussbaum's highly cognitivist theory

As the following section is about show, Roberts's as well as Prinz's account of emotion are compatible with Haidt's. In order to highlight that there are concepts of emotion that are NOT compatible with the SIM and the MFT, I chose Martha Nussbaum's cognitive theory of emotion. She emphasizes the fact that emotions necessitate beliefs and complex attitudes about the world much more heavily than Roberts. According to her, perceptive theories fail to appreciate several features of emotions that she regards as essential: "their aboutness, their intentionality, their basis in beliefs, their connection with evaluation. All this makes them look very much like thoughts after all(…)" (Nussbaum, 2004, p. 190). Aboutness, the property of emotions that they are –unlike moods, for example- generally directed at an object, that they incorporate an evaluative attitude towards concrete objects and situations is seen by her to be an important point in favor of regarding emotions as being a kind of judgment, a kind of thought, rather than just a reflex-based somatic change.

Nussbaum concludes that complex beliefs are not only causes but "constituent part of the emotion (which has non-belief parts as well)" (Nusbaum, 2004, p. 190). It is this highlighting of very complex beliefs and not just concept-mastery as a part of emotions that makes her approach more extreme than Roberts's:

> "In order to have anger, I must have an even more complex set of beliefs: that there has been some damage to me or to something or someone close to me; that the damage is not trivial but significant; that it was done by someone; that it was done willingly, that it would be right for the perpetrator of the damage to be punished. It is plausible to assume that each element of this set of beliefs is necessary in order for anger to be present: if I should discover that not x but y had done the damage, or that it was not done willingly, or that it was not serious, we would expect my anger to modify itself accordingly or recede."
> (Nussbaum, 2004, p. 188)

We can therefore classify Nussbaum's as a much more radical example of cognitive theories than Roberts's. She emphasizes not only the concept-dependence of emotions, but postulates that in order to undergo emotions, one necessitates beliefs of in some cases extreme complexity, and that undergoing an emotion also is a way of making a judgment about the world, mainly in the form of " acknowledgements of neediness and lack of self-sufficiency" (ibid., p.185).

In sum, the frontline between cognitive and perceptive theories can be described it in the following way:

- For perceptive theorists, emotions are essentially **felt**, while for cognitive theorists, the felt component is only one of several constituents of emotions.
- For perceptive theorists, emotions do not necessitate mastery of **concepts** or **beliefs** about the world, while for cognitive theorists, without desires and concepts (understood in a very liberal way) there can be no emotion.
- For perceptive theorists, emotions are just a vehicle of perception and only *raw material* for our **understanding** of the world, for a cognitive theorist, undergoing an emotion already means to *gain understanding* of the world.

The fundamental differences between Roberts's and Nussbaum's position however show just how much diversity there is to be found within the class of cognitive theories of emotion: how many claims are actually shared with perceptive theories is a matter of degree, and there are cognitive theories that come in many ways very close to perceptive theories (like Roberts's) while others can be regarded as opposed to perceptive accounts in most if not all points. Now that several differing philosophical standpoints on emotion have been roughly introduced, I proceed to the next section in which I demonstrate that while the two philosophical theories suggested by Roberts and Prinz work well with Haidt's theory, Nussbaum's theory of emotion does not. This then leads to the conclusion that Haidt's theory presumes at least the falsehood of strongly cognitive theories of emotion.

3.2.2. *Emotion* in philosophy and in the paradigmatic theory

In order to assess the fit of the different philosophical approaches with the example models and the example study, I check whether the philosophical concepts of *emotion* are in accordance with three features of *emotion* that are essential for the role they play in the SIM and the MFT. These features are:

- Moral emotions are a kind of intuition – that means they are effortless, uncontrolled, unintentional and the person undergoing the emotion is not aware of the cognitive steps that lead from the perception of a trigger stimulus to the emotional reaction. (Emotions as a form of intuition, see chapter 2.3.2)

- Whether an action has emotional valence for a person can be measured by asking the person whether that action bothered her. (Operationalization of emotionality in the example study, see chapter 2.4.1)
- Emotions are to be understood as an associative coupling between representations of somatic states and representations of external entities. (Somatic Marker Hypothesis, see chapter 2.3.1)

On the next pages, I assess whether the presented philosophical approaches agree with these features that Haidt's understanding of emotion postulates.

Intuitiveness of moral emotions

In MFT and SIM, emotions can *always* be regarded as intuitions (as I argue in 2.3.2). A concept of emotion must therefore be compatible with the intuitiveness of moral emotions in order to be compatible with Haidt's standpoint. That means that *emotion* has to be understandable as uncontrollable, unintended, effortless and process leading up to the emotion as not being accessible to the person undergoing the emotion.

- Prinz's account is per se completely independent of concept mastery or anything close to "reasoning" or conscious processing. In fact, it makes conscious processing about danger *dependent on* our capacity to undergo the emotion of fear.
- Roberts's account includes mastery of concepts and certain beliefs about the world as conditions for the capability to undergo emotions. Concept mastery is however the condition of a lot of intuitive processes and should therefore not be considered to stand in the way of recognizing emotions understood in Roberts's way as intuitions in Haidt's sense.
- Nussbaum's formulation of what is supposed to be part of an emotion does however explicitly include conscious processes into her definition of emotions – at least in *some* cases. Especially the part quoted in the previous section about complex beliefs being criterion for whether one is undergoing an emotion suggests such an interpretation, as this implies there being several steps to emotion, involving appreciation of the situation and one's own needs and capacities. In the case of her understanding of grief, one can prevent feeling it by refusing to accept that one has undergone a loss (intentionality of grief). One is aware why one is feeling grief, what the object of one's grief is and what about the object is the source of

one's grief (awareness). One can sometimes be diverted from grief by directing attention to other thoughts (efficiency). Therefore, in Nussbaum's case the constraint of the intuitiveness of moral emotions is **not** met.

Operationalization of "affect"

A second constraint would consist in the operationalization of "affect" in the exemplary study via the question "Would this bother you?" - here one would be able to grant all understandings of the term constraint satisfaction:

- Prinz's account would tie the meaning of "to bother" to the undergoing of emotions. The very meaning of "bothering" would be "eliciting evaluative emotions".
- Roberts's account would understand "bothering" as concern-related concept that would thusly apply exclusively for concern-based-construals or concern-based-beliefs. In the former case, the connection to emotions would be direct in the sense that "bothering", as in Prinz's understanding, would mean something like "eliciting evaluative emotions" – in the latter case, the connection would be indirect, but still given, with "bothering" meaning "touching your personal interests" and emotions being dependent on one's personal interests.
- Even Nussbaum explicitly acknowledges that there are constituents of emotions that are "non-belief parts". These non-belief parts seem to make out exactly what is asked for by the *Bothering*-question which would be the evaluative, caring, and affective aspect of emotion. The *Bothering* question would aim exactly at this part of emotion that leads Nussbaum to call emotion *upheaval of the soul*.

The Somatic Marker Hypothesis

The probably most refined constraints should be set by Damasio's Somatic Marker Hypothesis that stays surprisingly neutral towards the question of what is actually linked *to* the somatic marker. The main focus lies on the process of generating somatic markers - dispositions to show or "simulate"

certain somatic reactions as response to mental presence of certain perceptions or concepts. This feature is shown however by *both* Roberts's and Prinz's theory:

- In Prinz's case, the coupling happens pretty directly between the mental presence (through perception) of an entity and the consequent somatic response.
- In Robert's case, the somatic response is coupled to a rather complex network of beliefs, concerns and percepts, but nonetheless a necessary condition for undergoing an emotion.

How can the SMH be in accordance with both philosophical understandings of emotion? Let me answer this question by highlighting a very interesting feature of his account of emotion that he suggests in his book "Descartes's error" (Damasio, 1994): distinguishing between what he calls *primary* emotions and *secondary* emotions:

- **Primary emotions,** like fear of darkness, or fast moving objects are the "basic" set of somatic responses that we are determined by nature to show towards given simple stimuli like size, span or certain types of motion:

  > "Note that in order to cause a body response, one does not even need to 'recognize' the bear, the snake, or eagle, or such, or to know what, precisely, is causing pain. All that is required is that early sensory cortices detect and categorize the key features of a given entity […], and that structures such as the amygdala receive signals concerning their *conjunctive* presence. A baby chick in a nest does not know what eagles are, but promptly responds with alarm and by hiding its head when wide-winged objects fly overhead at a certain speed. "
  > (Damasio, 1994, pp. 131)

- **Secondary emotions**, like fear of a stock market crash, or jealousy of any kind, make use of the same *mechanism* (building of somatic markers) as primary emotions. However, the kind of representation connected to the somatic marker is completely different:

  > "The process begins with the conscious, deliberate considerations you entertain about a person or situation. […] At a nonconscious level, networks in the prefrontal cortex automatically and involuntarily respond to signals arising from the processing of those images. This prefrontal response comes from dispositional representations that embody knowledge pertaining to how certain types of situations usually have been paired with certain emotional responses, in your individual experience. […] Nonconsciously, automatically and involuntarily, the response of the prefrontal dis-

positional representations […] is signaled to the amygdala and the anterior cingulate."
(Ibid., pp. 136.)

Roberts's understanding of emotions fits especially well with Damasio's description of secondary emotions: as a first step, situations and objects have to be recognized (understood, if you like) as belonging to a certain class of object. Then, the somatic response is triggered.

Prinz's understanding of emotions does neither really fit exclusively to the concept of primary emotions nor does it really fit to the concept of secondary emotions. Prinz explicitly regards the "understanding and judging"-component of secondary emotions as redundant (Prinz, 2003, p.9): His approach does abstain from conscious mental presence or fitting into categories as a step for undergoing emotions. However, the aspect of secondary emotions, that some (for example moral) emotions are dependent on learning the *correct* eliciting stimuli, is present in his understanding of emotions, too. This is why it seems to me that Prinz's understanding of emotions can be understood as

(1) an extreme case of Damasio's understanding (he basically does not postulate any entities that Damasio would not say exist and he does not say Damasio is making any mistakes apart from making his definition too liberal)
(2) Incorporating both "kinds" of emotions in respect of there being innate somatic markers and learnt somatic markers.

Nussbaum's approach does not really fit with Damasio's model: her insistence on the propositional component (what the emotion "says about the world") as identity criterion does not really fit to the idea that the coupling of somatic response and eliciting stimulus are what defines an emotion. In the end, it is the specific somatic response that serves as the ultimate identity criterion for a specific emotion (like sadness, anger, or fear) for Damasio, while Nussbaum regards other, cognitively accessible aspects of the stimulus as the decisive factor , for example in the case of grief the insight of not being able to do anything against a heavy loss that is in fact realized as a heavy loss (Nussbaum 2004, 184f.). Even though she admits the existence of a phenomenal component of emotions, this component is neither the most important feature of an emotion (that would in her view be the realization of self-insufficience) nor suitable for differentiating different emotions properly (for this, she would regard the *kind* of self-insufficience better suited). This makes her account more or less incompatible with Damasio's account.

In the light of these considerations I conclude that Roberts's and Prinz's accounts of emotion fit very well with Haidt's approach, while Nussbaum's has to be rejected by Haidt as a) neither appreciating the somatic component of emotions nor the implicit character of involved beliefs and b) allowing for

non-intuitive emotions. Haidt can therefore be regarded as committing himself to a perceptive or minimally cognitive theory of emotion. Note that this is not an empirical finding of Haidt's but presupposed in his use of the concept *emotion*. This can be shown by substituting a strong cognitivist understanding of emotion into the example validation of his theory. Once this is done, Haidt's theory that excludes strong cognitive concepts of emotion loses its empirical support. Haidt's theory can therefore not be regarded as *showing* that strong cognitive accounts are wrong, but as *presuming* it.

3.2.1. Effects of a strongly cognitive concept of emotion on validation of SIM

What would we have to think about the exemplary study if we understood *emotion* in the way Nussbaum does? Most importantly, it would not be possible anymore to derive from the SIM an important prediction for the example study - that *Bothering* will be a better predictor for morality of moral judgment than *Harm*. Remember that the emotional character of judgments was tested through the *Bothering*-question. Emotionality in the form of affirmation of *Bothering* is seen as an indicator of an intuitive judgment. This is the key to deriving the fifth prediction for the outcomes of the example study from the SIM - it allows for drawing the conclusion that people will tend to be bothered when moralizing a judgment from the premise that moral judgments are intuitive.

This piece of thought clearly does not work if one adopts Nussbaum's idea of emotions. Earlier I showed how she explicitly makes the point that judgments that bother the judging can indeed be conscious acts – as in the case of anger at somebody for a certain action of hers that vanishes when we understand she did not commit the act in question by purpose. This however stands in direct opposition to Haidt's close coupling of moral emotion and moral intuition: moral intuition explicitly prohibits "conscious awareness of having gone through steps of searching, weighing evidence, or inferring a conclusion". Under the assumption of a strongly cognitive view of emotion, the sentence that the emotionality of a judgment is an indicator for its intuitiveness is not valid and prediction (5) ("*Bothering* will be a better indicator of moral judgment than *Harm*") cannot be derived from the model anymore. Without the capacity to derive that prediction the SIM cannot be supported by the exemplary study anymore – the rest of the predictions can be derived solely from the MFT.

I conclude therefore that the validation of the SIM through the example study does in fact rely on a certain philosophical standpoint about emotion, namely the idea that emotions do not incorporate conscious steps of inference or taking conclusions. Haidt can therefore be regarded as making a phil-

osophical commitment – he assumes a rather noncognitivist account of emotion, accepting some but certainly not all philosophical interpretations of the term *emotion*. This is direct evidence for my first hypothesis, namely that psychological theories adopt philosophical standpoints and make philosophical commitments. Let me show in the next chapters how this type of philosophical commitment affects the psychological interpretation of the example study as well as what conclusions empirical philosophy can draw from the example study.

3.2.2. Effects of meaning of *emotion* for empirical philosophy

Example #1: Greene's differing philosophical commitments lead to differing philosophical conclusions

The first example for influence of philosophical commitments in moral psychology on philosophical conclusions comes, just like the one in the intuition chapter, from Joshua Greene (Greene, 2007). According to him, by showing that deontological judgments are emotional judgments (which is supported by the SIM, as shown in the last chapter), he has shown that deontological[27] judgments are not only fixed input-output pairs, but also prepared by evolution - purely primary emotions, in Damasio's terms. Greene takes this claim and therefore the conclusions that follow from it to be supported by the exemplary study (that shows *Bothering* to be a better predictor of moral judgment than *Harm*). Note how this understanding of emotion is very different from Haidt's. The complete innateness of moral emotions postulated by Greene is in direct conflict with Damasio's somatic marker hypothesis that assumes that many emotional dispositions rely on *learnt* associations. (Damasio, 1994, pp.136) It is conflicting with both the SIM that assumes that moral intuition can be influenced by social interaction and the MFT that emphasizes the diversity of moral judgment as an indicator of the socially constructed nature of concrete moral norms. Note furthermore how Greene's different *conceptual* account of emotion completely changes the way the exact same *empirical* evidence is interpreted.

In the course of the argument that employs this alternative concept of emotion, Greene furthermore cites evidence that our deontological judgments take inputs into account that we actually think one

[27] Once again, I adopt Greene's particular style of using the terms *deontological* and *consequentialist* that I explained in the metaphilosohical part of the section about intuition. As its validity plays no role for the point I want to illustrate with Greene's example, I refrain from discussing it in this thesis.

should not care about when assessing the moral value of an action, like how physically close to us something is happening. The closer it gets, the stricter is our judgment. As moral judgment is emotion-based, our moral emotions are triggered and enforced by factors that should be irrelevant to moral judgment. Joshua Greene hypothesizes that this is very likely a result of changing circumstances between the time evolution "built" the emotions responsible for deontological judgment (small groups of hunter-gatherers) and today (big interconnected but anonymous communities). Evolution favored increased compassion with group members, because they were much more important for your own survival – this is why hurt done to people closer to us evokes stronger emotions and consequently stricter moral judgments .

These considerations suggest the conclusions that deontological judgment is a) constant over long periods of time, due to its emotional, genetically predisposed nature and b) not a good heuristic for choosing a path of action in our current environment (since its usefulness from the early days of mankind is outdated by now). Consequently, not only deontological judgments should be treated with a lot of suspicion, but also deontological philosophy should be regarded as not being about rights and duties but about justifying hard-wired dispositions to show reflex-like emotional reactions to certain stimuli that bring inevitably bad results.

An example that Greene uses to illustrate this view is the typical human longing for retributive punishment, which he sees as both a classical deontological standpoint and a typical emotion-driven kind of judgment:

> In other words, the emotions that drive us to punish are blunt biological instruments. They evolved because they drive us to punish in ways that lead to (biologically) good consequences. But, as a by-product of their simple and efficient design, they also lead us to punish in situations in which no (biologically) good consequences can be expected. Thus, it seems that as an evolutionary matter of fact, we have a taste for retribution, not because wrongdoers truly deserve to be punished regardless of the costs and benefits, but because retributive dispositions are an efficient way of inducing behavior that allows individuals living in social groups to more effectively spread their genes.
> (Greene, 2007, p. 71)

> The arguments presented here cast doubt on the moral intuitions in question regardless of whether one wishes to justify them in abstract theoretical terms. This is, once again, because these intuitions appear to have been shaped by morally irrelevant factors having to do with

> the constraints and circumstances of our evolutionary history. This is a problem for anyone who is inclined to stand by these intuitions, and that "anyone" includes nearly everyone. (Greene, 2007, p. 75)

One could summarize this line of thought in the following way: deontological judgment is emotional judgment. Emotional judgment is understood to be a genetically inherited trait. Therefore deontological judgments are the same as in the early times of human development. We can conclude that deontological judgment is tailor-fit to the environment of early mankind. Our present day environment is not the environment that deontological judgment is "designed for". Therefore, deontological judgment in our present day environment will lead to much worse outcomes than it used to. Therefore, it is rational to discard deontological moral judgment as a "blunt biological instrument".

What would Haidt, Prinz or Damasio say about this? Remember the distinction between primary and secondary emotions that Damasio introduced and that is generally accepted by Haidt and Prinz(in Prinz's case with the minor qualifications discussed in section 3.2.2): We have *learnt, rather cognitive* secondary emotions on the one hand and we have *innate, rather noncognitive* primary emotions on the other. If we now look at what Greene is telling us about emotions in general, we see that concerning the role of emotion in moral judgment he should, in Damasio's point of view, be referring to a rather secondary understanding of emotion (Damasio 1994, 134ff). The stimuli that elicit moral emotions have to be *learnt*. This is the case in perceptive theories as well as in cognitive theories. It is also one of the key points of the MFT. Greene however makes it clear that he understands emotion – including moral emotion- to be genetically preset –just like Damasio's primary emotions. Only this type of emotional reaction can be described as tailored by evolution in a sensible way.

We should by now see that given Damasio's understanding of emotion, the property of being innate like for example fear of big objects coming at you with great speed cannot apply to moral emotion: moral emotions can hardly be innate and evolutionarily acquired, because the acquisition of the stimulus part (the *elicitation file*, as Prinz would call it) of an emotion is considered by all three of the researchers to be a form of *learning*.

Assuming a concept of emotion that takes the difference between primary and secondary emotions into account would turn Greene's argument inconclusive: Moral emotion and evolutionarily preset emotion are two different types of emotion. Accordingly, one would have to change the first sentence of the summarized argument into "Deontological moral judgment is $emotional_{sec}$" and the second into "$Emotional_{prim}$ judgment is understood to be a product of evolution". As a consequence, one cannot derive deontological morality to be a purely genetically inherited anymore. It is merely his very particular concept of emotion that allows Greene to draw his conclusions. These conclusions

may be backed by a lot of scientific results – but only under the assumption of a highly debatable conceptual presumption.

I hope that this case nicely illustrates the importance the concept of emotion as a primitive for defining morality can play. Greene's concept of emotion contradicts the one of philosophers like Prinz, psychologists like Haidt and neuroscientists like Damasio – and so does therefore his philosophical conclusion that it would be rational to discard deontological moral judgment[28] as well as his psychological interpretation of the example study. This incompatibility is the result of radically differing philosophical commitments. But even quite subtle conceptual nuances can decide whether it is legitimate to make certain inferences or not. I demonstrate this in the upcoming section:

Example #2: Jesse Prinz's special concept of emotion allows for special interpretation of Haidt's results

In an attempt to build a bridge between empirical science of morality and moral philosophy, Jesse Prinz argues for a sentimentalist theory of morality. According to his approach, the very meaning of the term *moral* is dependent on the disposition to show certain emotional reactions to the object of our moral judgment, for example stealing. *Moral* thereby becomes a term that attributes a response dependent property, just like *funny*, *delicious* or *loud* (Prinz, 2006, p.34). Whether an entity outside the perceiver's body has a certain property is dependent on the perceiver's bodily response to the entity.

If one chooses to adopt Prinz's radical concept of emotion, empirical support for this idea of the meaning of *moral* can be drawn from the SIM and MFT[29]. Remember that according to Haidt, moral

[28] Let me at this place remind you that it is not my primary intention to show that Greene is wrong, but only to show how his philosophical conclusion hinges on his concept of emotion. Let me confess here that I am personally convinced that he is wrong. But others, among them much more accomplished philosophers than me, have already showed this to be the case in abundance (for example Kahane 2012).

[29] Prinz does not in fact deduct this hypothesis from the SIM and the MFT, only from the database that supports them. The reason for this is that according to Prinz (2007, p.99), Haidt postulates a causal relationship between moral emotion and moral judgment, while he himself postulates that emotions are an integral part of moral judgments. A moral judgment without involvements of emotions at least on the conceptual side cannot be called a moral judgment. I interpret Haidt as indifferent about this relationship. As Haidt regards moral intuition as a process and moral judgment as a form of belief, moral judgment can very easily be interpreted as being the causal result of the pro-

judgment is determined by moral emotion. The domain of moral emotions within a culture determines the domain of morality within that culture. Remember that emotions are elicited through an external stimulus. Haidt's moral theory stayed neutral to the question whether a classification of that stimulus in form of a judgment of concept application is necessary or not for a moral emotion – as long as this judgment is intuitive. Perceptive theories argue that no classification of the stimulus in the form of a judgment takes place. That means that the concept *moral judgment* becomes dependent on the concept *moral emotion,* as the eliciting of the moral emotion becomes the only way to find out whether a situation was morally relevant and whether a judgment was a *moral judgment*. This is a much stronger claim than any of the ones that Haidt makes. Haidt argues that moral judgments are emotion dependent. Prinz argues that the whole idea of what moral judgments are is emotion dependent. Note that this conclusion cannot be drawn if one adheres to a cognitive theory of emotion. Under this assumption, it is necessary to judge a certain situation as morally relevant to undergo a moral emotion. A judgment about moral valence of a situation therefore takes place independently from emotional valence, therefore for cognitivists, the concept *moral* must be independent from the concept *emotion*.

The decision whether to adopt a perceptivist or a cognitivist stance on emotion therefore influences directly how the SIM and the MFT can be interpreted philosophically. Like in the case of Greene's argument from the chapter about *intuition* that interpreted the SIM to be committed to a certain concept of intuition, the inference of a sentimentalism about the meaning of *moral* from the SIM and the MFT is dependent on a certain philosophical interpretation of the term *emotion*. As, unlike in Greene's argument from the chapter about *intuition*, no two incompatible meanings of emotion are confused in this case, the present argument is also legitimate.

3.2.3. Conclusion

Quite a few important points were discussed in this chapter. I hope I succeeded in showing that there is quite a lot of different understandings of *emotion* out there – in philosophy as well as in empirical

---

cess of intuitive inference, but only in the form of a last part within the causal chain called *moral intuition*. Moral judgment can be caused by moral intuition and be part of moral intuition understood in that way. If you like to understand *moral judgment* as *deriving moral attitudes*, moral judgment can be even interpreted as being moral intuition in Haidt's terms. This allows in my opinion for a "Prinzean" interpretation of the SIM as their differences are purely terminological in this point.

science. In the rather metascientific part of this section, I have demonstrated that through his theory and operationalization of emotion in the example study, Haidt commits himself to perceptivist and minimally cognitivist theories of emotion. His theory on the one hand regards emotions as intuitions and on the other hand relies on Damasio's Somatic Marker hypothesis that emphasizes the phenomenal, *felt* component of emotions, excluding strongly cognitive approaches to emotion. His practical test for emotionality of judgments in the example study does not exclude elaborate judgments to be involved in moral emotion. Haidt can therefore be concluded to *presume* perceptivist or weak cognitivist concepts of emotion both in theory and research practice.

The metaphilosophical part of this section demonstrated that the chosen concept of emotion will have direct consequences on what empirical data will tell us for our philosophical perspective on moral judgment. Experimental setups might not endorse what they are supposed to endorse (If one substitutes Nussbaum's perspective on *emotion* into the validation of the SIM), they might endorse something completely different (if we look at the example study with Greene's concept of emotion in mind), and they might allow us to draw additional philosophical inferences by refining the concept (if we look at the example study and Haidt's theory from Prinz's interpretation of *emotion*).

This suggests a conclusion that seems a little unsettling: the actual empirical data taken for itself does not tell us as much about moral judgment as we might have hoped - philosophical arguments that rely on empirical data can in principle be defeated by a priori reasoning about concept use. On the other hand this chapter also illustrated how fruitful a diligent crossdisciplinary argumentation can be: Jesse Prinz's sentimentalism gains a lot of strength from its empirical support. It associates itself with testable scientific theories and provides an explanation of the world that seems more rooted in evidence than others.

One of my hypotheses did however not so far figure in this chapter: I hypothesized that implicit philosophical commitments can undermine empirical research and so far I did not offer much in favor of that thought. It is however covered in the next section which deals with commitments about the syntactical structure of moral judgments. In this segment I show how different moral theories take different stances on what the syntactical structure of moral judgment is – and that these diverging syntactical patterns can be found to figure in empirical research, too. I furthermore establish that these patterns tend to be overlooked – which increases the risk that evidence about differing phenomena is lumped together.

3.3. Structure

There is a puzzling feature about morality that can be very nicely explained by appeal to Machiavelli's "Il Principe", Dante's "Inferno" or (to pick a very recent example) the immensely successful TV series "Game of Thrones" that without a doubt has been at least in terms of displayed viciousness and creativity in the subject of hurting people inspired by the earlier two. The puzzling feature is that while one would without a doubt regard most of the characters depicted there as immoral, we can easily recognize that some characters tend to be even more immoral than others: In "Il Principe", the military leader Cesare Borgia is more morally rotten than the cruel minister Remiro d'Orco that he first puts into place to "uphold the public order" in the freshly conquered province of Romagna by instantiating a terror regime, and whom he then has killed in order for the people of the Romagna to regard him as the one who freed them from the oppressor; in Dante's inferno, we find a hierarchical "ranking" of sinfulness through the circles of hell in which different persons from history are shown to undergo different just punishments for their misdeeds, from being stung by bees eternally to being grilled or even chewed on by Satan eternally; and in Game of Thrones, we find very vicious characters like the brutal and cruel Knight Gregor Clegane or the sneaky and cold blooded Lord Balish to be real nice fellows compared to the psychopathic child-king Joffrey and the just as psychopathic Ramsay Snow who both kill and torture for mere pleasure. It is an interesting question how this is possible: moral judgment is considered to be a judgment about whether to do something or not – and should therefore be rather binary. But if we are able to rank persons – and also acts – according to their wrongness in a sensible way, there must be more to moral judgment than just a do-or-do-not distinction.

In moral psychology, this puzzle is encountered as soon as one considers the actual operationalization of moral judgment in empirical studies: when asking people in a standardized way about their moral judgments one will have to give options how to answer – or one will need a way to classify their reaction if one does not ask but simply observes. This can either be a moral/immoral binary choice (perhaps with a "how sure are you?" scale attached) or a scale indicating "how bad is this, on a scale between 0 and [some number]?". These two judgments are logically quite different from each other – even though we might suppose to ask for the same kind of judgment. The answers to the different kinds of questions about moral judgment require quite different kinds of logical operations to be performed. It would be indeed surprising if these kinds of operations would be performed by the same cognitive mechanisms or, if they were part of the same mechanism, would happen at the same stage of the mechanism. One could in fact assume that both testing methods describe two

different kinds of moral judgment at play implying different kinds of cognitive processes. Let me therefore introduce the two ways of measurement in a nutshell.

On the one hand, one can ask for a binary judgment with a question like: “Is it ok? Yes or No?”. An answer would involve the ascription of a simple quality to an action. Asking for moral judgment in this way resembles logically the questions “Is this thing red?” or “Is this a coconut?”. On a computational level, this kind of judgment is a judgment whether the quality applies for a certain object or not. This is everything there is in terms of options. A plain *Yes* or *No*. And it is this distinct kind of judgment that the probands in the given example study in the first chapter were asked for.

On the other hand, asking for moral judgment in form of a graded rating requires a much more refined way of computation: in order to compute a meaningful scale rating to make sense, one would require an ordering relation.- There would have to be a concept of what it means that “A is better than B”, and this concept would have to fulfil certain criteria, like being transitive and non-circular. The capacity to give a sensible reply to a question about morality on a scale would also necessitate the presence of an idea of what should be at the endpoints of the scale(for example “dying for the sake of humanity” and “genocide”). Leaving aside the endpoint question, the syntactic structure of “A is better than B” is already quite different from the one of “A has the property x”. To apply it I need to set two entities into a relation instead of just affirming a property of a single entity. From a computational perspective, a moral judgment in form of a scale rating would be a sorting task: comparing the object (or action) of evaluation with certain (sometimes idealised) reference objects (or actions) and sorting it into that category or to that position in a list to which the comparison shows the best fit. An example of such a task would be asking to which place in an alphabetically ordered list (for example a bibliography) a new entry (for example a reference to a paper) would have to be put.

If we look at the relationship between both ways of computing a moral judgment, we see that we cannot derive a graded judgment from one binary judgment and vice versa by the means of pure logic- – only when given a certain additional information (for example a threshold or a maximization rule) we can compute a binary judgment with help of the relation that underlies one graded judgment.

A comparable case would be the translation of binary and graded measurements of *heat*. This translation would have to give us a way to transform a degree figure from a thermometer to the affirmation of a perception-related sentence “it is cold outside”. If I want to infer that people will affirm the sentence “it is cold outside” from a given temperature value in degree Celsius, I need some kind of threshold: if the temperature goes below that certain threshold, I can expect people to judge it to be

"cold outside". It is a remarkable property of human beings that this threshold is subject to quite radical changes, but nonetheless – without it, one could not infer a binary temperature judgment from a graded temperature judgment even though both judgments are about the same property (heat). This means that even if I have an exact thermometer value I will need some kind of "translation manual" in order to know whether people will say "it's cold outside". The same thing can be expected for graded and binary forms of moral judgment.

3.3.1. Structure of moral judgment in philosophy

If we look for philosophical positions that endorse one of these two syntactical analyses of moral judgment, we do not have to dig very deep. Already among the most famous philosophers we can pinpoint proponents who would rather tend to the one side and proponents who would rather tend to the other. I chose as examples on the one hand Kant's deontological and on the other Mill's Utilitarian conception of moral judgment. Let me however issue a further disclaimer at this point: I am presenting Kantian and Utilitarian moral philosophy in a very sketchy way in order to give an example of how graded and binary moral judgment figure in philosophical tradition. Do not intend to make any exegetic claims concerning the works of Immanuel Kant, Jon Stewart Mill or Jeremy Bentham. As this is not of importance for my argument but rather for illustrative purposes, is regard this way of presenting Kantian and Utilitarian claims as legitimate.

Example #1: Kantian binary moral judgment

The foundation of the Kantian test for the morality of an action is the famous categorical imperative. "Act only according to that maxim whereby you can, at the same time, will that it should become a universal law" (Kant, 1993, p.30). If any action is prohibited by this universal law and it is performed willingly, this action is hence immoral[30].

[30] There is an additional condition for an action to be truly morally worthy, which is that it is performed out of respect for the moral law and not as an effect of pure affect, not only done *in conformity with duty but from duty* (Kant, 1996: p. 53). Doing good just because one is for example in good mood is not regarded as an achievement. However, even though important for any true understand-

Let me illustrate this: If I am lying to someone, I am using language in order to deceive another person; I am affirming something that I know not to be the case. The maxim behind this action is that I do not use language to share information, but to manipulate people for my personal gain. The very point of language is however information sharing – otherwise there would be no language. In a world in which nobody used language to share information, nobody would have reason to believe in the utterances of another person. It would therefore not be possible to make the maxim behind lying a universal law: that would go against the very idea of language itself. One can therefore conclude that lying cannot be right.

This imperative is a rule that divides actions into two classes: actions that are permissible and actions that are morally impermissible. We can label this test and hence the underlying concept of morality as "binary" without any problems.

Of course, Kant also noticed the curious capacity of ours to make graded moral judgments. In addition to the binary moral-immoral classification given by the categorical imperative, he argues us to have a more fine-grained way of evaluating an action: assessing whether or not it is in line with so called *imperfect duties*. These duties are also in a way given by the categorical imperative and therefore duties, but only in a more indirect way, that allows for flexibility and proportioning in application. Being a valuable part of society for example is an imperfect duty against oneself that can be realized in very different ways. (Kant, 1902, VI-p. 309, IV-p.421) Failing to adhere to imperfect duties is only be regarded as failure to be praiseworthy, but not as an immoral act in itself, as one might be a valuable member of society in one way or another (Kant, 1902, VI-pp. 444). Praiseworthiness can be expected to vary according to how sure of an indicator of adherence to imperfect duties a given act is. The more we can regard a given act to be an indicator of being a valuable member of society, the more praiseworthy it is. An example of an imperfect duty closely reaqlted to being a valuable part of society is friendliness. There can hardly be any doubt that a sensible assessment of friendliness would have to be graded and not binary. But friendliness can look very unalike in different persons or different cultures, and each individual will have another way of being friendly, making levels of friendliness hardly comparable between persons. While the necessity of friendliness is objective, its implementation is highly subjective. This variance in ways to fulfil imperfect duties and the fact that one act is only an indicator of adherence to an imperfect duty makes it impossible to call failure of acting friendly in one single situation a full blown immoral act. Imperfect duties are just too abstract and long-term to regard nonobservance of them in a single situation as a clear violation of moral duty. This makes them in a certain way peripheral when combined with a clear and direct ap-

---

ing of Kantian ethics, this condition will add no additional value to this chapter which is why it has been exiled to this footnote.

plication of the categorical imperative in clear cases, like murder, theft or similar misdeeds. In these situations, imperfect duties are not in themselves relevant for the moral worth of an action, understood in a binary way. It is wrong to murder someone, even though you might be "friendly" to your victim by killing it quickly and painless instead of torturing it to death for hours. But imperfect duties are still helpful in assessing the goodness of an action once we ask for a kind of graded judgment: one would probably agree that being "friendly" to one's victim is after all *better* than being "unfriendly". However, this goodness assessment has no influence over the *overall* moral worth of an action, which is determined by conflicting with the categorical imperative.

Let me get back to the example of lying. I concluded before that lying cannot be right as it would not be possible to make the maxim behind lying to a universal law: that would go against the very idea of language itself.

Imperfect duties will now for example allow us to make a distinction between lying to make somebody feel good and lying to make somebody feel bad: if a doctor wants to motivate a patient with an unhealthy lifestyle to exercise more, she can for example choose to untruthfully tell her that blood levels of, say cholesterol are already improving – or she can try to motivate by untruthfully saying that levels even got worse. From a Kantian perspective one would conclude that (given the motivational force is equal) even though both acts of lying are immoral, the first option is somehow better than the second, just because it is nicer to treat people kindly.

So, imperfect duties allow a Kantian to draw distinctions in goodness, or praiseworthiness, between different immoral acts or between different morally permissible acts. The story a Kantian might therefore tell about rating scale ratings of moral value would be the following: the *actual* moral judgment determines towards which end of a moral judgment rating scale a person orientates when making the judgment through application of the categorical imperative. The *exact* position on the scale is however determined by procedures within our mind that are at best only indirectly *moral* considerations – an assessment of adherence to imperfect duties. Scale ratings are a form of moral judgment, but only one that is secondary to binary judgment. Binary Judgment can therefore be said to be the *actual* or at least *primary* moral judgment.

I conclude that from a Kantian perspective, moral judgment is primarily a binary choice. Graded judgments about the goodness of an action are meaningful, but they are not directly but at most indirectly about the moral worth of an action. From a Kantian perspective one might conclude scale ratings to indicate mainly the fit of an action to imperfect duties, only the very rough tendency towards one of the endpoints of the scale with a threshold somewhere in the middle of the scale *really*

tells us anything about the actual moral worth that the questioned study participant attributes to an action.

Example #2: Utilitarian graded morality

This binary view can be contrasted in a nice way with the Utilitarian picture of moral goodness: here, it is quantitative measures that decide about the goodness of an action. (Bentham, 1996, I-II, pp.11) The most famous formulation of this measure, the "Principle of Utility" or "Greatest Happiness Principle" has been formulated by John Stuart Mill: "the Greatest Happiness Principle, holds that actions are right in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness." (Mill, 1972, p.7). Happiness is of course to be understood as happiness of everyone, not merely the happiness of the agent or one individual.

The decisive point here is that utility of actions can be used to put these actions in a clear ordering relation. In microeconomics, this property of utility as well as the possibility to compare utility of actions and goods with utility of a specified sum of money is what makes the demand part in the derivation of the market price of a good or an action possible. This is the case because in microeconomics as well as in Utilitarianism the goodness of an action is rooted on a utility value that is in itself not binary at all but based on a "more useful than"-relation.

At this point it seems worthwhile to give an example. Lying can be generally considered as detrimental to the overall utility of everyone: if one person lies to another, she is trying to induce a wrong understanding of the world into the other persons mind. Having a wrong understanding of the world is normally detrimental, mainly because acting or not acting on wrong premises tends to yield undesired consequences. It is not implausible to assume that the utility of telling the truth is generally higher than the utility of lying: the potential liar might in the end end up with less than if she lied, but the potential victim would normally benefit more from being told the truth than the liar would gain from lying. If a doctor lied to a patient about her physical health to motivate her to exercise more, we could expect that the overall consequences would be worse than if the doctor told the truth and would motivate otherwise. In the worst case, the patient would find out and the relationship of trust between doctor and patient would be over. In the best case, lying will have had no better consequences than if the doctor had not lied but tried to motivate otherwise. Whether on the other hand the doctor had untruthfully said something that made the patient feel good or bad would make indeed a *moral* difference: Being proud of imaginary achievements is without a doubt a better consequence than feeling depressed because of imaginary shortcomings. A Utilitarian would definitely affirm that also small changes in rating scale ratings imply changes in assumed moral worth of the judged act. A problem however would be how to translate judgments of utility into binary judgments of right and wrong. There are several options, for and against which there are valuable arguments:

A first obvious choice would be to introduce some maximization rule: of all the possible options that we have in a given situation, the best one is the right one, and all others are wrong. This rule is pretty nice and simple, but it has the shortcoming of introducing a very high standard of morality – one that probably cannot be met by anyone. On the other hand, also virtue ethicist and deontologist moral theory works with idealist and unreachable pictures of morality. The worse problem for this translation rule is the following: it seems hardly possible to realistically assume ALL consequences of one's action in a fashion detailed enough to make a judgment concerning which particular action will yield the BEST outcome. While it might indeed be possible to judge whether one action has better outcomes than the other, it would be very difficult to know which choice of action has the *best* outcomes. On the one hand because this would necessitate considering *all* possible choices of action, an almost impossible task, on the other hand it is, as soon as quantitative measures are considered, always possible to do more or better. Just take a beggar who you are giving a five Euro bill. According to a maximization rule, this would be worse than giving ten Euros and therefore not the morally permissible action.

The introduction of some sort of threshold seems to be a more realistic choice than a maximization rule, just like in the temperature case introduced above. But the problem that one would be facing here would be the arbitrariness of such a threshold: neither would it be enough to count an action as morally praiseworthy if it would generate a net gain in overall utility, nor would the utility payoff of doing nothing count as a good threshold. Both would be far too liberal in calling something the morally right choice. Doing nothing is often in itself terrible, so that a little improvement just cannot be enough to make an action plausibly morally alright. A net gain of overall utility is very quickly arrived at, even in situations that one would not regard as morally relevant like when one gives a smile to a stranger. This difficulty in translating graded utility or happiness ratings into a binary right/wrong judgment has led Norcross (Norcross, 2006) to conclude that a translation from graded to right/wrong is not sensible after all. Moral judgment should rather be revised to be actually a graded judgment that we attach labels like "right" or "wrong" to in very clear cases, similar to our practice of saying that it is "hot" or "cold" only in very clear cases.

Coming back to the scale-or-binary-choice-question: if one understands moral goodness as based on utility, one will be able to transfer this relation to a "morally better than"-relation very easily. The greatest happiness principle of Utilitarianism implies a *graded* judgment of moral value – this graded judgment might possibly be transduced into a binary judgment, even though people tend to disagree whether and how this is possible.

The differences between the two approaches

Let me put both pictures together: while Utilitarians regard moral judgment as dependent on our judgment of the (comparative) utility in the form of happiness resulting from an action and therefore as *primarily* (or in extreme cases *solely*) graded or comparative, Kantians regard the judgment of the moral rightness of an act in principle as a binary choice. Both sides have developed theoretical vehicles for translating their idea of the *primary* syntactical form of moral judgment into statements of an alternative syntactical form: For Kantians, the actual (binary) moral judgment is the basis of a more fine grained evaluation of that act based on its accordance to imperfect duties. For Utilitarians, there are basically two choices for transducing an actual (graded) moral judgment into a binary judgment, which is on the one hand introduction of a utility threshold and on the other hand the introduction of maximization rule. In the next section, I demonstrate how this distinction can be mapped on different operationalizations of moral judgment: I introduce a study that makes use of a comparative concept of moral judgment and explain the limits set by this conceptual decision.

3.3.2. *An example of graded judgment in psychology*

The case I chose to exemplify how the syntactic structure of moral judgment affects empirical research is a study by Thalia Wheatley and Jonathan Haidt (2007): here, people were hypnotized to feel a flash of disgust when reading or hearing the neutral trigger words “often” and “take”. Whenever persons were confronted with these words, a flash of disgust was triggered. The stimuli were written in a way that made them contain the trigger words in order to evoke disgust in the hypnotized participants. This evoking of disgust led to a significant change in the moral evaluation of situations that were given by making an indication on a scale with the end points “Not at all morally wrong” and “extremely morally wrong” in comparison to non-hypnotized control subjects. This result is interpreted by Wheatley and Haidt as a confirmation that disgust plays an important causal role in moral judgment.

We can recognize effortlessly that this study and its interpretation make a philosophical commitment about the syntactical structure of moral judgment: it is supposed to be evaluable through a rating

scale which implies the existence of a "morally better than"-relation with more than two equivalence classes[31].

A Kantian would have to refute the result that disgust modifies moral judgment, as the study has a fundamentally different idea of what moral judgment (or a change of moral judgment) is. For a Kantian, the best explanation of scale rating effects would be that scale ratings reflect the rating changes in terms of imperfect duties. A change in judgment about imperfect duties is however not a change in the actual (binary) moral judgment of the displayed behavior.

A Utilitarian on the other hand would have no problems with Wheatley's and Haidt's result: as the morality of an action is indeed assessed in a graded way (via "utility" or "goodness"), their conclusion is absolutely sensible. It is even perfectly in line with the rather Utilitarian point that moral judgment is less a rational choice but more a decision that "feels right" (Mill, 1972, pp.28).

The point that I want to make here is that different concepts of *moral judgment* can imply different takes on syntax and computation of moral judgments and vice versa. It is therefore extremely important to be clear about what one wants to be understood as moral judgment and how one is measuring it. Otherwise, two measurements that are allegedly of the same phenomenon, but are in fact not, could be thrown together, producing a bunch of possibly incongruent data. Unfortunately, this is exactly what happens in psychological practice: both ways of judging are thrown together again and again when citing literature in favour of one's own point – for example by Haidt and Wheatley who relate their study to the results of the (binary) example study. It would be very important to increase the clarity of such remarks in order to not getting things twisted.

3.3.3. MFT and SIM presuppose binary judgment and cannot predict graded judgment

I mentioned earlier that the example study, which asks for binary moral judgment, is regarded to be about the same kind of judgment as the hypnosis study, which asks for graded moral judgment. From my considerations unto this point it follows that this cannot be regarded to be the case as long as there is no rule for translating graded moral judgments into binary moral judgments.

[31] Which means that there is more to morality than actions or behaviors being „just right" or „just wrong" – there are many more possible evaluations in between those poles

It is therefore time to analyse the MFT and the SIM in regard of two questions: What is the syntactical structure of moral judgment presumed by Haidt's moral theory? And does it offer us a way of translating one type of judgment into the other?

The answer to the latter question is: no, there is no translation manual for relating binary judgments to graded judgments. But as moral judgment is dependent on emotion, it seems worthwhile to explore the possibility of introducing an "emotionality scale": after all, emotions consist in action guiding somatic states that are triggered by external stimuli. Is it possible to introduce a scale for emotions as understood by Haidt? In some sense, the answer should very likely be: "yes". Somatic changes can come in different intensity, and this intensity could be rated on a scale. Take for example the emotion of disgust. The somatic changes associated with disgust can go from a minor shiver to full-blown nausea to even more than that. On the other hand, emotions play the role that they drive action and attention in a certain direction. Either they do so successfully or not. Whether or not an emotion is in the end attention and action guiding and the decisive part of our judgment of a situation is a binary question: either attention is guided or not – either action is guided or not. The MFT attributes to moral judgment and moral emotion in general a clear purpose: omission of (biologically) detrimental behaviour, enforcement of (biologically) beneficial behaviour. This role in my opinion suggests a binary role of emotion in moral judgment. Either emotion is "on" and causes a certain bias, or it is "off" and does not[32].

Furthermore, the MFT regards moral judgments as judgments according to norms that are taken the most important in a given society. These norms are about either displaying certain virtues or showing certain types of behaviour. Concordance to a norm is hardly expressible as a graded judgment. Either I refrain from stealing or not. Of course one might assume norms as idealizations of character prototypes like "be like Achilles/Francis of Assisi/Emperor Yu". In this case, closeness to the unreachable normative ideal state could be rated on a scale. However, norms of that character would be very hard to fit with the sanction-criterion introduced in the SIM. Remember that in order for a norm to be a moral norm, noncompliance needs to be sanctioned by society. Noncompliance to an ideal standard is however hardly a possible foundation of sanction. Therefore, also the norm compliance

[32] One could object here that if it is possible to undergo several emotions at once, it could be possible to undergo two emotions at once that pull in opposite directions. To map this "being pulled in two directions" and the degree to which one side wins, a scale would indeed seem apt. I would like to refute this objection by pointing out that if emotion as a concept is dependent on an action bias, it is the action bias that decides about which emotions one is truly undergoing in the end. As one cannot be biased to do and not to do something (by definition – otherwise it would not be a bias), the idea of undergoing two emotions pulling in opposite directions at the same time is in conflict with Haidt's very idea of emotion.

aspect of morality that can be found both in the SIM and the MFT indicates a commitment to a binary view of morality in Haidt's moral theory.

This makes it impossible to derive predictions in rating scale terms that go beyond "tendency to the *morally right* end of the scale" or "tendency to the *morally wrong* end of the scale". If there is a negative moral emotion, there is a negative moral judgment. If there is a positive moral emotion, there is a positive moral judgment. Change of moral judgment should correspond with a change of emotion and change of emotion with a change of moral judgment. But the changes here should not be expected to be gradual but categorical: If the judgment changes from pro to contra, the emotion should change from positively valenced to negatively valenced – and vice versa. A gradual change like in the Wheatley and Haidt-study just cannot be predicted from the MFT and SIM alone. It is therefore a mistake to assume that Wheatley and Haidt's study supports the SIM or the MFT or that the SIM or MFT would predict the results that can be found in Wheatley and Haidt's study.

3.3.4. Effect on Jesse Prinz's argument for emotional nature of morality

The same issue can be spotted in regard to the philosophical use of empirical results. The main threat to philosophical arguments posed by commitments about the structure of moral judgments should be clear by now: if no way or method is offered how to get from a moral judgment with one syntactical structure to a moral judgment with another, a philosophical conclusion is only valid if it assumes the very same syntactical structure of moral judgment as its premises. Otherwise *moral judgment* in the premises means something else than *moral judgment* in the conclusion. Please note that once again, I have turned towards metaphilosophy.

An example of getting things twisted in philosophy is an empirically based argument that Jesse Prinz offers in one of his papers about the necessity of emotions for moral judgment (Prinz, 2006): here, he refers to scale-results (for example the study above) as evidence for his moral theory that strongly suggests a binary concept of moral judgment –he compares judging wrongness to judging whether "cherries are red", a prototypical binary judgment (Prinz, 2006, p. 34). Even if one looks very thoroughly, one will not be able to find an explanation how graded moral judgments relate to binary moral judgments and *how exactly* the empirical data is supposed to support Prinz's idea of a binary moral judgment. At least in this particular publication, one could therefore accuse Prinz of making an inconclusive argument by offering us a non-sequitur as a conclusion. Please keep in mind that I am

not arguing against his conclusion, I am just stating that the argument is flawed, especially if the argument aims against Kantian theorists who could not be addressed using rating scale results as arguments anyway and who might be the only ones really shocked about emotions playing a part in the evolvement of moral concepts in the first place.

3.3.5. Conclusion

This chapter aimed to highlight how subtle and easy to overlook philosophical commitments in psychology of moral judgment can be. I explained how *moral judgment* can be understood differently in regard to its primary syntactical structure through the examples of Utilitarianism and Kantianism and demonstrated that that these diverging understandings can both be found in psychological practice. Problems that result for psychology and any empirical investigation of moral judgment were highlighted by showing how predictions that are assumed to be derivable from the MFT and SIM in fact are not due to implied commitments about syntactical structure of moral judgments. I furthermore introduced the thought that this situation could be defused through the introduction of a kind of "translation rule". It turned out that without such a translation rule, neither a confirmation relation between study and empirical model nor a justification relation between premise and conclusion of an argument can be sensibly upheld. What such a translation rule could look like is the subject of a later chapter of this book (chapter 4.3).

This kind of philosophical presumption is one of the trickier ones that I am presenting in this dissertation. It is extremely easy to overlook which gives it a lot of damage potential, especially since it is practically unavoidable. And it is (just like the value transitivity of norms in the case of virtue ethics that figure in the next section) a presumption that also trained philosophers are not going to come up with to begin with because it does not really play much of a role in philosophical discourse about moral judgment.

The next section figures a similarly tricky aspect of *moral judgment*: What is it that is actually judged? Is it actions or rather agents? Just like in the case of structure, this does not seem to make much of a difference on first sight – but allow me to convince you that it does. I demonstrate how choosing the take on what is judged can undermine empirical support and in the worst case the coherence of a theory itself. In this section, I furthermore have the opportunity to investigate the validity of a philosophical argument that stems from Jonathan Haidt himself. This is why the pattern of the next section is not as clearly separated into a metascientific and a metaphilosophical part, but rather oscil-

lates between both. Of course I still try to be as clear as possible about what kind of observation I am making at which point. However, especially in this section, the borderline between both disciplines proves to be pretty blurry.

3.4. Object

It is a very interesting feature of religion (which is quite regularly regarded as the single true source of morality[33]) that divine commands alone are often not seen as giving enough guidance on how to get by in one's daily life. The most institutionalized form of fixing this and providing additional ways of deciding what is right and what is wrong independently of direct divine commands can be found in Islam: here, written records of the life of prophet Mohammed called *hadiths* take an important role in defining moral and religious duties. Accordingly the *hadiths* are ranked very carefully concerning the likeliness of their correctness. The idea behind the importance of the *hadiths* in Islamic faith is that Mohammed, as one that is known to have certainly lived a morally and religiously valuable life, can without any doubt be regarded as a role model for any faithful person. In a doubtful situation where the divine commandments from the Quran cannot offer clear guidelines, the question how Mohammed would have dealt with the situation offers moral guidance. A similar, even though not as strongly institutionalized practice can be found in Christianity, where the accounts of the life and the behaviour of Jesus Christ are just as important as God's direct commandments for determining the morally right choice in a given situation. The conflict about the poverty of Jesus Christ and his apostles between pope John XXII and the Franciscan order in medieval times made famous by the movie "The Name of the Rose" was for example considered to entail massive normative consequences, for example concerning the issue whether the church should be allowed worldly possessions or not. The right thing to do is what Jesus and the apostles have supposedly done.

Two of the largest of world's religions make use of role models to give guidance in daily life. And indeed, role models often offer the advantage of being handier than mere rules, norms and laws, since they seem easier applicable to specific situations in life. From a philosophical point of view, they are particularly interesting as they offer an alternative type of justification for moral judgments compared to general law-like principles. Instead of deriving an imperative from a law-like general principle, one derives imperatives from the character traits of a role model figure. There has been quite a

[33] An opinion that I do not share but the existence of which I find quite useful to mention here

bit of philosophical discussion about which way is the more efficient, elegant and suitable to justify moral judgments.

Given these considerations, the focus on judgments of acts in the exemplary study can be regarded as a prototypical self-imposed philosophical constraint. The object of moral judgment is regarded to be a particular *kind of act* committed by somebody. To adopt Prinz's way of speaking, the "elicitation file" of a moral judgment would be a *kind of act*, instead of, say, a *kind of character trait*. As this section is to demonstrate, the example study is in good company here: many proponents of traditional western philosophy focus on the goodness of kinds of acts when assessing the goodness of an action.

However, as soon as we look at the classical psychological literature about classification and judgment of behaviour in general, we find very plausible alternatives, for example the account of behaviour classification and explanation provided by Nisbett and Jones as early as the Seventies (Nisbett & Jones, 1973). According to that approach, one will tend to explain one's own behaviour rather by the means of circumstances while the behaviour of others is rather explained by traits of character that are attributed to them. This finding - that explanation of others' behaviour relies mainly on attributing character traits - suggests that instead of a cluster of elicitation files with the label "bad kinds of actions", one might as well postulate a cluster of elicitation files with the label "bad kinds of character traits" that serves as our moral compass. And indeed there is a philosophical approach to moral judgment that could under certain conditions be combined perfectly with this psychological theory: *virtue ethics*.

In the present chapter, I proceed in the familiar way: I introduce virtue ethics and situationalism as two competing philosophical standpoints about the object of moral judgment; I assess which position the theories and the study take on this matter; I point out some empirical findings in favour of virtue ethics; I explain how inserting the noncompatible concept would affect Haidt's rationale and model. Finally, I turn to Haidt's own philosophical argument: I investigate how Haidt's proposed virtue ethics approach will do if we actually assume a virtue ethicist view on his evidence and come to the conclusion that this philosophical view would undermine not only the empirical foundation of his theory, but even the very core idea of the MFT.

3.4.1. The object of moral judgment in philosophy

Examples # 1&2: Kantianism ad Utilitarianism as typical situationalist approaches

As I already mentioned, most classical ethical theories center either on the consequences of or the intention behind specific kinds of acts and less on whether these acts would, in a given situation, instantiate certain character traits.

Immanuel Kant for example focuses on the intention of the actor when it comes to determine the goodness of an action. The Kantian test for the morality of an action, "Act only according to that maxim whereby you can, at the same time, will that it should become a universal law" (Kant, 1993, p.30), the moral law, is a rule about acts. The reason that somebody is judged to have done something immoral is that the committed act was prohibited by the moral law – not that the act instantiates a character trait prohibited by the moral law.

Note that the maxim of an action is not a character trait but rather a sort of intention. An intention attributed to a type of act. Kantian morality works fine without any mention of character traits[34].

The same counts for Utilitarianism: here the guiding principle of morality is the "Greatest Happiness Principle", which the reader might still remember from the last section: "(…)actions are right in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness." (Mill, 1972, p.7). Just like in the case of Kantianism, it is very clear here that it is the *kind of action* that decides about its moral value, not its indication of the moral praiseworthiness of the person committing it.

What do both approaches have in common concerning the object? They agree that the decisive attribute of the morality of an action is independent of the character of the person committing it. It is an *intrinsic*[35] property of that act. This is not the case for virtue ethicist approaches to moral judgment as the next section is to point out.

[34] This is of course an oversimplification. The importance of *virtue* for Kantian ethics remains debatable (Sherman, 1998; Johnson, 2008). For the illustrative purposes served by Kantian ethics in this section however, an oversimplification will do.

[35] A property is intrinsic if it is a property that lies "within" the entity. *Chairs* for example have the intrinsic property that you can sit on them. *Being wooden* is an example for a non-intrinsic property of chairs. Being wooden is however an intrinsic property of tree stems.

Example #3: Martha Nussbaum's Aristotelean virtue ethics

When deliberating about the goodness of an act committed by a certain person, a virtue ethicist will not ask "Is this action permissible?" like an Utilitarian or Kantian would but rather "Is this something a person with a good character would do?". The clue about this approach to ethics is that it is not a property of a type of action that makes an act a morally good act – but that it is the kind of behavior a morally good person would display. Moral goodness is primarily a property of persons, not of actions – judgment of an action as morally valuable is therefore derivative of a judgment of character of the actor, as actions are mere *indicators* of character. In this section I present the virtue theorist account of Martha Nussbaum, whom Jonathan Haidt explicitly presents as a virtue theorist that he is sympathetic with (Joseph & Haidt, 2004).

Virtue ethics in the sense of Aristotle and Nussbaum can generally be traced back to Aristotle's *Nicomachean Ethics* (Aristotle, 1999). Here, Aristotle lays down his theory of what it needs for a person to live a good and successful life[36]. He comes to the conclusion that whether a person is well equipped to live a good life is dependent on the character of that person. This character should show the stable disposition to display certain kinds of behavior that enable a person to live a good life. This global disposition to behave in a virtuous way can be spelled out as a set of specific dispositions to behave in specific ways. These specific kinds of virtuous behavior are "kinds of goodness", *arête,* commonly translated as *virtue* or *excellence* (for example in Aristotle, 1999). These kinds of excellence pertain to specific types of reactions to specific situations: courage, moderation, justice, generosity, hospitality, self-esteem, kindness of temper, truthfulness, friendliness, and intellectual virtues like smartness and wisdom. It is a very important point about these domains of excellence that what for one person may be virtuous or excellent behavior, may be unvirtuous for another person: what would be courage for a strong man can be foolish hybris for a weak man – what is generous for a poor man can be very avaricious for a rich man.(Aristotle, 1999, 1120b) Excellence in the aspects of human life where one needs to be virtuous in order to flourish is therefore person-relative and objective at the same time: while Aristotle regards the importance of courage for a successful life as objective, the issue what *courage* means in specific behavioral terms has to be assessed individually from person to person.

[36] This is a very important difference to modern ethics. While modern ethics focusses on moral behavior, Aristotle's ethics takes a more holistic approach to good life in general. Instead of being a good person, the aim is to live a good and happy life.

Whether an action is good and honorable turns from a question of adherence to generally applicable rules to a question like "would a virtuous person in my position and with my capacities do this?". This has the advantage that some of our moral intuitions, for example about generosity are met: it would be ridiculous to expect the same amount of generosity from a rich and from a poor person – to say that the rich person is better because he adhered to the moral norm that one ought to give to the poor while the poor person did not, strikes one as extremely counterintuitive.

On the basis of this approach to what moral judgment is actually about, modern philosopher Martha Nussbaum makes her own point that is about how to determine whether a given institution is just or not: there are certain aspects of human life about which humans just *have to* make a choice how to deal with them:

- mortality, that is (via our biological disposition to die one day) something that affects every human being
- the body, the vulnerabilities and capabilities of which are very stable across cultures
- pleasure and pain, that have about the same effects on behavior and the same phenomenal appearance everywhere
- cognitive capability, both in the form of the capacity for problem solving and the desire to increase one's knowledge about the world
- practical reason, that is the practice of planning ahead in order to reach certain goals
- Early infant development, which means certain experiences of helplessness and "growing up"
- Affiliation, which means the general tendency to live in groups and to cherish especially intimate relationships to one's loved ones
- Humor, that means the practice of joking and laughing together.
  (Nussbaum, 1993)

To each of these parts of life, we can attribute a way "to choose appropriately in that area of experience" – this is what Nussbaum wants to be understood as *virtues* (Nussbaum, 1993, pp. 246). The choices about one's death, one's body, one's pleasure and pain, one's cognitive capability and so on play the decisive role in organizing human institutions that Aristotle attributes to virtues.

Nussbaum takes this concept of virtue that is integral for living a good life mainly to base her picture of political justice on it: a political system is just if the capabilities to flourish as a human being – in other words to develop the virtues associated with these parts of life – are distributed in a just way. This approach is an alternative to those which regard justice to be about distribution of rights and duties, for example the one that John Rawls argues for in his "A Theory of Justice".

But how does Nussbaum combine *virtue* with *moral evaluation* of particular acts? Mostly indirectly, I am afraid – even though the indirect remarks point to a clear direction. She clearly states that the job of ethics is assessing what positions towards these essential parts of life are the *right* ones. The job of ethics is therefore to define the *prototype of a moral character*. It will be only against this prototype that any moral evaluation can be made. (ibid., p. 247)

The important point of virtue ethics (represented here through the exemplar of Nussbaum's account) for its combination with Haidt's moral theory is the way that virtues are regarded as *individual* traits (that entail differences of standards of evaluation depending on individual or environmental factors) that are pertaining to *universal* areas of application (*domains* in MFT terms, that define the areas in which excellence can possibly be considered morally praiseworthy). It is this way of combining an objective domain with relative virtues that makes Haidt and Joseph regard virtue ethics, and explicitly Nussbaum's brand, as the approach to moral philosophy that best matches the most important feature of the MFT.

3.4.2. Haidt's own philosophical position on virtue ethics

If we look for character in the SIM, we find it in a very explicit form that was mentioned already before: "Moral Judgments are therefore defined as evaluations (good vs. bad) of the actions or character of a person that are made with respect to a set of virtues held to be obligatory by a culture or subculture" (Haidt 2001, p 817). In the case of the SIM, we can therefore assume that moral judgment as judgment of character is indeed a very realistic option. As seen above, the distinction between moral norms and conventional norms even refers to the term "virtue".

When it comes to the MFT, Haidt goes even further. Together with philosopher Craig Joseph he explicitly endorses an account of virtue theory:

> A virtuous person is one who has the proper automatic reactions to ethically relevant events and states of affairs, for example, another person's suffering, an unfair distribution of a good, a dangerous but necessary mission. Part of the appeal of virtue theory has always been that it sees morality as embodied in the very structure of the self, not merely as one of the activities of the self.
> (Haidt, 2004, p. 61)

> Such theories fit more neatly with what we know about moral development, judgment, and behavior than do theories that focus on moral reasoning or on the acceptance of high-level moral principles such as justice.
> (Haidt, 2004, p. 62)

What is the reason for Haidt's embrace of virtue ethics? Most importantly, as the first quote suggests, the idea of describing morality as an assembly of character traits fits better than situationalist accounts to both the intuitive as well as to the domain-specific pluralistic picture of morality that the MFT draws. Character as a collection of dispositions to behave in a certain way captures Haidt's idea of automatic processes as a basic source of moral judgment much more elegantly than the constructs that situationalist ethicists have to postulate in order to account for the capacity of humans to behave morally, such as pure reason (Kant) or the feeling of sympathy (Hume). On the other hand, the description of virtues as domain-dependent kinds of excellence fits very nicely with Haidt's idea of morality as a conglomerate of different moral domains. The domains that Nussbaum postulates differ quite radically from those that Haidt suggests, but the foundational idea of moral goodness being goodness of different kinds fits much better to Haidt's modular view of morality than situationalist accounts that typically try to derive ONE moral principle that will be suitable for ALL situations.

The second and almost equally important point is that especially Aristotelean accounts like the one of Martha Nussbaum with a clear-cut appeal to human nature capture the MFT's idea of certain areas in which moral judgment is bound to happen as a matter of our biology:

> As Aristotle pointed out, and as current virtue ethicists have elaborated (Nussbaum, 1993), what it means for a personality characteristic to be a virtue, and not simply a behavioral regularity, is largely that it consists in functioning well in a specific "sphere of existence." And what Aristotle and Nussbaum mean by "spheres of existence" is similar to what evolutionary biologists would recognize as persistent adaptive challenges and other types of environmental constraint. Virtues are therefore quite at home in a scientific theory of moral functioning based on evolutionary psychology and cultural psychology.
> (Haidt & Joseph, 2007, p. 387)

> Aristotle's and Nussbaum's approach is also a nativist one, albeit one that locates the innate moral content in both the organism and the environment. Our four modules of intuitive ethics are in a sense a pursuit of this Aristotelian project. Like Aristotle, we are seeking a deeper structure to our moral functioning, though in the form of a smaller number of phenomena that are located more in the organism than in the environment
> (Haidt & Joseph, 2004, p. 63)

The quotes indicate what I mentioned before in the section about Nussbaum's virtue ethics: it is the general limitation of which spheres of life are possible candidates for moral evaluation (domains) combined with a sensitivity for environmental factors and individual differences when it comes to defining what excellence in these spheres consists in (culture dependence of domain expression) that strike Haidt and Joseph as a particularly good match between both understandings of morality.

I therefore regard Haidt and Joseph as advertising a virtue ethicist standpoint on the foundation of the MFT and therefore seeing *character traits* instead of *kinds of action* as the primary carrier of moral value. But as I already pointed out in the beginning of the chapter, the example study that is taken to support both MFT and SIM focusses explicitly on the judgment of *kinds of actions* and not of *character traits*.

Remember that the questions that the participants were asked in the study are designed in the following, action-related way:

> "What do you think about this? Is it very wrong, a little wrong, or is it perfectly OK to **[act specified]?"**

Note that by asking about kinds of actions, the question abstracts from the concrete situation and tests rather for a rule-compliance understanding of morality than a trait dependent one. In the upcoming section, I focus on the consequences of this contrast.

3.4.3. Negative effects of virtue ethicist *moral judgment* on validation of MFT/SIM

As I said, the example study is framing moral judgment as a kind of rule compliance. The different moral domains are mainly separated by the different kinds of situations that rules or norms are about. What a moral judgment is depends on the importance of the norm that is judged to apply: the most important norms are moral norms. Moral judgments are judgments concerning moral norms. Now let these norms change from legislative rule-like norms to a valuing of different character traits in the sense of behavioural reactions in certain situations. One can illustrate the difference between the two with a typically Aristotelean example: rule-like norms in the manner of "Never lose your shield in battle", "Never turn your back to the enemy while in battle", etc. can be regarded as instantiations of a single virtue norm "Be brave".

If we want to build this conceptual change into the derivation of the prediction of study results from Haidt's theory, on the first glance, nothing really changes in the sentence that deals with norms:

The moral norms of a culture A are

a. the set of norms that are seen as most important by its representatives

The sentence just becomes a statement about the value that is attributed to traits instead of value attributed to rules. This value attribution can still be measured in the same way as rule evaluation. I can test whether someone values courage in battle by asking about her opinion on losing one's shield in battle. But below the surface, one little trap sneaks into the rationale that can be easily overlooked:

Remember that traits are helping us to predict other persons' behaviour. Therefore, traits of a certain type can be an indicator of another trait. Having good table manners might for example be predictive for self-control and humbleness. This means that even if a trait that has no intrinsic value (like table manners) it can become a reliable indicator of a very valuable trait (like loyalty), and we will develop an affective attitude towards the indicating trait – if we assume affective attitudes to work in the way Haidt does, namely as associations between body states and representation of situations and objects[37].

Note that this kind of value- transition does not work with action rules: valuing one rule because complying with it will indicate complying with other, more important rules does not seem like a sound way of deriving an action's goodness. As I mentioned table manners might for example be valued because they indicate self-control and humbleness. To say however that the importance of "It is wrong to start eating before everyone's plate has been filled" is a direct implication of the importance of "it is wrong to make risky decisions" or "It is wrong to aim for dictatorship" seems downright absurd.

If we have a trait-based norm concept at hand, the moral value of a norm becomes therefore *transitive* in a certain way. This value-transitivity exists however only under the condition of accepting a trait-based concept of norm, not under the assumption of norms being law-like rules about types of action. And this leads me to an important point that many virtue theorists make: *real* virtues are always valued intrinsically, never as an indicator for some other, greater trait (Aristotle, 1999, pp. 1097a).

Within an Aristotelean framework of morality, the fact that a norm is merely valued highly does not constitute the morality of a norm, since a moral norm will have to be valued *intrinsically* – a subtlety that the operationalization of importance of norms within our exemplary study cannot possibly account for. We now have to confront a very inconvenient consequence: the study that once support-

[37] For a more elaborate analysis of Haidt's view on emotions, please have a look at section 2.3

ed MFT and SIM does not do so anymore under the light of the kind of Aristotelean virtue ethics that Haidt embraces. From the universalisation of the respective judgment and the will to sanction the judged action one cannot conclude whether the character norm pertaining to that judgment is valued *intrinsically* or only because the behaviour is seen as indicative for another trait that is *actually* morally valued.

This philosophical chain of thoughts is inspired and related very narrowly to a psychological hypothesis tested in a study by Tannenbaum et al. (Tannebaum et al, 2011) that I introduce in the upcoming section. The results of this study undermine the MFT in a way that is viable only to the virtue ethicist.

3.4.4. Virtue ethics undermines the MFT

Evidence that non-harm-based moral judgments stem from character evaluations: Tannenbaum et al. (2011)

In their study, Tannenbaum et al. (2011) investigated in how far certain actions affect judgments of morally relevant character traits and in how far harmless actions can lead to negative judgments of moral character. For this, they conducted a survey in which they confronted study participants with two stimuli: one of these stimuli displayed a moral transgression involving harm done to other human beings and the other one displayed a moral transgression involving no harm done to persons. Both scenarios featured young men "who learned that their respective girlfriends had been unfaithful, and reacted violently to the news. The *woman-beater* scenario read as follows: 'John learns that his girlfriend of 8 years has been sleeping around with another man. Upon hearing this, John becomes overwhelmed with rage and beats up his girlfriend.' The *cat-beater* scenario replaced 'beats up his girlfriend' with 'beats up his girlfriend's cat.'" (Tannenbaum et al, 2011, p. 1251)

The test persons were then to judge which behavior was *more immoral* on a 7 Point Likert Scale with the endpoints 7="The cat-beater's behavior" and 1="the woman-beater's behavior". Additionally, they were to compare in the same way several "character attributes, including which person was more empathic, sadistic, 'sick and twisted,' 'screwed up,' and likely to feel sorry for the homeless, help the homeless, enjoy the suffering of others, and have normal human feelings." Finally they had to judge in the same way, which behavior was "more common".

The behavior of the "woman-beater" was rated significantly worse than that of the "cat-beater", *while it was the cat beater that was assigned a slightly worse character*. 75% of the participants dis-

played this act-person dissociation. The results of this first survey with direct comparisons were replicated in an online survey employing the same scenarios but avoiding direct comparison by asking for the judgment of only one scenario.

These findings suggest that character traits are indeed attributed via indicative actions – even if these actions happen to be judged as "morally better" than harmful actions in itself. The outcome relates very neatly to my point about the intrinsic value of moral norms. Character traits are indeed attributed on the base of harmless violations – but it is still harmful violations that are to be considered worse. The best explanation for this contrast is that bad character traits are harmful character traits. A person that is hurting a cat is – sooner or later – going to hurt a person. This explication of bad character traits as being indicators of a harmful personality would however stand in contrast to the acclaimed aim of Haidt to formulate a moral theory that is not based on harmfulness as final criterion for moral value. It starts to look like embracing virtue ethics might cost more than just narrowing the empirical foundation – it might even cost the central point of MFT!

How virtue ethics undermine one of the MFT's central claims

Remember that Haidt argued against harm based morality empirically: he argued that non-western cultures assign a much greater value to non-harm-related norms than we do and therefore a harm-centred view of morality were eurocentristic. And now remember the fact that in European moral thought, due to general situationalist tendencies, value-transitivity between norms does not seem to make much sense while in other, virtue focussed cultures it might very well do. These two pieces of information would allow to explain the moralization of non-harm-related offenses outside Europe in terms of a harm-based picture of morality. Valuing non-harm-related norms becomes a quite likely consequence of valuing harm-related character norms. Please take into account the result of the Tannenbaum study that showed how certain offenses, even harmless ones, are taken to offer insights about the immoral character of a person. And now imagine that sins against the gods and against the order of society could serve the same purpose: nonconformity to public or religious order could be perceived as an indicator of a harmful character. If moral judgments are about the character of the actor, and transgressions in the purity and the authority domains indicate a harmful character, the likeliness of a person to inflict harm to oneself or others can all of a sudden be pinned down as the decisive identity criterion of moral judgments. Haidt's picture of the proper domain of moral disgust (do not eat poison) and moralized social norms (obey the ones in power) fits into this picture, as eating poison and evoking anger of those stronger than you can indeed be indicating harmful character traits - harmful to yourself, but also to the ones close to you. This value-transitivity can be

instantiated very easily by a purely associative, non-deductive, intuitive connection between harmless and harmful rule violations. Character based ethics paves the way for explaining Haidt's most important points (domains of moral judgment, emotional valence of morality) from the perspective of one of his most important target of criticism, which is harm-based concepts of morality.

From these thoughts I draw the conclusion that Aristotelean virtue ethics and the MFT as well as the operationalization in our example study are plainly incompatible. Adopting virtue ethics and primacy of character as a carrier of moral value will come at an unacceptable price to the MFT.

3.4.5. Conclusion

This chapter brought further new insights into the remarkable facets of the interactions between philosophy and psychology:

I found that taking a philosophical standpoint because of superficial coherence with one's empirical claims does not necessarily work out. I considered Haidt's argument in favour of virtue ethics that is based on the superficial similarity between the MFT and the take of virtue ethics on relativity of moral rightness. In his adherence to virtue ethics, Haidt neglects two important features of virtue ethics – which are on the one hand the intrinsic goodness of moral value and on the other the value transitivity between character traits. This makes him overlook the fact that virtue ethics not only undermines the empirical foundation of his theory (as shown in 3.4.3) but even the advantage in explanatory power that it assumes to have against harm-based approaches to morality (as shown in 3.4.4).

On the metacientific side, I showed that the operationalization of moral judgment in the example study binds Haidt to a situationalist account of moral judgment. On the philosophical side, I demonstrated how this commitment undermined the empirical foundation of Haidt's very own philosophical argument for virtue ethics, validating my hypotheses 1 and 2.

Furthermore, I observed from a metascientific perspective that committing to a virtue theorist account of moral judgment would make the MFT vulnerable for falsification through a chain of reasoning involving value transitivity and specific empirical results, validating hypothesis 3.

The sections about structure and object of moral judgment aimed to highlight the importance of being clear about conceptualisation and operationalization in the practice of social psychology and especially in research about moral judgment. The remaining two sections serve mainly the cause of

highlighting additional difficulties for philosophers employing empirical results..The first of these points relates to the issue of the role of intention in moral judgment. In the upcoming segment, I argue that neglecting intention attribution as a factor in moral judgment all together does in fact imply a commitment to a very unusual philosophical position about moral judgment. In fact, this position would be so unusual that Haidt's moral theory would become incompatible with philosophical tradition and therefore uninteresting for many empirical philosophers. I suggest two ways to interpret Haidt's theory favourably by "implanting" intention attribution into his models – and show how the viability of these ways will in the end depend from one's philosophical standpoint. In the end, I explain how a choice for one of the options allows for new ways of empirical argumentation.

3.5. Primitives III: intention

It is – on a first glance - a very convenient feature of the MFT and the SIM to be neutral towards any philosophical debates about other philosophically loaded primitives for explaining morality than intuition, emotion and reasoning. By claiming that morality is just the set of norms held most important by a given society, Haidt explicitly leaves aside all "fixed" criteria for morality, like whether an act was performed intentionally or committed purely accidentally. But a point that I wish to press in this chapter is that even if one explicitly stays indifferent about a certain point (in this case the agent's "intention") this indifference can itself be regarded as a philosophical commitment, with all the consequences and implications that explicit philosophical commitments tend to bring along. In the MFT and the SIM, as well as in the exemplary study confirming them, the term *intention* does not figure. I regard this as a clear sign that Haidt's theory is in fact indifferent about intention when it comes to define morality. Staying indifferent about intention in matters of morality IS however a very unusual standpoint, even within more than 2000 years of (European) philosophical discussion.[38]

[38] However, this does not mean that the topic of unintended moral guilt or he feeling of being guilty would not feature prominently in European literature, for example in *Sophie's Choice*, or *Ödipus*. It is not that the thought of moral guilt without intention would be unknown to European thought. It is rather in philosophy and law that it has traditionally been regarded as a mistake not to take intention into account when judging a certain behavior.

3.5.1. Intention as a factor in ethical theory

The notion of *intention* as an important factor in agency is without any doubt one of the oldest and most important primitives to be found in explanations of what moral judgment is. The main idea behind it is that most actions happen voluntarily, with the agent successfully acting in order to reach a certain aim. On the other hand, there are actions the outcome of which was not the originally intended result. If I give a "thumbs up" to an Arabian person, I might have unintentionally insulted him, not knowing that the gesture means "up yours" in several Arabian countries. It was not the desired outcome of my action to offend anyone. The undesired result of my action is due to unfortunate circumstances. The notion that intention plays a role for moral judgment is first introduced systematically by Aristotle in his Nicomachean Ethics (Aristotle, 1999):

> "Virtue, then, is about feelings and action. These receive praise or blame if they are voluntary, but pardon, sometimes even pity, if they are involuntary. […] Now it seems that things coming about by force or because of ignorance are involuntary."
> (Aristotle, 1999, p. 1109b)

The idea behind the distinction between voluntary and involuntary action is having a criterion about whom to blame for a certain damage. If something is done without intention and unwillingly, the agent cannot be held responsible for the outcome or the action. If an athlete throws a spear for exercise[39] and hits somebody with it involuntarily, and the spear was sharpened and not blunt as she thought, she is not responsible for the effects of her action and therefore not to blame (Aristotle, 1999, p. 1111a). This passage can of course be regarded as a mere sociological description of a certain set of moral norms at a certain place at a certain time. But the fact that this distinction is made to investigate the nature (or concept) of virtue makes it quite plausible that this is not an empirical, but a conceptual point. It is made not to describe what people think about blame and praise (even though this is also a part of Aristotelean method), but to actually set limits to the responsibilities of the individual concerning her virtuousness. Blame and praise is about intended acts.

This conceptual point of the importance of intentionality for the goodness or badness of behavior remains crucial throughout the ages: Immanuel Kant for example states that whether an act is morally praiseworthy or despicable is dependent on the maxim of one's **will** (Kant, 1902) and not on the outcome of an action. If you did anything you could in order not to do something, took any precau-

[39] While, of course, following all necessary safety measures

tion, and still do it, because the action could not be avoided – you are not to blame. Moral judgment is about **intended** acts.

And even consequentialist theories like Utilitarianism (that stay rather indifferent towards intention when it comes to the formulation of what morality **is**[40]) assign a key role to intention: although the intention does not affect the goodness of an act, the presence of an intention determines whether the immoral act is a certain person's act or not – and therefore whether the person who caused damage is to be judged as an agent. The Utilitarian Jeremy Bentham is very explicit about this: "If the act be not intentional in the first stage, it is no act of yours" (Bentham, 1996, VIII-V, p. 85). And so is also John Stuart Mill as the reader might recall fom the last sections' quote "the morality of the action depends entirely upon the intention—that is, upon what the agent wills to do." (Mill, 1972, p. 19). Acts are only acts if they are intended. To judge a person's behaviour necessitates judging her intentions.

The examples of Aristotle, Kant, Bentham and Mill show us that intention is deeply intertwined with how western philosophical tradition understands morality. The intention of the agent decides whether it was an act of her own or not. Moral judgment necessitates a judgment about intention. Intention ascription should also affect what Haidt calls the moral domain: If an action is condemned, even though it was clearly not intended, it would be very difficult to call this judgment a true moral judgment in the sense of any of the thinkers mentioned. This condemnation would either be a wrong moral judgment, or the concept of moral judgment applied would significantly differ from what is normally understood as moral judgment.

It seems that concerning philosophical takes on morality, Haidt is fairly alone with the standpoint that intention is no a priori conceptual factor in moral judgment. Let me stress again here that it is Haidt's *choice* and a *conceptual* foundation of his theory not to include intention in his definition of morality. It is not an empirical result. But even though Haidt does not mention intention, we might manage to "smuggle it" into his theory. With "smuggling it" in, I mean finding ways to interpret the SIM and MFT that make Haidt's take on moral judgment compatible with an important role of intention in morality. In my opinion, there a two options to do so. However, there is a philosophical price to pay for each version.

[40] The outcomes of an action and not the motive of an action decide about its moral praiseworthiness

3.5.2. A role for intention in the models – two scenarios

In the last sentence of the previous section, I wrote about "smuggling in" *intention*. Let me specify this a bit further: the vehicle through which intention is to be smuggled into Haidt's theory is something that I would like to refer to as *intention-sensitivity of moral judgment*. With this I mean the idea that moral judgment has to *imply something like intention ascription* and that information about intention is *relevant for moral judgment*. The way I see it, we could account for intention- sensitivity within the framework of Haidt's models in two ways:

- taking intention- sensitivity as mediated *by* intuitive-emotional processing
- taking intention- sensitivity as mediated by an act of self control *in the aftermath* of the actual moral judgment

Both ways of accounting would change an aspect of Haidt's theory. The former would have an effect on what accounts of emotions could be fitted into the framework, the latter would have an effect on whether moral judgment is a purely intuitive process after all. I investigate both possibilities in the upcoming part of this segment.

Scenario 1: a cognitivist account of emotions

Above, I showed that Haidt postulates moral judgment to be emotional judgment. An emotion consists of two parts: A stimulus representation and an ensuing representation of a somatic response. Intention- sensitivity seems very unlikely to be mediated by representations of somatic responses. But it does not seem unlikely that intention-sensitivity is a component of a stimulus representation part of a moral emotion. Triggering of a moral emotion would in this case be dependent on recognition of intention, just as it would be dependent on recognition of the presence of an agent. Recognition of intention would become component of moral emotions and therefore of moral judgment. Of course, intention is a relatively complicated notion. It has a lot of theory of mind in it and judging intention requires a lot of experience and knowledge. And here comes the philosophical price one would have to pay: if intention ascription requires knowledge, intention ascription cannot be fit into a radically perceptive theory of emotion. As long as perceiving is not to entail judgments of whether something is the case, and as long as judgment, but not pure perception requires knowledge of the world, then intention-sensitivity of moral emotions entails that moral emotions are not perceptive.

The introduction of intention sensitivity of moral emotions entails a more cognitivist interpretation of emotion and therefore of Haidt's theory. It seems needless to mention that such an option would seem very unattractive to perceptionist philosophers like Jesse Prinz who argue based on Haidt's theory: either they have to bid farewell to intention sensitivity and to using the concept *moral judgment* as people normally use it, or they would have to explain the problem away which seems hardly possible. In the case of Jesse Prinz, this would be especially drastic, since his whole approach to moral philosophy hinges on his perceptionist account of moral judgment as described in section 3.2. Luckily for perceptionsts, there is another way of smuggling intention-sensitivity into Haidt's theory. Even though actually, it is rather attaching intention sensitivity to Haidt's theory as a kind of footnote. Let me explain how it could work.

Scenario 2: a phenomenalist interpretation of MFT and SIM

In the previous section I figured out how to sneak intention into Haidt's moral theory. I found that not everyone would be happy with this way of handling things, though. Accordingly I present an alternative way of bringing intention together with the SIM and the MFT. Let me tell you two stories that could help me to do so:

When I recently talked to a saleswoman in a shop while trying on a suit: "I like this cut, but I really would like something that is not black, like the one I am trying on right now." She replied that this suit actually was not black, but a dark shade of blue. I should go in the sun outside of the shop to convince myself. I did, and indeed, the suit was blue. The lady said: "This happens a lot, in the lighting of the shop a dark shade of blue or grey looks black, even though actually, it is not."

In the winter before, I was participating at a students' negotiation competition. The team of my university was negotiating a case with another team and we did a fairly good job at not giving away too much confidential information. After the negotiation session, one of the two judges attending the session gave feedback to our team as being "the biggest liars she had ever seen". I was immediately infuriated by this remark and was going to complain about this insulting behaviour. Only minutes later I learnt that the judge had used the expression in a way that I was not familiar with: as a compliment[41]. I instantly revised my judgment: her behaviour just seemed offensive because I was not

[41] "You are the biggest liar I have ever seen" is, as far as I am aware by now, a compliment concerning professional attitude in legal affairs with a slightly ironical touch.

familiar with her particular use of the term “biggest liar I have ever seen”, but actually it was not - it was not her intention to insult our team.

There is a parallel between these two stories and I suggest it is one that helps to find a place for intuition in moral judgment. Just like my colour judgment was impaired by the bad lighting in the shop, my judgment of the behaviour of the judge was impaired by my ignorance of her intention. My point is that knowledge about the true intention of an agent can be understood to play the same role in proper judgment of behaviour as proper lighting plays in proper judgment of colour. Accordingly, the second way to introduce intention to the SIM and the MFT would be reinterpreting them as an account of *appearances* of morality. The key idea is the following: we might agree that after having derived a moral judgment intuitively we still might want to subject this judgment to some assessment of reliability in order to safeguard ourselves from something like “moral illusions” understood analogously to optical illusions. This would not alter the importance that intuitive processes play in deriving a moral judgment nor the existence of moral domains that can but do not need to be nourished by a culture. After all, we could still agree that the *actual* moral judgment is still intuitive. The cognitively controlled, assessing part is just some sort of quality control that does not add anything new to the judgment but just works as a quality control for intuitively appealing judgments -just like our capacity to spot optical illusions does not make our vision a cognitively controlled process. As we draw the analogy between colour perception and emotion further, we can employ an observation of Wilfrid Sellars (Sellars, 1997) about the meaning of a sentence like “it is red, but it looks black in green light”. It seems like also our perception of colour is at times erratic – just like our perception of moral properties. And just like this occurs in specified situations in the cases of colour perception (normally unusual or substandard lighting, sometimes also physiological abnormalities or semitransparent obstacles), this occurs with the perception of moral qualities in situations in which an action is committed unintentionally. In colour perception, the term “looks” can now be introduced to conceptually grasp the difference between sensing the redness of an object and the object being red. In moral judgment, this distinction is rather uncommon, but might be of some use here: it allows for rephrasing Haidt’s theory as a theory about “looking moral” instead of “being moral” with the “looking”-part still being the decisive part of moral judgment.

The meaning of the term *moral* would, in this scenario, be similarly dependent on application rules for our moral sense like the meaning of the term “yellow” would be dependent on the application rules of our colour vision. And we can of course be erratic about

1. whether the conditions specified by these rules hold or not
2. what these rules are exactly.

In contrast to a revision of moral emotion, this result would be one that perceptionists like Jesse Prinz could very well live with. Their philosophical commitment to the perceptional nature of emotion forces them to adopt this reading of the SIM and the MFT, given that they want to keep a place for intention in their understanding of morality.

3.5.3. *Intention-sensitivity* entails additional predictions of MFT/SIM

Now that I introduced two ways of slipping *intuition* into Haidt's moral theory as a matter of interpretation, the question pops up whether their empirical predictions might change and new ways could be found to empirically assess which of the interpretations is the most empirically adequate. The example study

of course not be a suitable candidate. After all, it does not control for intention of the depicted agents, for example through including "unintentional" conditions or an explicit remark that the persons in the scenarios knew what they were doing. And indeed, we know that in doubtful cases, people tend to attribute intention to actors: There is a bunch of empirical data about the so called *Knobe Effect*[42]*, the tendency to attribute intention* when condemning an act the intentionality of which is explicitly left unclear. It is tested for with stimuli in which somebody does explicitly not care about potentially negative side effects of his actions. If the side effects are harmful, intention is attributed and the situation is moralized. If side effects are beneficent, no intention is attributed and the action is not seen as particularly morally valuable. It seems therefore sensible to assume that in the case of moral outrage, an action is considered as intentional if there is no explicit hint to the contrary (Knobe, 2003). If a behaviour therefore falls into a moral domain and the intention of the actor is unclear, the judging person can be expected to assume the behaviour of the agent as intentional as a default assumption. Applied to the paradigmatic study, this effect implies that very probably, test persons automatically ascribed the agents in the stimuli bad intentions.

So, if no control for intention sensitivity is introduced, everything seems to work out very nicely for both *intentionalized* versions of MFT and SIM. However, there is an empirical finding that implies exactly that moral judgments from the non-harm domains are not intention-sensitive at all. And it is here that the two options for interpretation suggest different ways of handling the data.

[42] Named after its discoverer Joshua Knobe

3.5.4. A curious finding

In a study that could become a problem for the intention-sensitivity MFT and SIM, Young and Saxe (2011) made an online survey in which every participant made a judgment for a single scenario as depicted in the figure below.

Haidt's domains were represented by different groups of scenarios: harm, incest or ingestion (see figure6). There were two versions of each scenario: one in which a breach of norm relating to the domain was performed intentionally and one where it was performed accidentally.

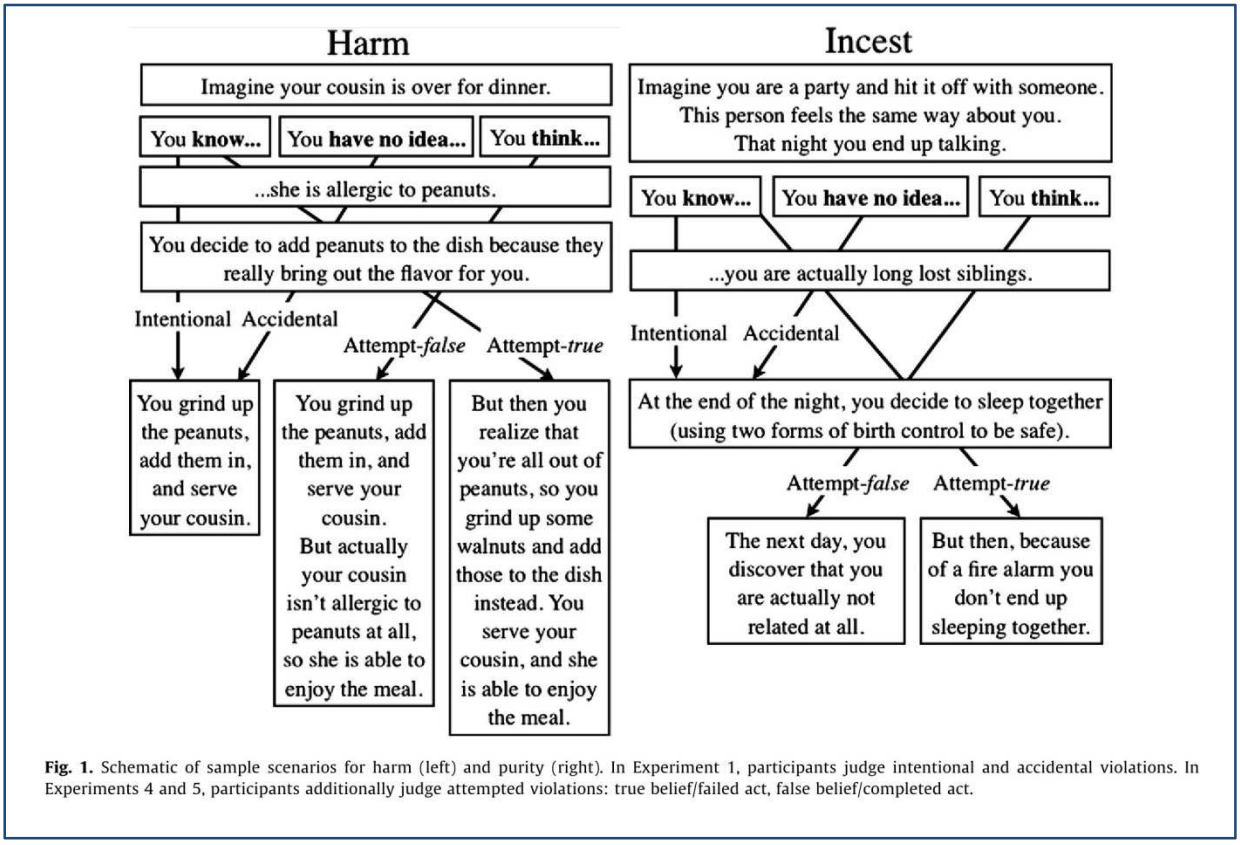


Fig. 1. Schematic of sample scenarios for harm (left) and purity (right). In Experiment 1, participants judge intentional and accidental violations. In Experiments 4 and 5, participants additionally judge attempted violations: true belief/failed act, false belief/completed act.

Figure 6: *Schematic of sample scenarios from Young & Saxe (2011)*

The results showed that if committed intentionally, all behaviours were judged as very morally wrong. Accidental transgressions involving harm however were judged as almost not at all morally wrong while on the other hand the "purity domain" transgressions were still judged as more or less grave moral transgressions. See figure 6 for a visualization of the contrast between harm-related norm transgressions and non-harmful norm transgressions.

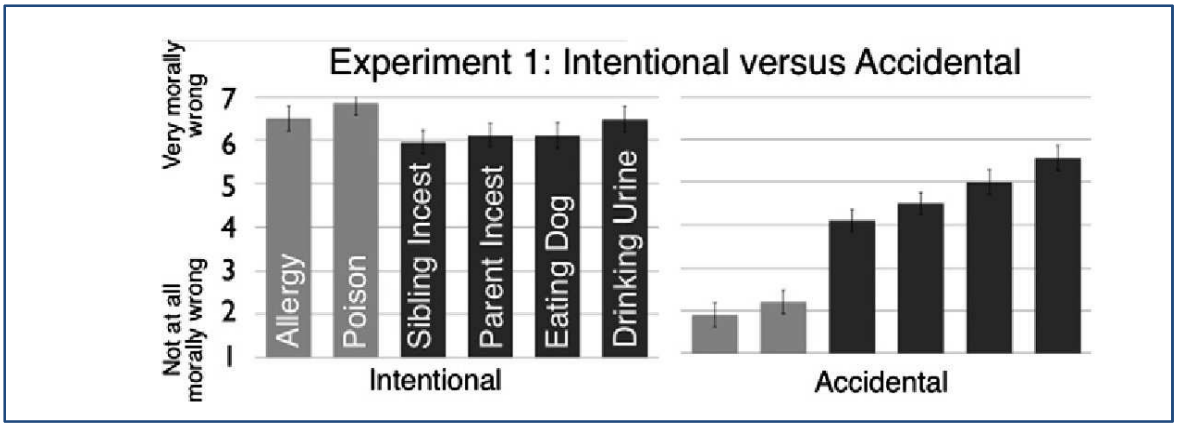


Figure 7: *Visualisation of results from Young & Saxe (2011)*

In several follow up studies it could be excluded that the effect was produced by a) a tendency to assume higher agent control in the incest – situation b) a higher emotional valence of accidental incest compared to accidental harm or c) moral judgments as disgust ratings.

Young and Saxe have shown in their study that people tend to make moral judgments in the divinity and hierarchy domain without regard to intention. Dependent on our stance about the role that intention plays for our concept of moral judgment, we face two options:

Option number 1: The way moral philosophy describes moral judgment is not empirically adequate

This choice is the one we would have to make if we stayed indifferent about intention. As Young and Saxe show that judgments that are explicitly called *moral judgment* do not take intention into account, the idea that intention decides about the morality of an act is downright refuted. This would be a really revolutionary conclusion – but one that would implicitly cut the conceptual tie of *moral judgment* to European moral philosophy and penal law in general[43]: intention as the condition for agency can hardly be taken away from moral philosophy. The concept of moral judgment changes dramatically, if the concept of agency changes. It is a judgment about another thing. The moment one regards intention as irrelevant for moral judgment, one is not talking about the same thing as Kantians, Utilitarians or Aristoteleans anymore.

[43] At least to accounts of penal law that regard law as closely linked to justice.

Option number 2: the *intention-sensitivity* of the MFT must be adapted

If we are to agree with philosophical tradition that intention is a conceptual determinant of moral judgment, Haidt's moral theory will have to bite a bullet. How tough the bullet will be, depends on the way we introduced intention into the theory. The tougher bullet would without any doubt be the one that an intention sensitive notion of moral emotion would entail: the domains of purity and hierarchy do not work like moral judgment as commonly understood works and therefore they would in fact not be *moral judgment* in the sense most people use the term. A model like the MFT that postulates such a conclusion would cease to be a theory of moral judgment in the eyes of Kantians, Utilitarians, Aristoteleans, and probably also adherents of the common language meaning of *morality*. However, this option seems pretty unattractive. Or vice versa: Saving the MFT would entail that its concept of morality is something completely different from the one of traditional moral thought.

The picture changes if one chooses to phenomenalize the MFT and the SIM: As soon they are about how a situation "looks" to be, it is very easy to explain away the result. One just needs to argue that the participants that judged the displayed behavior as immoral were victims to a *moral illusion*, analogous to optical illusion. Just like in the Müller-Lyer-illusion, the illusion that one line appears longer will not entail that we have to change our concept of *longer*, but just that we will have to note that appearances do not always align with what is reasonable judgment. Handling intention through phenomenalization of the theory would deal much better with additional evidence than treating intention-sensitivity as a component of moral emotion. It would allow to sensibly keep intention out of MFT and SIM while still allowing for western moral philosophy and Haidt to be compatible.

Note however (assumed that the empirical evidence given is in fact valid and not subject to artefacts) that the choice between cutting all conceptual ties to moral philosophy on the one hand and phenomenalizing Haidt's theory is a *forced choice*. As far as my argument goes, there is no alternative. Every philosopher who applies Haidt's thought to his theory will have to assume its phenomenal nature. Otherwise he or she would make a mistake to take the MFT or the SIM to be about morality in the classical philosophical sense – which it cannot be, as long as there is no place for intention-sensitivity.

3.5.5. Conclusion

In this chapter, I made several points. I started with ascertaining that intention is – even by consequentialist philosophers – generally seen as a condition for agency and therefore for the moral valence of an action. Haidt however stays entirely neutral about intention in his formulation of moral judgment.

However, I explained why being neutral about intention does not mean that one is neutral about intention: taking intention into account as a decisive factor in one's concept of agency does matter quite a lot to what is considered the domain(s) of morality. Whether something has been *done by somebody* or is just a result of chance is of utmost importance to philosophical tradition. I called this the *intention sensitivity* of moral judgment. Under the assumption that intention sensitivity of moral judgment was somehow to be implied by Haidt's theory, I looked for ways to implant intention sensitivity into the exemplary empirical models. I found two ways to do so and I demonstrated how the appeal of each of them varied depending on what philosophical standpoint one was committed to in the first place. Then I demonstrated how one of my newly introduced rivalling philosophical interpretations of Haidt's theory did a noticeably better job explaining certain additional empirical evidence than another – a clear indication that the decision how to philosophically interpret the model does affect its empirical validity. Given the validity of the empirical results cited, the choice of the interpreter was in the end between phenomenalizing Haidt's theory or accepting it to be about something different than morality as understood in the common language sense of the term.

After investigating how neglecting a factor about moral judgment does in fact entail a philosophical commitment, let me demonstrate how a clear standpoint that on first sight seems like a clear philosophical commitment can turn out not to be one after all. This is the case with Haidt's insistence on the relativity of moral judgment that seems relatable to another piece of philosophical thought about morality: moral relativism. This is generally considered to be the idea that two conflicting moral judgments can be true at the same time. This on first sight counterintuitive idea can be explained in a variety of ways. In the next chapter I investigate whether Haidt's approach does indeed commit itself to moral relativism as understood in philosophy and whether one might find ways to make Haidt's assumed philosophical commitment less strict. In both cases, I come to a positive conclusion, showing that Haidt's theory is in fact compatible with objectivism about moral truth once one is willing to accept minor changes to common philosophical interpretation of the theory.

3.6. Truth

So far, we have seen that the MFT, SIM and empirical research rely on some quite specific understandings of certain aspects about moral judgments. One aspect that is especially evident though has so far not been covered, even though it has been appearing on the roadside here and there: Whether there is something like the truth of moral statements and whether it is culture dependent or not.

In our everyday life we are confronted with the issue of moral relativism all the time: moral judgments are considered as universal, even when they are explicitly said not to be. Vegetarians for example can be very aggressive and evangelist about their moral conviction that killing and eating animals is wrong. But even if they are not and explicitly state that they are ok with other people eating meat, people who eat meat still sometimes feel accused by their mere insistence of not eating meat for moral reasons. I suggest this effect occurs since it is very difficult to imagine somebody abstaining from something for moral reasons without condemning those who do not abstain. This sometimes leads to the absurd situation that carnivores accuse vegetarians of moral superstition, even though the latter had explicitly stated to have nothing against people eating meat. Moral judgment just has something objective and universal in it, analogous to judgments about things being inflammable or water soluble.

On the other hand we often have no problem to say that for example Aristotle was not a bad person just because he owned slaves - we know that owning slaves was not a moral issue in Athens during the fourth century B.C. Moral judgment has something relative in it, just like the sentence “For me, strawberry ice cream is the best. For John however, chocolate has the best taste.” Moral judgments seem to cover a very peculiar middle ground between judgments of taste that we would regard as true only relative to a person and her environment and judgments of physical facts, the truth of which is not relative to anything but objective and universal.

There exists a considerable body of philosophical work about this aspect of moral judgment that is mainly circling around the question of what it means for a moral judgment to be true; whether moral judgments are true relative to a time, to a space, to a culture or to a person; or whether moral judgments are true objectively.

In this chapter, I assess whether Haidt’s approach can be regarded as taking sides within this question and I come to the result that Haidt’s position should be regarded as leaning towards moral relativism. I however also show that it is also compatible with a certain understanding of moral objectivity and that a minor qualification is enough to get rid of the problem of moral truth all together.

3.6.1. Several remarks about truth in general

As I remarked in the introduction chapter, this work is based on a concept of meaning that emphasizes that the meaning of a term is dependent on the application of this term in theory and practice. Whether a term is applied correctly is dependent on the theoretical context in which it is used. *Sea fish* might have different meanings when used in the context of different gastronomical theories: In some, it might include whale-sushi, cuttlefish and oysters, in others it might not.

There is a certain idea of truth connected to this understanding of meaning that I myself have to presume: if the true meaning of a term is – in some way – context dependent, the true application of a term is context-dependent. This however entails that the truth of all sentences is theory-dependent. Truth of a sentence is something that is dependent on the picture of the world (or language, or conceptual scheme, or theory of the world) that the onlooker is sharing. This is much less crazy than it seems on the first view: As I already mentioned, whether an utterance of the sentence "this is red" is truthful depends on the conditions that are assumed to be necessary for truthfully assessing the colour of an object. Therefore, just like whether "the object is red" is *true* depends on these conditions, so does the question whether the object *is* in fact red.

This implies a certain idea about truth to be a non-starter in this chapter: The idea of true sentences to be mapping the real picture of the world (like a photograph or picture mapping the "real" look of a scene on paper or screen) does not work in its most naive interpretation. Whether a sentence is depicting the world correctly is dependent on the conditions for such a mapping to count as correct. Such conditions however are theory (or language, or conceptual scheme) dependent. There is therefore no objective truth in the sense of an objectively true mapping of the world.

The objectivity of moral truth is therefore always *objectivity within a certain picture of the world*. The idea of objective moral truth is not the idea of the possibility of an objective mapping of some theory-independent moral values floating around somewhere in Platonic heaven. As I mentioned, this idea would be a non-starter, just like the idea of a theory-independent mapping of which objects are *red*. Whether truth of moral judgment is relative or objective is rather a question that deals with the following alternatives: on the one hand moral behaviour can be explained with appeal to a theoretical entity called *moral values* that is objective within a certain picture of the world. One could compare this to the concept of *mass* in a world view that employs Newtonian Mechanics[44]. On the other hand, moral values can be regarded as a theoretical entity that is relative to culture, time, etc. An

[44] Mass cannot be measured directly. Volume, speed or weight can. It is from measurements like these that mass is calculated from.

analogous case would be the species-relative concept of *sexually attractiveness*: what is *truly* sexually attractive is very much dependent on the species that you are looking at. *Sexually attractive for baboons* is radically different from *sexually attractive for humans.* I argue that the MFT and the SIM are compatible with both ways of understanding *moral truth*. It is compatible with understanding moral truth as relative like statements about sexiness and it is compatible with an understanding of moral truth as objective like statements about the mass of an object.

3.6.2. Moral truth in the MFT

A surprising finding about Haidt's writings is that the issue of relativity of moral truth is never addressed. He might have skipped this part because it is just too trivial a fact that he will not for example adopt a strict ethical objectivism assuming that each moral sentence is in fact either true or false - and that if everyone in a certain culture believes that A is wrong, but in fact, it is not, then everyone in this culture is wrong. This just seems to fly straight in the face of the whole idea of the Moral Foundations Theory. After all, the idea behind that theory is that morality is culture dependent. But there is a subtle difference between the claim that varying cultures endorse different, partly conflicting values and the claim that statements pertaining to these values are all true. Take two examples in which value-judgments are different but truth relativity is not adopted.

In Traditional Chinese Medicine (TCM), the coherence of the diagnosis with several epistemical values (for example cosmic resonance theory, the existence of Qi, the five elements and the circle of transition between these elements) plays an important role for subscribing treatments and identifying the cause of an illness. Values or principles of Chinese medical theory can therefore coherently be described as something quite different to values or principles of western medical theory that lead it to perform studies or experiments that follow very strict methodological rules based on a cause-effect picture of the world. This does however not prevent that the values of western medical theory might offer better results on cancer treatment, immunization, treatment of bacterial infections through antibiotics etc. It is surely possible to state that principles of medicine are fundamentally different between cultures and still not all equally right.

In fascist or monarchist theories of the state it is assumed that the state can be managed more effectively if there is a strict top-down chain of commands. There are many places in the world of which it is sensible to assume that this is the common idea of how the state and society works – just like there are many places where the democratic, bottom-up concept of democratic government is re-

garded as the most efficient way to organize a state. However, despite acknowledging diverging cultural takes on formal hierarchy in society, one can still insist that democracies have proven much more effective in preventing famines than authoritative systems. Therefore the principles of democracy are more in line with a picture of the state that regards as a main job of the state to keep its subjects alive – and (given that qualification) the principles of authoritative rule are mistaken.

These were two examples how underlying principles of cultural world views can be acknowledged as different, but that this appreciation does not automatically entail relativism about which of these principles are true. I assume that just like in the case of principles of medicine and principles of the stately hierarchy, cultural relativism and relativism about truth are independent. This independence is substantial. The SIM and the MFT are explicitly dealing with why some people come to share certain moral beliefs and other people do not – this does however not entail that they have to accept that all these beliefs are true.

The mentioned findings should leave the philosopher who would like to employ Haidt's results pretty excited: after all, by categorically refuting objective moral truth as a premise, Haidt's theory of morality would be a taboo for a considerable number of philosophers. Treating moral judgments like judgments of taste has some very tricky consequences: for example, it becomes challenging to define what honest moral disagreement with cultural practices of another culture is actually about – since it is not about objective moral truth and not about truth relative to one's own moral values and definitely not about truth relative to the other culture's values. There are approaches to solve this puzzle, though, and in the final part of the segment about truth, I present one that seems to fit perfectly with Haidt's theory. But needless to say: as these approaches are philosophical theories, they are themselves heavily criticized. So in addition to checking for consistency with its natural ally moral relativism, I explore whether there is a way out of the relativist urchin gang for Haidt – which I believe there is. Even though for this way out, one would have to accept some other strong premises in exchange.

In order to assess which philosophical standpoint fits well to the MFT and SIM, let's recapitulate which points about moral truth we can possibly get out of SIM and MFT:

To begin with, there is the concept of norms that are culture dependent (see 2.1.1 and 2.2). If one regards norms as normative sentences then it seems fair to assume that the culture-relativity of norms entails the culture dependence of the truth of sentences like "it is polite to stand up while greeting a lady". These sentences are true only in respect to a given culture. As moral laws are classified explicitly as a kind of norms, and these norms are considered to be empirically assessable, one can assume that the same holds for a sentence like "in culture A, it is morally obligatory to help peo-

ple in distress". This would bring us two first clues: it is possible to make true statements about the moral worth of an action relative to a certain culture.

Given this idea of culture relative moral truth, social links within the SIM entail the possibility of change of truth value of culture- relative moral statements over time. Please keep in mind: moral statements are true relative to the set of most important norms in a given culture at some point in time. The SIM claims furthermore that moral intuition is subject to change through links 5 and 6, the links accounting for interpersonal influence of each other's moral intuition. This makes shifts in common moral intuition possible. That means that the importance of norms may change over time, changing the background to which the truth of moral judgment would have to be assessed. As moral intuition of a group changes, the culture relative truth of moral judgments changes with it. If for example you had the possibility to assess the importance of many norms concerning sexuality in Europe from 1900 to 2000, you would very likely find striking differences in the importance that is attributed to these norms. The SIM could explain these shifts via the mentioned links 5 and 6. And even though relative to the point in time, different behaviours will be regarded as morally permissible and morally despicable, both descriptions of what is morally right and what is morally wrong will be *true*.

We can therefore count three constraints that Haidt's theory does impose on the truth of moral judgment:

- Moral judgments can be true or wrong.
- Truth of moral judgment is culture dependent.
- Truth of moral judgment can be subject to change over time.

3.6.3. Philosophical standpoints about moral truth and their fit with MFT/SIM

Example #1: Gilbert Harman's moral relativism

A philosophical position that fits excellently with this kind of approach is Gilbert Harman's moral relativism. According to Harman, moral sentences like „X is wrong" are always elliptical for „in the value-system of a group of people A, X is wrong.". A value system is supposed to stand for a set of *motivational attitudes* like "do not harm others" (Harman, Moral Relativism Defended, pp. 8) – something that the MFT refers to as *norms*. A sentence "X is wrong" is only true in relation to a given set of motivational attitudes – a set of norms, in Haidt's terms. We might therefore call this set of norms in Haidt's terminology a morality.

Morality is consequently regarded by Harman as a group phenomenon based on affective valences of its members (Harman &Thompson, 1979, p. 9). This set of motivational attitudes is derived through an implicit bargaining-like process (Harman, 1975, p. 13) that can be understood to work in the following way: It is primarily the affective valences of individuals that make certain behaviors and general rules valuable and desirable. Wealthy people for example will regard the importance of property rights as much higher than poor people. Common people for example will regard the merit principle in public administration and military as more important than privileged noble people. And as human beings are *social animals*, they would have to find a compromise about the norms that should govern society. Harman calls the process through which such an implicit compromise is achieved *Moral Bargaining.* This is the process through which group relative sets of motivational attitudes that can be called moralities come into being. These rules and values will differ from one group of people to the other along the lines that the interests of its constituents and its environment will differ. Rules for marriages and organizing gender roles would differ in a belligerent society populated by much more women than men compared to a peaceful society in which the genders are more balanced. Rules of who counts as a citizen and is allowed civil rights will depend on how big differences in wealth within the society are and how much workforce there is in relation to the amount work that needs to be done. The bigger the differences in wealth and the more exchangeable the workforce becomes due to a mismatch between workforce and workload, the less egalitarian will the distribution of rights be. Accordingly, the set of motivational attitudes in relation to which the truth of moral judgment has to be assessed will change from one culture to the other.

Note that this claim is equivalent with making a point about the meaning of *moral*: the meaning of the term *moral* is dependent on how the people are using the term – which is dependent on their respective set of motivational attitudes. As language is a public affair, people will always *somehow* have to agree about what *moral* means. How to use the term *moral* is a pragmatic decision. However, it is a pragmatic decision that is dependent on who is part of the speaker community and who is not. A society of vegetarians might consider animal rights a moral value, while a society of carnivores might not do so. With there being different speaker communities, there will be different meanings of the term *moral*. It is a classical orthodoxy of philosophy of language that difference in meaning does entail (or is entailed by) differences about which uses of the word are considered correct and which ones are not. According to this theory of the meaning of *moral judgment*, "Eating cows is wrong" would be true in many contexts in India where there are many Hindus and vegetarians, but not in Europe. Therefore the meaning of the term *morally wrong* would diverge substantially between both speaker communities. Note furthermore that Muslims and Europeans living in India are not sanctioned for eating beef as long as they do not do so publicly. This can be regarded as a compromise achieved through moral bargaining: the Hindu community benefits from Muslims since they are able

to fulfill jobs which Hindus are not allowed to perform due to religious belief (e.g. burying dead cows).

Note that these considerations about the truth of moral sentences match the points that Haidt makes about moral truth: there is moral truth, as truth in the sense of Harman is "rightful use of the word". Moral truth is culture dependent, as long as culture is regarded as a speaker community. And the truth value of sentences containing moral judgments is subject to change over time, as the composition of the speaker community is so, too.

A highly thought-provoking feature about this idea of moral bargaining is that it is in fact fairly difficult to imagine as a process – Harman stays quite vague about that point. The fact that the SIM can be interpreted to tell exactly the story how moral bargaining works proves from my point of view how excellently both approaches to morality fit into each other.. While Harman is rather taking care of aspects of philosophy of language ("how can moral judgments be true or false without clear cut rules of meaning of moral terms?") – Haidt explains the ways and workings of the social and psychological interactions that are so essential for the societal mutual agreement on values assumed by Harman's theory. Moral intuition is what drives moral judgment and intuition is influenced by the intuitions and reasoning of others. If we keep debating about morality as a society, according to the SIM our moral intuitions will converge to a certain degree. The set of motivational attitudes towards which a society would eventually converge would therefore have to depend on the intuitions of its members, and these will depend on their interests and on the environment of the society. The parallel to Harman's concept of moral bargaining is striking.

So moral relativism and Haidt seem like the perfect couple. They are not only compatible; they even seem to help each other out. Is there however any chance to combine Haidt's theory with an objectivist point of view? Let me attempt to convince you that surprisingly, moral objectivism can be combined with the MFT and the SIM, too.

Example # 2: Judith J. Thomson's objectivism

Moral relativism is refuted for a number of reasons by moral objectivists like Jarvis Thomson: she is arguing from the standpoint that it is indeed possible to find out whether a moral statement is true objectively – and that many arguments against that standpoint can be countered quite easily (Harman and Thomson, 1996, pp. 65). If it is however possible to find out about moral truth, there must

be something like a moral truth. Therefore, a statement like "In relation to A, X is wrong" is not even a moral judgment (ibid., pp. 188), as moral judgments are in principle not relational.

This contradicts on first sight the basic assumption of the MFT that the morality of a group of people consists in the set of norms considered most important by this group: if there were a different way to asking people to find out what is moral truth and what is not, then asking people for what is right and what is not would not be the ideal way to find out about what is morality. From the objectivist point of view, Haidt's method is hopelessly flawed: Finding out how *true* moral judgment works is impossible if so many ideas about morality that are conflicting with each other are incorporated into the database. This would be like the attempt to find out how chess expertise works while remaining agnostic of what makes somebody a good chess player and making no difference between data gathered from professional chess players' behavior and data gathered from laymen's behavior.

However, I think we might regard this kind of disagreement about philosophical premises more superficial than one might initially assume. Remember that I introduced a way of dealing with assessment of moral truth that is independent of our moral intuition earlier when dealing with intention as a factor within morality: I highlighted that understanding SIM and MFT as an account of the phenomenology of moral judgment would keep its main points alive while considerably reducing its philosophical baggage. Instead of being an account how moral sentences (understood as culture-relative) come to be true, it can also be understood as an account of why certain moral truths (understood as non-culture relative) are not recognized by large groups of persons, for example why slavery and oppression of colored people in the Southern states of the US was so widely accepted through a considerable part of US history. Haidt's theory would become a theory about moral opinion instead of moral truth. One could take this approach to neutralize the MFT in terms of moral relativity and moral objectivity, too. If one regarded the MFT and the SIM as theories about moral opinion, then the method of treating all judgments the same, even if they contradict, would cease to be troublesome for objectivists. In contrast to polluting the data about *true* morality and obstructing findings about how true morality works, incorporating wrong moral judgments into the database of moral *opinion* is very legitimate. It even helps to explain why certain wrong moral sentences seem so appealing to us. Of course there would be a tough bullet to bite: A substantial amount of people would adhere to wrong beliefs about morality, at least when it comes to whether to regard norm violations in the disgust domain as morally wrong or not - regardless of what will come out to be the moral truth about this matter. On the other hand, if we take a diachronical perspective, we will find that today's moral values – no matter what culture we are talking about - would force us to regard a lot of beliefs of our ancestors as wrong not only beliefs about morality. Just take beliefs about physics as an example. But if physics proceeds, why should morality not do so, too? This thought is a good

catchword for an objectivist account of morality and value in general that tries to incorporate the intuitive advantages of relativism into a "objectivism (or: *realism*) with a human face".

Example #3: Hilary Putnam's "realism with a human face"

Before I conclude this chapter, I would like to introduce an account of moral truth that I personally find very promising for making Haidt's theory acceptable for open minded objectivists and realists – even in a non-phenomenalized version. It is Hilary Putnam's "realism with a human face". As its slogan-like name suggests it is promoting (among other aspects of truth) a kind of realism that brings all the benefits of realism/objectivism (Nazis *are* wrong and not only in *our* system of moral values but as a matter of fact) together with the benefits of moral relativism (how to deal with questions about morality where there is no consensus or knock down argument?). Unfortunately, one has to accept some rather strong premises to accept this approach to truth. Please understand that I run through these premises without spending too much time justifying them - I merely want to introduce this approach in general, not in detail.

The *first point* to introduce is that according to Putnam, whether a theory is true is dependent on its fit to certain *epistemic values* like *coherency*, *relevance* and *Ockham's razor*. Here, he is leaning strongly to an account of truth that is comparable to Quine's in "On what there is" (Quine, 1953): there is a set of criteria that tell us which description of the world is the best. The best (a.k.a. true) theory about the world is the one that fits best to these epistemic values.

What is special in Putnam's case (and the *second point*) is that Putnam does not regard this set of values as fixed points. Instead, he sees epistemic values not only as constituting what is to be regarded as a fact but also as being influenced by exactly what these facts are. The measure for truth is determined by what we know about the world. Take for example the epistemic value that in order to be true, a sentence must not be in direct contradiction to the works of Aristotle and the Bible that was commonly held in the Middle Ages. Most people will probably argue that this value has been discarded by now due to our knowledge of facts and not due to an unjustified shift of values. The relationship between facts and values according to Putnam is one of mutual justification and improvement. The more we now about the facts of the world, the better we will be able to scrutinize our set of epistemic values, and as our refined epistemic values give us new clues where before we were clueless, the more we will then in turn come to know about the world.

This is the foundation of Putnam's concepts of truth and reason (*third point*): *reason* is the capacity to assess sentences about the world according to a given set of epistemic values (this normative nature of reason is why Putnam regards it as not reducible to science) – *truth* is an idealistic concept of what our account of the facts about the world would be given that we had refined our values infinitely.

This brings us to the *fourth and last point*: moral values behave just like epistemic values. There is no reason to treat both kinds of values differently. They are just as fact-dependent as epistemic values and they "create" facts about the world that we can judge implausible.

If we accept all four points, we can give the following account of moral truth: "real" moral truth is an idealistic entity, just like the "ideal physics" that is referred to by reductionists about scientific theories all the time[45]. However, it gives us the capability to make sense of something that one might call "moral progress": certain moral sentences and ideas just fly in the face of reason (as soon as our epistemic values have developed far enough) and can be regarded as brutally false in hindsight.

Likewise, we can regard our current moral value systems as competing theories about the world, as sets of declarative AND normative sentences that compete to describe the world in the best way. If we accept a notion of "rudimentary truth", as "true according to a set of facts and values given by a certain theory" and regard moral theories as culture dependent, we end up with something that can get along very nicely with Haidt's presumptions about moral truth that is still objectivist. Note by the way that the concept *reason* as defined by Putnam fits very nicely to what I discussed in the chapter about intuition: "reason" is here a standard of justification, not a psychological category.

Through such a perspective, MFT and the SIM would become an account of how folk theories of morality work and develop. They would become an important part in explaining how moral progress works, which would become an important part in making Putnam's brand of objectivism more plausible. This brand of objectivism would allow to read Haidt's moral theory as being about true and false moral judgments, but only in the context of a set of norms (or theory of morality). Which of these sets of norms is the best will be decided in the course of history, and most of them will land on the same pile as phlogiston theory, the Ptolemaic world view, and alchemy.

I hope that this section demonstrates how, under the assumption of fitting additional premises, Haidt's theory of morality can be interpreted in an objectivist way without changing its scope or main message. It can then even be employed to make an objectivist standpoint like Putnam's more plausible by giving explanations for how competing moral theories develop and interact.

[45] This ideal physics has unlimited predictive and explanatory power, so high that it is going to make psychological terms superfluous – at least according to Churchland (1981). It is highly unfortunate that this *ideal* physics it is sometimes thought to be *actual* physics by careless students who thereby misunderstand reductionism.

3.6.4. Conclusion

Let's sum up what to take home from this chapter. The points made here served mainly to show that Haidt's tendency towards a relativist view is in fact *just a tendency*. It is pretty unproblematic to interpret Haidt's theory in a classical objectivist way as soon as one phenomenalizes Haidt's theory as shown in the chapter about intention. If one is willing to accept certain premises, there is even a way to adopt an objectivist reading of Haidt's moral theory without phenomenalizing it. However, the character and the topic of the MFT and the SIM change dramatically depending on what metaethical standpoint we choose. While from the perspective of a Putnam-branded objectivist, it would be a theory about folk moral theory, from the perspective of a tough minded objectivist it would be a theory about mere opinion. To claim its character as being about really (and not just "the best theories that we have at the moment") true and false moral judgments, one would have to ascribe to Haidt's moral theory a kind of moral relativism like the one that Gilbert Harman envisions.

Let me point out that MFT and SIM can therefore be considered to touch the philosophical matter of moral relativity only peripherally, even though they might look like doing so more radically on first sight. Already minor interpretative twists immunize Haidt's theory of moral judgment against attacks based on objectivist accounts of moral judgment. On the other hand, The MFT and SIM could help to make certain aspects of metaethical thought about moral truth more tangible, like in the case of "moral bargaining" – a process that the SIM could help us to get a much better idea of, and in the case of "moral progress", a process that would be necessary for Putnam's realism to make sense.

3.7. Summing up a few points

This concludes the section of this work that deals with the detection of philosophical commitments as presented in the introduction. Let me recapitulate briefly the results and put them in context with the theoretical groundwork laid in the introductory chapter.

In this introductory chapter I made the point that empirical research of moral judgment is bound to make what I called *philosophical commitments*. With this label I want to refer to the exclusion of certain interpretations of a term by using, defining or operationalizing it in a specific way contrary to these interpretations. Even though these commitments are primarily conceptual as they merely imply that one is using a word in a certain way and not in another, when it comes to the matter of moral judgment we can often associate certain philosophical positions with this linguistic decision. The

decision to use the term moral judgment in a certain way therefore implies the rejection of certain ways to understand the term that are suggested by philosophical positions.

I went out to look for such conceptual decisions and I found some of them. However, this was not enough for me. I was also interested in the effects that these philosophical commitments have for empirical research on one side, and for philosophical arguments that rely on this empirical research on the other. I wanted to explore the interconnections between the empirical and the philosophical approach to investigating moral judgment – on both sides. So let me review what I found out:

I found different kinds of philosophical commitments that appear in psychological and neuroscientific research about moral judgment, I found that these commitments can have consequences for psychological and neuroscientific research, I found that they do indeed affect philosophical arguments trying to exploit empirical results and I found that choosing the right philosophical foundation can strengthen the empirical point made by a theory.

There are philosophical commitments in moral psychology

I showed that Haidt's approach to moral judgment does in fact exclude a lot of possible understandings of morality and its constituents *by definition*. These commitments were: the exclusion of certain cognitive theories of emotion through the adaption of Damasio's neuropsychological Somatic Marker Hypothesis; the presumption of a situationalist picture of morality through the decision to regard judgments about actions to be pertaining to situations (and not to character traits) in operationalization; conflicting presumptions about the syntactical structure of moral judgment in different experiments through different operationalizations. The terms *intuition* and *reason* showed to have a double meaning in common language and philosophy of which Haidt only adopts one side - the way he uses the concept not relatable to all facets of philosophical use of the term.

I also showed that certain aspects of morality that could be considered philosophically challenging in the SIM, namely moral truth and intention sensitivity to be much less problematic than they look like on a first glance: moral truth showed to be a non-issue once a sufficiently lenient interpretation of the SIM and MFT was adopted; intention sensitivity proved to be an interpretational issue, too, mainly because no presumptions were made on Haidt's side.

These philosophical commitments do affect empirical philosophy of moral judgment

I showed that Haidt's philosophical commitments have significant effects on philosophical applicability of his results: misunderstanding Haidt's concept of intuition led to wrong conclusions about the nature of morality; differing concepts of emotion applied to the example study led to contrary conclusions from the same results; overlooking certain commitments in operationalization of moral judgment made Haidt's embrace of virtue ethics self-refuting; overlooking differing syntactical standards turned one argument of Jesse Prinz's for sentimentalist ethics into a non-sequitur.

Overlooking these philosophical commitments affects psychological research

I showed that the understanding of *emotion* determined whether empirical data does indeed suggest moral judgment being evolutionary preset; I showed that a virtue ethicist understanding of moral judgment would undermine the empirical validation of the MFT; I showed that by overlooking the unclear relationship between graded moral judgments and binary moral judgments studies that are technically not supporting the MFT are regarded as doing so.

I consider my research hypotheses as validated from this point. In what remains of this book, I focus on how to get moral philosophy and moral psychology to cooperate more effectively. A very unsettling prospect that these observations suggest is that research about moral judgment is eventually something quite arbitrary and purely dependent on use of debatable concepts. In what follows I offer a way to escape such a conclusion. As an alternative, I point out that we will be able to refine our methods by interconnecting different ways of measuring moral judgment, broadening the empirical base for a joint theory of morality. I hope thereby to give some idea how different kinds of empirical evidence could be used to tackle certain philosophical points much better when combined in a way that does not allow the philosophical theory to "escape" by simply reinterpreting certain foundational terms. Furthermore, I introduce an empirical approach to bridge conceptual gaps in moral judgment, using data gathered by myself and colleagues as an exemplar.

4. **Pinning down morality – an outlook**

So far, this thesis has mainly featured destructive arguments. Its main topic has been the revelation of incoherencies and contradictions in seemingly promising attempts to successfully merge empirical with philosophical research of moral judgment. One might say that my global target of criticism has been an "anything goes" kind of concept usage in research about moral judgment. The corresponding understanding of *moral judgment* could be framed as the conjunct of two ideas: firstly that moral judgment embraces very different types of judgments, and secondly that drawing conclusions about *moral judgment* from sentences about *moral judgment* is generally legitimate. In the introduction, I specified why the first idea does not necessarily entail the second idea but rather the contrary. I offered reasons why in the context of interdisciplinary research on moral judgment, the second idea is very likely to be wrong. The third chapter demonstrated that the second idea of the "anything goes" mindset is in fact wrong and that taking an "anything goes" stance in empirical philosophy, but also in empirical research itself leads to self-contradictions, incoherencies and inconclusiveness. In sum, chapter 3 gave plenty of evidence that the "anything goes" understanding of moral judgment is simply inadequate for interdisciplinary research.

Let me point out at this point once again that I remain explicitly agnostic about in which situations the term *moral judgment* is correctly applied. I do not want to prescribe anyone how to use the term *moral judgment*. If however I want to refrain from arguing that there is one and only one correct way of using the word *moral judgment* but still want to uphold my refutation of "anything goes", I am faced with a challenge. How should interdisciplinary research proceed? How should interdisciplinary research deal with the new obstacles that I presented in the third chapter? This brief last chapter points out some answers. It suggests how a pluralistic[46] view of moral judgment can not only deliver coherent results but how it can in fact enhance our capacity to successfully conduct interdisciplinary research.

To illustrate this thought, I would like to come back to the example of the biologist from the first chapter who defined *sea fish* much more narrowly than the gastronomist who defined it as "animals you fish out of the sea". When asked about the reason for this more narrow definition, he might do

[46] With *pluralistic view of moral judgment* I mean exactly that there are different ways to use the term *moral judgment* that are equally legitimate but at times divergent or even mutually exclusive. Understanding *moral judgment* as being about character and understanding *moral judgment* as being about actions imply two different ways of using the term *moral judgment*. Taking a pluralistic stance would mean to acknowledge the resulting difference in meaning while not assuming that one way to use or to understand the concept is *wrong*.

exactly the same as I did in the last chapter. He might explain that the gastronomist's "anything goes" concept of *sea fish* is incoherent in a number of important ways: that it incorporates animals with gills as well as animals with lungs; that it incorporates vertebrates as well as invertebrates. He would certainly suggest that it would be more fruitful to acknowledge these differences than to ignore them. "Animals you fish out of the sea" is a much too broad definition for drawing conclusions from one specimen about the other. What is needed to describe the ways and workings of marine life adequately is an understanding of *sea animals* that incorporates more specific definitions like crustaceans, mammals, reptiles, and the like and acknowledges differences and communalities between them.

The view on *moral judgment* that I developed in the first three chapters can be likened to the last point of the biologist, the one about sea animals. There are different types of sea animals - just like there are different types of moral judgments. All of them can be called *sea animals* legitimately – just like there are various legitimate ways of applying the term *moral judgment*. However, our capacity to draw conclusions about sea animals in general from one type of sea animal is very limited. There are important differences between sea animals that have to be acknowledged. In regard to moral judgment, this was the main claim behind my research hypotheses. The most important point of the biologist however is the one about scientific fruitfulness: recognizing differences between sea animals allows for explaining them better. This part of the thesis is dedicated to explain in how far drawing differences between types of moral judgments helps to explain moral judgment better.

My argument proceeds as follows: Firstly, I introduce a framework about what it means to explain something better. Secondly, I present two ways in which appreciating differences between types of moral judgments enables researchers to discover new results. This is on the one hand by actively embracing the conceptual scrutiny that I showed to be necessary in interdisciplinary research in order to argue in much more sophisticated ways. On the other hand it is by empirically relating different nuances of moral judgment to each other. The latter is exemplified by an empirical study that I conducted with Professor Stephan Glasauer from the Center for Sensorimotor Research at the Department of Neurology and Gloria Benson from the Neuro-Cognitive Psychology master program of the Ludwig-Maximilians-University Munich.

So let me begin this last part of my thesis by coming back to the idea of *explication* as understood by Rudolf Carnap.

4.1. Explication again

The attentive reader might still remember from the introduction that there are four determinants of explication: *scientific fruitfulness*, *simplicity*, *exactness* and *similarity*. For the purposes of the introduction, it sufficed to explain only *scientific fruitfulness* and *similarity*, leaving the other two for later. The moment has now come to have a closer look on the two remaining determinants while also refreshing the reader's memory about *scientific fruitfulness* and *similarity*.

Take again the biologist's differentiated account of *sea animals* and the gastronomer's coextensive[47] term *sea fish*. It is a mixture of knowledge about the world and knowledge about our needs for describing the world that make the biologist's differentiated account attractive and possible.

As the differentiated concept of *sea animals* can for example be used more effectively to formulate scientific laws than the gastronomer's *sea fish,* it has the advantage of higher scientific fruitfulness. This is the reason why zoologists chose to introduce all the differentiations of sea animals: it made their job (explaining behavior of animals) easier.

If two explicata are of the same fruitfulness, it is again the practicability of the concept that will decide which one is adopted. This time, it is the **simplicity** of the explicatum that decides which one is better. There are two kinds of simplicity:

> "The simplicity of a concept may be measured, in the first place, by the simplicity of the form of its definition and, second, by the simplicity of the forms of the laws connecting it with the other concepts." (Ibid., p. 7)

Admittedly, the definition of the differentiated term *sea animals* is more complex than the pretheoretic term *sea fish*. After all, it has to contain the important taxonomic differentiations that allow the biologist to do his work. However, as scientific fruitfulness is taken to be of higher priority than simplicity, I take this criterion to be overruled in our case.

The last one of the "pragmatic" rules for explicata is the one that an explicatum will have to display a level of **exactness** that allows it to play a decisive role in scientific explanation. For a differentiated concept of *sea animal*, this is clearly the case: not only is it possible to exactly determine whether or not an animal is a sea animal, but also what kind of sea animal it is.

[47] Everything the gastronomer calls *sea fish* is called a sea animal by the biologist and vice versa.

Last but not least on the list: **similarity**. In our case, we see that even though there is a clear difference between *sea animals* and *sea fish*, the terms remain coextensive.

In consideration of these four requirements of explication and their ranking of importance, the key condition for a differentiated account of moral judgment to explain moral judgment better than the "anything goes" account or nonpluralistic views consists in increasing scientific fruitfulness. The other criteria are either secondary in their nature, as simplicity, or they can be easily recognized to be met: the criterion of exactness is met in so far as that a more fine grained pluralistic approach of moral judgment allows for the same exactness as the "anything goes" approach **plus** a taxonomy of moral judgments along clear rules . The criterion of similarity is met in so far as what counted as a moral judgment in the "anything goes" account still counts as such in the differentiated account.

In the next two sections, I introduce two ways in which this more fine grained, subtle conceptual approach to moral judgment promises to be much more fruitful than the status quo of "anything goes": on the one hand the more fine-grained construction of empirically minded philosophical arguments, on the other side the empirical investigation of relations between different types of moral judgment. I furthermore assess whether they can indeed be regarded as instantiations of scientific fruitfulness. On the basis of these assessments, I argue for the differentiated account's superiority.

4.2. Empirical knockout through recognition of plurality of meaning

4.2.1. The idea

The first benefit of the more fine-grained understanding of moral judgment that I want to present is the capability for more sophisticated arguments in empirical philosophy. As could be seen in the third chapter, these arguments often tend to take the form "According to philosophical theory X, moral judgment has property *a*. Study Y shows that moral judgment does not have property *a.* Accordingly, theory X is wrong – it fails to account for the fact moral judgment has property *a*". What I have demonstrated above (also in chapter 3) is the issue that *moral judgment* understood as by theory X can be something different from *moral judgment* understood as by study Y without people noticing it. The most evident and most important change for the methodology of empirical philosophy would have to be that the first step of an argument would have to be assuring that the cited empirical data and the philosophical theory in question share the same understanding of *moral judgment*. Of course, on the surface this makes empirical philosophy weaker: If I have to argue for something this normally implies that there are reasons to be found against my conclusion. Furthermore, philosophical theories can be customized easily to be unfeasible for a given empirical approach: assume someone wants to defend her philosophical theory X. All she will have to do is to state something like: "I see you thought that with *moral judgment* I mean only this: [one kind of definition]. But actually, what I mean with *moral judgment* is also that: [another type of definition]". The empirical argument will be in vain. But this reinterpretation of her philosophical theory will come at a prize. In order to evade the attack from one side, she will have to become more explicit about certain aspects of morality at other places of her theory. It is here where the empirical philosopher will have to attack next. And even then, she might adapt her theory further. But this will entail more clarifications and refinements at other places, and the empirical philosopher will attack at these places - until there is no refinement to make anymore. And then, finally the empirical philosopher would have reached her goal of showing that theory X is wrong. This is what I mean when I state that empirical philosophy needs a more refined way of arguing in order to be actually effective. Another way of being more refined would be to look where philosophical theories are already very concise and make this spot the spot to strike first. An example for such an argument could be constructed against Jesse Prinz's sentimentalist theory that can be based on the conceptual scrutinizations of the previous chapter.

4.2.2. An example: intuition insensitivity in emotionally impaired persons and Prinz's sentimentalism

Remember the section about *emotion* from the last chapter: In this section, I presented Prinz's perceptionist theory of emotion as a conceptualization of *emotion* that is on the one hand more extreme than the one employed in SIM and MFT, but which allows on the other hand to draw several additional philosophical conclusions from the database of both approaches. The most precious philosophical outcome is a sentimentalism about the concept of morality itself. Note furthermore that as stated in 3.2.2, Prinz's perceptionist concept of emotion is a prerequisite for this metaethical theory.

Recall further section 3.5, the one about *intention*. In this chapter I found two ways of introducing intention-sensitivity to the MFT and the SIM – on the one hand, intention-sensitivity could be introduced as intention-sensitivity of emotion. On the other hand, it could be introduced as a post-hoc evaluation of the judgment itself. I pointed out that for a perceptionist account, the former way would not be viable, as it would entail a prototypical cognitive understanding of *emotion* by making judgments part of moral emotion. An assessment of mental states as a prerequisite of moral emotion would entail that this particular moral emotion would not be atomic[48] in the way that Prinz's theory would need it to be.

Through this line of thought it can be established that Prinz's sentimentalism predicts moral emotions to be intention-**in**sensitive. Note how this prediction is a result of recognizing differentiations in concept use: on the one hand Prinz's dependence on a perceptivist concept of emotion stems is necessary to gain empirical support from studies like Wheatley & Haidt's (005) and Haidt et al. (1993). On the other the necessity of intention-sensitivity of moral judgment originates from the need to keep his concept of morality coherent with philosophical tradition, common language and jurisdiction.

But now Prinz's theory has a problem: in contrast to its newly derived prediction, emotionally impaired persons have been shown to be far less intention-sensitive in their moral judgments than the comparison group in a study by Young et al. (2010).[49]

[48] With atomic, I mean "not further analyzable"

[49] The reader will probably have noticed that in the chapter about intention, I argued along evidence for the contrary. I fear that there is indeed evidence for both sides – I am sorry to have to offer as the only comfort that I am not particularly interested which sort of evidence is the *right* one. I merely take the particular pieces of evidence as illustrations of my theoretical point.

Young et al. let people with lesions in the VMPFC area who were emotionally impaired judge intended and unintended harms, similar to the stimuli used to test for the Knobe Effect discussed in chapter 3.5. The VMPFC patients rated the scenario in which a person intended to hurt another but accidentally failed to do so almost as permissible as not intending to harm and not harming and even less permissible than harming by accident. This was in strong contrast to both control groups of people with other brain lesions and unimpaired persons. The fact that people with impaired emotion did perform so badly at incorporating intention into their moral judgment suggests strongly that intention is assessed as a part of an emotion[50].

In Prinz (2006), Prinz explicitly cites studies with emotionally impaired patients as evidence for the importance of emotion in moral judgment. The term "emotionally impaired" therefore is highly probable to be applicable to Prinz's concept of emotion – if not emotion theorist Jesse Prinz has misattributed his own theory. This brings Prinz into trouble, as the data suggests that his radical perceptivist theory of emotion might not work out, as the cited results indicate that intention detection *does* play a role in moral emotion. This is a serious problem for his sentimentalist moral theory since the important role that his restrictive understanding of *emotion* played cannot be upheld by a concept compatible with Young et al.'s results: according to these results, we need knowledge about intention for moral emotions. In Prinz's terms: in order to have the proper elicitation files at hand we need an assessment of intention.

Of course, this is only one study and there are many ways to explain away empirical results. The study could be flawed, Prinz could speculate that VMPFC-patients also have other deficits that we do not yet know about, etc. However, I regard this case a quite strong (at least stronger than many of the exemplary cases that I criticised) since the concept use matches the one of Prinz quite closely. And as I said, adapting his own theory would come at the prize of becoming more vulnerable to empirical data at another point of his theory.

4.2.3. Scientific fruitfulness of recognizing plurality

This rather ping-pongy argument that oscillates between philosophy and psychology shows how empirical philosophy in the end benefits from greater conceptual scrutiny. It helps immensely at specify-

[50] Under the premise that moral judgment is indeed based on an emotional response. This is however a premise that Prinz will have to accept.

ing empirical predictions of philosophical standpoints – and offers therefore new ways to argue for or against them.

Does this example show that a more complex concept of moral judgment is more scientifically fruitful? I argue that it does. Remember Carnap's definition: "A scientific concept is more fruitful the more it can be brought into connection with other concepts on the basis of observed facts; in other words, the more it can be used for the formulation of laws.". On the first glance, it might seem dubious that the section above showed that a more complex concept of moral judgment could be used for the formulation of laws. But remember how I got to the point that Prinz's theory is incompatible with Young et al.'s results: the key was that I was able to conclude that Prinz's theory necessitates the intention-insensitivity of emotion. I was only able to do so by acknowledging on the one hand the specific aspects of Prinz's concept of emotion, the necessity of these aspects for the validity of his metaethical theory and on the other hand the fact his need to incorporate intention-sensitivity outside of the process of undergoing an emotion.[51] The statement "all moral emotions are intention-insensitive" that I was able to derive from this line of thought is exactly the type of statement that one could describe as law-like. It is an empirically testable generalization, the terms *moral emotion* and *intention-sensitive* are referring to observable and measurable properties.

Prinz's concept of emotion and its meaning for his metaethical theory allowed to derive a law-like statement – and to have it falsified. This strongly suggests the higher scientific fruitfulness of a more complex conception of moral judgment. But let me however give you some even stronger evidence: in the next section, I present a way to create quantitative laws on the basis of a more refined concept of moral judgment.

4.3. Psychophysics of moral judgment

In chapter3.3, when dealing with binary and graded moral judgment, I introduced the similarly different approaches to understanding the nature of *heat*. While "warm", "cold" and "hot" are binary concepts that are normally used by persons, "35° Celsius" or "72° Fahrenheit" are quantitative concepts that are normally read from some form of thermometer. Note that with quantitative concepts

[51] If I took an "anything goes" stance on either his concept of emotion, its importance for his ethical theory or the importance of intention sensitivity for moral judgment, the results would not affect Prinz's sentimentalism as a whole. It would be enough of a reaction to allow for intention sensitivity of moral emotions to evade criticism.

we can do things that we cannot do with classificatory concepts. We can formulate much more precise laws, having the whole power of mathematics applicable to our concept. Generally, quantitative concepts can be regarded as more scientifically fruitful. But now imagine we are not interested in *physical* phenomena but *psychological* phenomena concerning heat. On the one hand the phenomenal concept of heat is now more easily relatable to other psychological states – on the other, its classificatory nature makes work very unsatisfying, compared to the work the physical quantitative concept. If there were a way to make the phenomenal concept more precise by somehow pinning it to the physical concept via a set of "translation rules", we could make use of the practicality of the physical concept in order to better understand the working of our phenomenal grasp of heat. And indeed, there is a branch of psychology that is doing exactly this: Psychophysics.

4.3.1. Psychophysics

Sharpening of everyday concepts through empirical investigation is tangible. But how can we expect the sharpening of psychological or mental concepts to look like? On the following pages I present a branch of psychological research that has a very long tradition – even though psychological behaviorism did in fact interrupt the development of the field for a few years in the 20th century. I am talking about the field of psychophysics.

The basic idea of psychophysics was to find out lawlike relations between the subjective appearance and the actual properties of physical entities. Gustav Fechner described it in the following way:

> Psychophysics should be understood here as an exact theory of the functionally dependent relations of body and soul or, more generally, of the material and the mental, of the physical and psychological worlds.
>
> We count as mental, psychological, or belonging to the soul, all that can be grasped by introspective observation or that can be abstracted from it; as bodily, corporeal, physical, or material, all that can be grasped by observation from outside or abstracted from it.
>
> Fechner (1966), p. 7.

Note that what is important is a *functional dependency*, the term function understood here in its formal-logical way as an unambiguous relation between two classes of entities. The purpose of the

whole endeavor is to specify *exactly* and in an unambiguous way how external, physical entities and their mental representations are related:

> In general, we call the psychic a dependent function of the physical, and vice versa, insofar as there exists between them such a constant and lawful relationship that, from the presence and changes of one, we can deduce those of the other.
>
> Fechner (1966), p. 7.

The aim of psychophysics is to be able to conclude specific properties of the mental representation of a stimulus from specific properties of that given stimulus. The properties of the stimulus are controlled; the actual target of scientific investigation is the representation. The entity that is described by psychophysics is therefore how certain features of the world are represented. What is explained by pschophysics is the representation of a feature of the world – for example of brightness, colour, or heat.

As an illustration of the methods of psychophysics, consider the following example: the representation of length in the case of a Müller-Lyer illusion[52].

The Müller-Lyer illusion is originally known as the illusion that two lines of the same length are perceived as being different dependent on whether there are fins attached to the line or not: The smaller the angle between the fins and the line, the shorter the perceived length of the line.

Imagine we want to find out more about the perceived length of a line in a Müller-Lyer illusion. Imagine we would like to have a more refined idea of what a Müller-Lyer-Illusion is. A typical psychophysical approach would look like this:

In a number of trials, a subject is asked to make a forced choice judgment which of two lines is longer: line A (fins in 45 degree angle) or line B (fins in 135 degree angle). While the length of line A stays constant over all trials, the lengths of line B varies.

In the end, the experimenter will be able to derive from the answers of the testee a function of the length ratio between both lines to the proportion of "longer" answers given by the testee. A typical result would be a sigmoid function like in Figure 7.

[52] This hypothetical example comes from *Psychophysics – a Practical Introduction* (Kingdom, Prins, 2009).

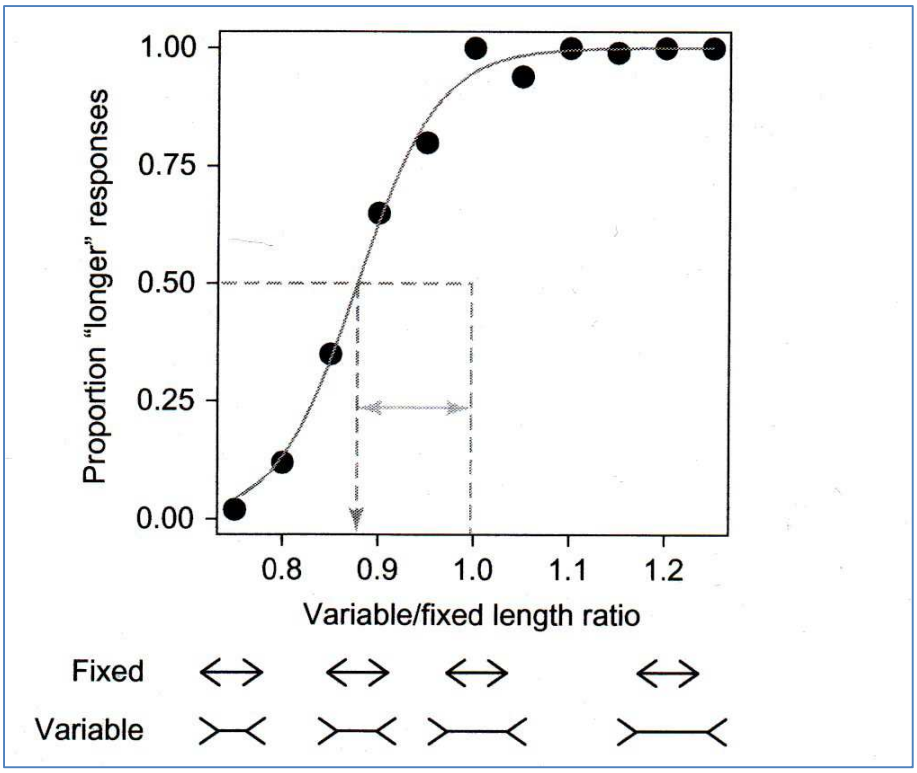


Figure 7: Hypothetical psychophysical curve for variating line lengths in the Müller-Lyer-illusion from (Kingdom & Prins, 2009), p .19

I would like to argue that this understanding of the Müller-Lyer illusion offers a better explanation than the original one. Let us have a look at Carnap's criteria one last time: the original idea that we had of the Müller-Lyer-illusion could in Carnap's terminology best be described as "pretheoretical": the illusion was described with the help of classificatory concepts "fins in less than 90° angle" and "fins in more than 90° angle", as well as the comparative concept "appear longer". While this vocabulary was exact enough to describe that there is a psychological effect and to differentiate this effect from other psychological phenomena it was by far not enough to understand how the Müller Lyer Illusion *works*.

The new description/definition of the phenomenon will be much **exacter** and easier to correlate to neurobiological unnderpinnings and therefore be **more scientifically fruitful**, while at the same time referring to the same psychological phenomenon as the pretheoretical concept. The **similarity** condition for a better explanation is therefore met, too.

In summary, psychophysics work like this: psychological concepts are set into relation to physical events and become describable in terms of these physical events. As physical events often allow for more fine grained description, psychophysics will allow for more fine grained description of psychological entities. This more finegrained description of the psychological process can be expected to

yield increased scientific fruitfulness as it becomes easier relatable to quantitative measures of other research methods (for example the blood oxygenation level dependent (BOLD) signal measured by functional Magnetic Resonance Tomography (fMRI)). In the next section I argue that this method allows to increase the scientific fruitfulness of an understanding of moral judgment, too.

4.3.2. Psychophysics of moral judgment – why

Let me assemble the puzzle pieces which I mentioned so far. On the one hand, my arguments and exemplifications from chapters 1 and 3 have shown that there is a plurality of concepts of *moral judgment* in philosophy as well as in psychology. I demonstrated that this plurality can make interdisciplinary arguments as well as empirical research itself a tricky business.

Now followers of Carnap would very likely suggest to let science make progress on its own: indeed a steady ongoing precisation of the concept according to practical considerations as a quasi-automatic process would be a very nice prospect. Philosophers could easily lean back in this case and let science do its work. But I hope to have shown that science does not work that way. People indeed overlook important distinctions between the objects of their research and mistakenly treat them as the same. And there is still another reason why I am skeptical about the convenient idea that the problem will solve itself as science strives to the most practical, most scientifically fruitful concepts: especially in the research about neuropsychological phenomena, there is a bunch of different disciplines that employ different methods to discover new facts about the world: there are neurobiological methods that are invasive and noninvasive, there are psychological methods that focus on behavior, and those that focus on verbal behavior, and in the case of morality there is even philosophy with its own, non-experimental method. If we regard concepts as tools to describe the world, and I take Carnap to agree with this idea, then we should expect that disciplines employing different methods will find different conceptual tools most scientifically fruitful. Neurobiologists love concepts that can be operationalized independently of verbal behavior because it allows for animal research. On the other hand, exactly these verbal behaviors will allow psychologists to differentiate between psychological processes that would not seem differentiable otherwise. On the other hand, as I showed, different ways to differentiate verbal behavior have different advantages: Regarding moral judgment as binary allows us to easier relate moral judgment to behavior. Regarding it as graded facilitates determining how quantifiable external factors affect verbal moral judgment. Different interests and different methods make it hardly conceivable that philosophy, psychology and neurobiology will end up using

the same concept of moral judgment someday *automatically*. To illustrate the matter with the concept of *sea fish*: it seems equally unlikely that zoology and gastronomy will end up with the same classification of *sea fish.* Seashells, squids and prawns have all the gastronomic properties that one would expect of *sea fish*: they are light and easily digestible (at least compared to a meat plate), cannot be stored for long, go best with white wine. Except in a very specialized restaurant, it would just not be sensible to differentiate the concept of *sea fish*.

And this is where psychophysics appears on the stage. I introduced psychophysics as allowing for higher precision in the description of psychological concepts. Now let me explain why psychophysics of moral judgment *could* solve the problem described above.

If we assume that moral judgments are about states of our environment, we can assume that we could relate changes in the environment to the psychological process of moral judgment. But remember: there are different ways to operationalize moral judgment, different concepts of *moral judgment:* Moral Judgment1, Moral Judgment2, etc. Accordingly there are different relations from world to *moral judgment*, one for each concept of moral judgment. The trick that I now suggest is quite simple: instead of aiming at measuring the relation "moraljudgment1-world" or "moraljudgment2-world", moral psychophysics aims at building a psychophysical function "moraljudgment1-moraljudgment2" by comparing results of both to the same *world* stimulus. A comparable case would be the alignment of a Fahrenheit-thermometer to a Celsius-thermometer. Here, one would proceed just in the suggested way: Put the thermometers into a mass with the same temperature and see how the measurements relate to each other. This procedure is then repeated with masses of different temperature until a relation of sufficient exactness has been established.

Having such a relation at hand in the case of measurements of moral judgments would allow scientists and philosophers to bypass the difficulties that I described in the main part: I mentioned for example in section 3.3.3 that there is no translation rule that would allow Haidt to derive graded moral judgment predictions from his binary moral judgment model. A psychometric function would allow for exactly such a procedure. It would aim at finding out at which point and under which circumstances changes in scale measurements will have effects on binary judgment. Together with Stephan Glasauer and Gloria Benson, I undertook a first step towards such a psychometric function and towards demonstrating its scientific fruitfulness. I present its results on the upcoming pages.

4.3.3. Psychophysics of moral judgment – how

The idea behind the study originated in the need to find a way to relate different ways of measuring moral judgment to each other. For this purpose, we took a selection of stimuli from the literature of which we expected that they would cover the whole range of scale judgments from "absolutely morally impermissible" to "absolutely morally permissible" (see Appendix 2). The stimuli each consisted of a description of a situation that left a given agent with the choice between two options. These options were explicitly presented to the participant as "option A" and "option B". We divided participants into four groups:

- The first group was asked for a *binary* rating of *option A.*
- The second group was asked for a *binary* rating of *option B*.
- The third group was asked for a *graded* judgment of *option A* on a sliding scale with the endpoints 0 and 100.
- The fourth group was asked for a *graded* judgment of *option B* in the same way.

We constructed the relation "mean graded rating vs. percentage of positive binary judgments" on the basis of two kinds of stimuli:

- On the one hand, "normal" moral judgments, that is situations in which one option was in line with moral obligation while the other was in line with the self-interest of the agent.
- On the other hand "classical" moral dilemmas: situations in which both choices for action would have to be considered as doing something morally wrong.

A typical **dilemma** case would be the following case:

> You are a doctor. You have two patients who are both critically ill waiting for the same organ transplant operation. Both patients will die if nothing is done immediately.
> Patient A has been waiting for many years and has suffered a long time due to his illness. You feel he deserved the organ although his illness has weakened him so much that his chance of survival after surgery are very low.
> Patient B's chance of surviving the operation is much higher, since he has been waiting for only a year and endured far less suffering from his defective organ.
>
> You are faced with a choice between two options:

Option A: Even though Patient B has a much higher chance of surviving surgery - you decide to give the organ to Person A because he has waited much longer for the organ.

Option B: Even though Person A has waited much longer - you decide to give the organ to Patient B because he has a much higher chance of surviving surgery.

In the *normal* stimuli, there was no conflict between two moral obligations but rather a conflict between self-interest and moral obligation like in this stimulus adapted from Sommer et al. (2010):

At a department store you discover your dream clothes. On the way to the cash register you remember a report on child labor which you have recently seen on TV. The brand name of the clothes you want to buy was mentioned there, too.
You are faced with a choice between two options:

Option A: Even though this means financing child labor - you decide to buy the clothes

Option B: Even though this means not buying the clothes - you decide to refrain from financing child labor

The different types of stimuli, options and response options are presented in a graphical way in figure 9. All stimuli can be found in Appendix 2.

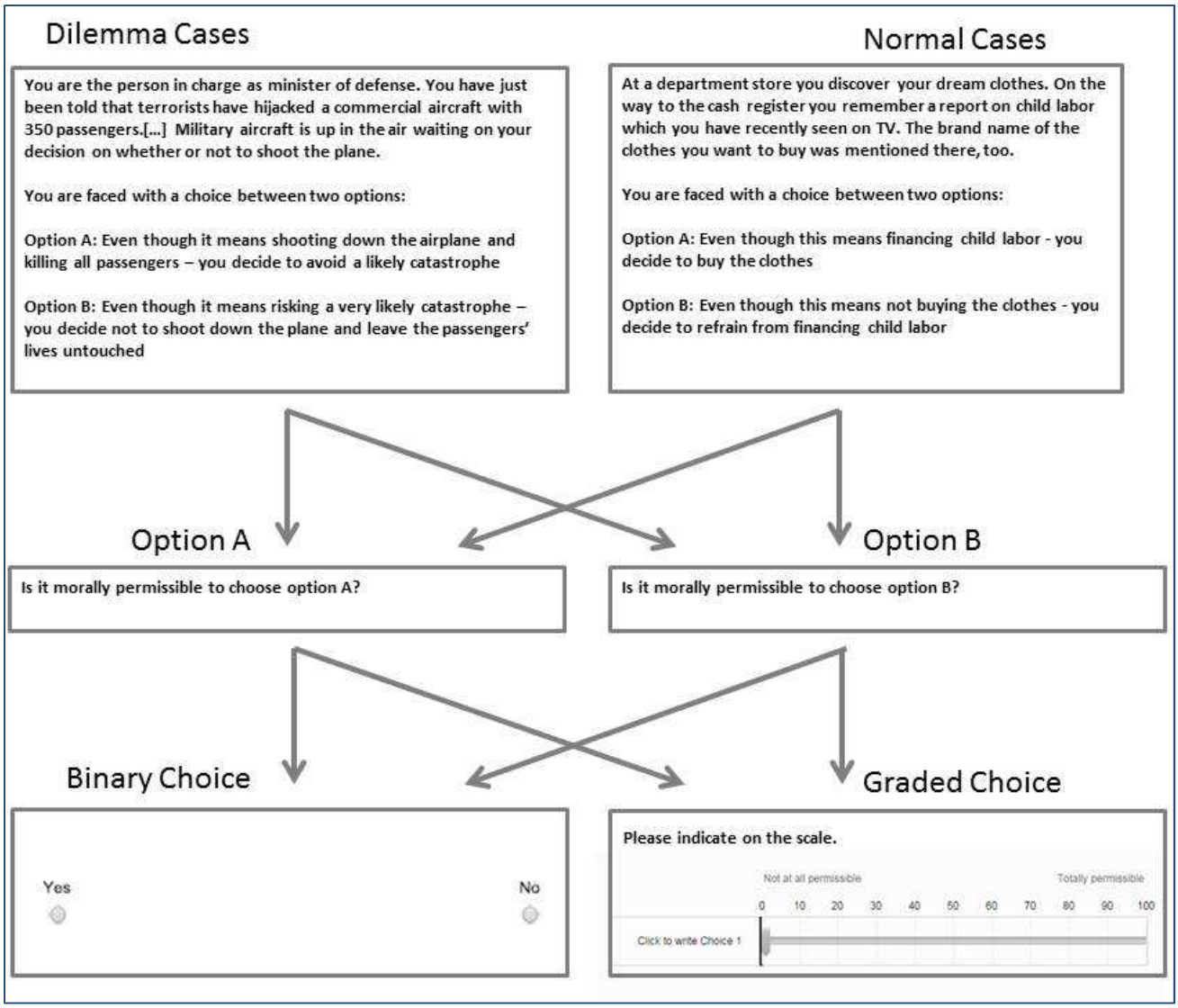


Figure 9: Construction of survey questions in the example study.

In order to demonstrate the scientific fruitfulness of a concept of moral judgment refined by our psychophysical survey, we decided to use our new found function to give us insights into the issue of which type of judgment is the primary one in computation of moral judgments. This idea is based on the following consideration: if there is indeed a stable mutual dependence of binary and graded judgment, one of them can be expected to be *primary* while the other is *derivative*. Either binary judgment is derived from a more basic graded judgment, or vice versa. Depending on which type of judgment is considered as primary, different predictions concerning the psychometric function can be expected. We therefore built two simple models in order to run computer simulations that we could compare with the actual results:

- One of the models assumed that each graded judgment was derived from a binary judgment with the help of a graded "imperfect duty value" of that given situation that would vary between subjects around a given mean in form of a normal distribution. Due to its similarity to

the Kantian syntax of moral judgments, we called it the "Kantian Model"[53]. In this model, binary judgments were the primary form of moral judgment.

- The other model derived binary judgments from graded judgments that were normally distributed around a given mean. Given the similarity to the Utilitarian concept of moral judgment presented in chapter 3.3, we called this model the "Utilitarian Model". In this model, graded judgments were the primary form of moral judgment.

Both simulations as well as detailed predictions of these simulations can be found in Appendix 4.

In summary, we expected three major benefits from a systematic, mathematical description of the relation between graded and binary judgments:

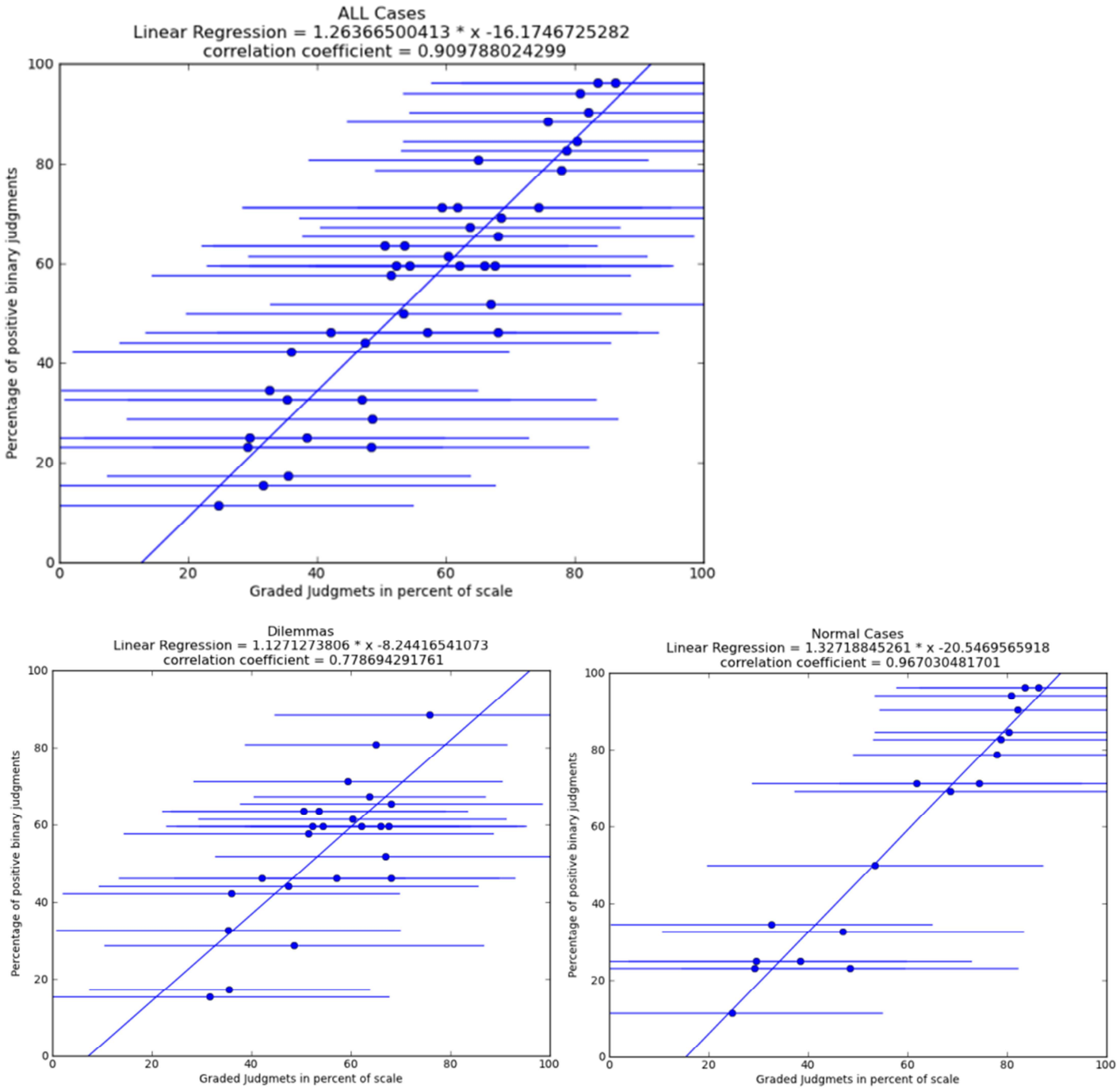
- Easier translation of experimental findings in the sense discussed above
- Insights into psychological processes behind the binary – graded extrapolation
- Insights into potential differences in processing between dilemma cases and normal cases

The results of the study (details in Appendix 3 and Appendix 5) indicate a surprisingly stable correlation between binary and graded judgment with slight differences of the psychophysical function between dilemma cases and normal cases (figure 10) .

The psychometric Function

When all cases are considered, the correlation has a linear regression of 1.26 *x – 16.17 and a correlation coefficient of .91. The correlation stays very clear when dilemma cases and normal cases are considered separately: The correlation coefficient of the dilemma correlation is 0.78, the one of the normal cases correlation is at 0.97. In the normal cases-condition, the slope of the linear regression ($y = 1.33*x – 20.5$) is slightly steeper than in the dilemma-condition ($y = 1.13*x – 8.24$) .

[53] That being said I want to emphasize here that neither Kant nor any Utilitarian does to my knowledge make any explicit statements about the issue. Their moral theories serve merely as a source of inspiration. What is especially important is the fact that their points are about the (in the broadest sense) *logical* derivation of *true* normative statements while this study deals with the causal-psychological process of deriving moral opinions. I discussed a very similar issue in chapter 3.1 about *intuition*.



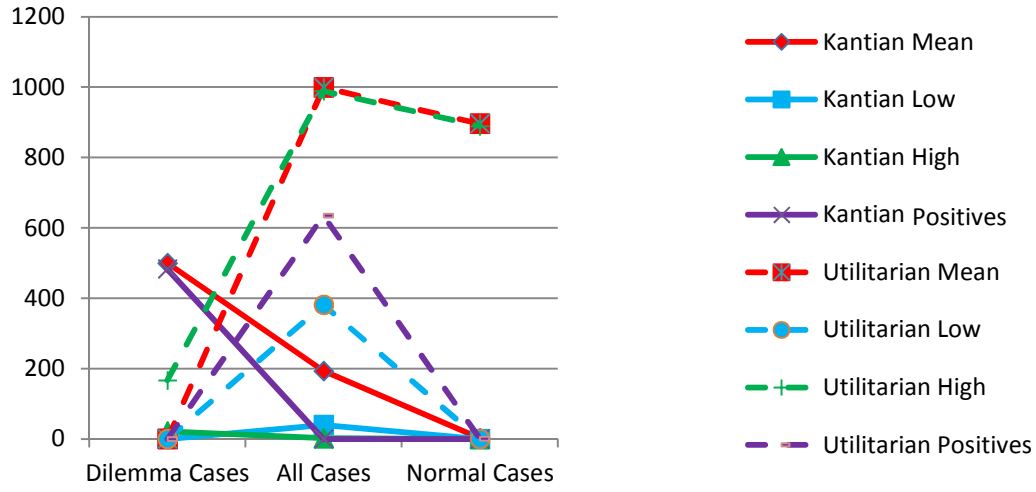
Linear regressions and coefficients of correlation of relation *„Mean graded judgment vs. Percentage of positive binary judgments"* for all cases (upper left), dilemma cases (lower left), and normal cases (lower right).

Figure 10: Psychometric functions *„Mean graded judgment vs. Percentage of positive binary judgments"* for different groups of cases

A highly interesting feature of the result is the "tolerance zone" in binary judgments for cases that are judged very low or very high in graded terms: For an average graded judgment between 0 and 12,8 as well as between 92,2 and 100, no change in binary judgment is to be expected for a change in graded judgment. This effect intensifies in normal cases while it is less pronounced in dilemma cases.

Fit of simulations suggests primacy of graded judgments

Concerning the computational order of binary and graded judgment, the predictions of the Utilitarian model proved to be much closer to the actual result in terms of coefficient of correlation and the linear regression of the discussed psychometric function as well as three other mathematical functions describing the distribution of graded judgments in the individual cases in dependence of percentage of positive binary judgments[54] (see figure 11).



Matches of predicted correlation coefficient (+- 0.1) and linear regression (y = +- 10 for x=0 and x= 100) in a thousand simulation runs for Kantian and Utilitarian model for *"Percentage of positive binary vs. Mean graded judgment"* ("Mean"), *"Percentage of positive binary vs. Percentage of graded judgment below10 %" ("Low"), "Percentage of positive binary vs. Percentage of graded judgment above 90 %" ("High"), "Percentage of positive binary vs. Percentage of graded judgment above 50 %"* ("Positives").

Figure 11: Correct predictions of psychometric functions in 1000 trials for different simulations

That being said, the same psychophysical function within dilemma cases alone proved extremely hard to predict and no satisfying results were delivered by any of the models. This suggests that the

[54] These functions were „Percentage of graded judgments above 50% vs. Percentage of positive binary judgments", "Percentage of positive binary judgments vs. Percentage of graded judgments above 90%" as well as "Percentage of positive binary judgments vs. Percentage of graded judgments below 10%"

small variations in terms of linear regression and correlation coefficient between dilemma cases and normal cases functions are not an artefact but do indeed imply some substantial difference in processing between both types of stimuli. Here, further research will be needed.

Differences between dilemma cases and normal cases

Between dilemma cases and normal cases, there were remarkable differences in the slope of psychometric functions, in the correlation coefficient, as well as in the fits of the different simulations. While some Kantian predictions proved to be better than the Utilitarian ones for the dilemma cases, the psychometric functions were described more adequately by the Utilitarian prediction when only the normal cases were considered. Another interesting difference between responses to the different stimuli types was that the shapes of distributions of graded judgments tend to look differently, with normal cases showing a much higher tendency of graded answers to be in the extreme low (0-10 out of 100) or in the extremely high (90-100 out of 100) part of the scale. This tendency can be expressed in the form of the functions "Percentage of positive binary judgments vs. percentage of graded judgments below 10" and "Percentage of positive binary judgments vs. percentage of graded judgments above 90", which show significant differences in correlation coefficient as well as in slope of the linear regression in the case of "Percentage of positive binary judgments vs. percentage of graded judgments above 90". These functions can be found in Appendix 3, the histograms of all cases displaying the distributions of graded judgments per case can be found in Appendix 5.

4.3.4. Benefits of psychophysics of moral judgment

So what is the resulting surplus of scientific fruitfulness about the meaning of *moral* given the results of the psychophysical example study made possible by a revised understanding of *moral judgment*?

In terms of facilitation of the formulation of laws, acknowledging the difference between graded and binary judgment must be considered extremely scientifically fruitful. Not only can the relationship between graded moral judgments and binary moral judgments be formulated in a lawlike way, it can even be quantified in form of a mathematical function.

This psychophysical function allows to draw new inferences: for example in Wheatley & Haidt (2005), the effect on disgust on graded ratings of morality was investigated (see section 3.3.2). In this study, a peculiar finding was made that is taken by Haidt as evidence for his (binary) model of moral judgment and by Prinz for his (binary) sentimentalist theory. A stimulus that contained no moral violation at all was rated a 2.7 out of 100 in absence of disgust. This number increased significantly to a 14.0 in the disgust condition. Prinz and Haidt regard this finding to support the hypothesis that (binary) moral judgment is substantially driven by emotion. Our study however shows that, based on the psychometric function derived from all stimuli, cases this change in moral judgment can be translated to a mere 1,57 percent increase in percentage of positive binary judgments at best. If the psychometric function that describes only normal cases is taken to calculate this change, **no increase of positive binary judgment is predicted at all**. Our results therefore show that superficial support for Prinz's and Haidt's theory vanishes as soon as a more complex conception of moral judgment is put in place. **Emotional manipulation can in this particular example expected to have no or only negligible effect on binary moral judgment.**

The psychophysical function furthermore allows for assessing predictions of two simple models of which one derived graded judgments form binary judgments and the other vice versa. This allowed for recognizing the utilitarian-syntax model of moral judgment as a much better explanation for our results. This finding could prove a challenge for moral grammar theories (for example in Mikhail, 2007) that explain (binary) moral judgment purely as the result of a series of binary computations. But this would take us too far at this point.

Lastly, psychophysical methods allow for describing differences between responses to *moral dilemmas* and responses to *normal morally relevant situation*s in new ways. They can be spelled out in terms of the function "mean graded judgment vs. percentage of positive binary judgments", but also in terms of other functions, namely "positive binary judgments vs. graded judgment below10 %" "positive binary judgments vs. graded judgment above 90 %", and "positive binary judgments vs. graded judgment above 50 %" (see Appendix 3 for more information).

In conclusion, refining our understanding of moral judgment by explicitly accepting graded and binary moral judgments as two distinct types of moral judgments allows for assessing empirically the relationship between both types of judgments. Due to its statistical nature, this assessment allows not only the evaluation of purported validation of models and theories, but even the validation of models of moral judgment on its own and furthermore the description of possibly different types of psychological processes involved. I take this situation to be a considerable improvement compared to the "anything goes" approach of moral judgment that leaves these conceptual nuances untouched.

4.4. Conclusion

In this chapter I introduced the idea that a more differentiated picture of moral judgment does not only allow to avoid inconsistencies in interdisciplinary concept use, but also to increase our understanding of moral judgment. After introducing the concept of scientific fruitfulness as a standard of evaluation of scientific concepts, I suggested two consequences of a more refined conceptualization of moral judgment: introduction of psychophysics of morality on the one hand and a more fine grained style of argumentation in empirical philosophy on the other. The latter makes use of conceptual subtleties in order to find spots where to test for the empirical adequateness of philosophical approaches. The former bridges the conceptual gap between different measurements of the same quality. Both allow to assess existing data in a more coherent way than in the "anything goes" account of morality. Both increase our capability to formulate laws and predictions. Both should therefore prove useful in bringing forward psychology of moral judgment and empirical philosophy as a method and I can only hope that interdisciplinary research on the matter will develop in this direction. A more refined conceptualization of moral judgment can therefore be regarded as more scientifically fruitful than an "anything goes" approach.

5. Advancing empirical research on moral judgment

This concludes my thesis on what *moral judgment* is - or rather, about the plurality of meanings of *moral judgment* and its consequences in the context of empirical and interdisciplinary research. I have come a long way: In chapter 1, I laid out the philosophical groundwork for this investigation.

It came out that in the light of certain understanding of philosophy of language the combination of the vagueness of the prescientific concept moral judgment and science's need for conceptual precision leads to a possibly troublesome consequence. That consequence consists in the prospect of different empirical approaches assuming different meanings for *moral judgment*, not in the light of explicit considerations but mainly through implicit concept use. Due to their implicit nature these conceptual divergences could be assumed to be easily overlooked in interdisciplinary practice as well as within a single discipline, resulting in counterproductive misunderstandings. From these thoughts I derived three hypotheses about usage of the term *moral judgment* in interdisciplinary and empirical research: (1) notions of *moral judgment* with differing implicit meanings are applied in different research projects and approaches, and these implicit differences have negative effects on (2) interdisciplinary philosophical arguments and (3) empirical research on moral judgment itself.

In chapter 2, I introduced the reader to the exemplary theory of morality that is supposed to be my "model theory" (analogous to a *model organism* in biology).

Haidt's Social Intuitive Model of moral judgment was introduced and its key concepts analyzed. In the same way, his Moral Foundation Theory was presented along with its key concepts and its claims. As a last step of presenting Haidt's theory of moral judgment, I established the point that the meanings of *intuition* in the SIM and *emotion* in the MFT are so closely related that both models can be understood as one integrated theory about moral judgment. Finally, I introduced an exemplary study that is taken to support the SIM and the MFT and demonstrated how the specific predictions for the study's results can be deduced from the general claims of the MFT and the SIM.

Chapter 3 showed how easily divergences in understanding the term *moral judgment* can be overlooked, that they are in fact overlooked and that this sabotages philosophical as well as scientific practice.

In the first part of chapter 3 I elaborated how the concept *intuition*, even though philosophers and psychologists define it in a similar way, serves different purposes in philosophy and in psychology. I showed how this turns an exemplary philosophical argument based on empirical results inconclusive, validating hypotheses 1and 2.

In the next section, Haidt was shown to be committed to a perceptivist or slightly cognitivist take on *moral emotion*. The validating study was established to be unable to validate the SIM under the assumption of a strongly cognitive concept of emotion. I demonstrated how different philosophical understandings of *emotion* lead to different conclusions about the philosophical interpretations of the exemplary study's findings. This provided further validation for hypotheses 1 and 2.

Furthermore, it was shown that the SIM and the MFT can be regarded as sharing a binary take on the syntactic structure of moral judgment. It was also highlighted that unawareness of this point leads to the erroneous conclusion of Haidt's theory being validated by a further exemplary study employing a graded understanding of syntax of moral judgment. It was also demonstrated that philosophical approaches are prone to the same mistake. Along with first validation for hypothesis 3, this section provided further support for hypotheses 1 and 2.

In the next step, it was demonstrated that even though Haidt argues for a virtue ethicist account of morality his approach depends on a situationalist view. I highlighted that consequently applying a virtue ethicist view would undermine the empirical support for his theory and (in connection with specific empirical results) the main claims of the MFT. Accordingly, this section provided further support for all three hypotheses.

It came out that intention sensitivity as a key feature of moral judgment has to be implanted into Haidt's theory to make it compatible with folk and traditional academic concepts of moral judgment. Two ways to do so were demonstrated to involve differing philosophical consequences. One of the ways was found to be in conflict with results obtained with specific empirical results. Hypotheses 1 and 2 were further supported by these findings.

Lastly, it was shown that in spite of its embrace of culture relative morality, Haidt's theory is interpretable in both relativist and in objectivist terms. I demonstrated how specific objectivist and relativist theories of moral truth can be made more plausible through combination with Haidt's theory.

In chapter 4, I demonstrated how taking differences of meanings of scrutinized theoretical concepts of moral judgment seriously can help to explain the phenomenon of moral judgment (understood as broadly as in common language terms) much better. Two ways of achieving better results in research on moral judgment through adopting a pluralistic understanding of *moral judgment* were presented:

I introduced an argument that derived implications for a given philosophical account's understanding of *moral judgment* from its conceptual stances on *emotion* and *intuition*. These implications were shown to allow for derivation of experimental predictions that contradict given empirical results.

Furthermore, a way of empirically relating diverging concepts of moral judgment to each other was presented. Using my own study as exemplification, it could be demonstrated that results from this kind of approach do allow to translate findings pertaining to one understanding of moral judgments to theories employing a different concept of moral judgment. It was also exemplified how with comparably easy methods, new types of empirical predictions concerning moral judgment can be tested.

As I mentioned, I have come a long way since I have started this thesis with an anecdote from the book "Hitchhiker's Guide to the Galaxy". This episode was intended to show what can go wrong if people use the same term to refer to different entities and they do not become aware of it in time. In the same book, a famous philosopher is mentioned who has published three bestselling books: "Where God went wrong", "Some more of Gods greatest mistakes" and "Who is this God person anyway?". If I replaced *God* with *Moral Psychology* and *Empirical Moral Philosophy*, they could as well be the title for the first three, rather destructive chapters of this thesis. But especially with the fourth chapter I hope to have offered more than that. I hope to have offered ways how interdisciplinary arguments and interdisciplinary research can in fact *work* and that interdisciplinary research about moral judgment still bears huge potential for development of new methods and spectacular findings.

I did however not make any metaethical points. Nor did I make any ethical points. I might have made some psychological points, though – but only on the side. I just opened a big box of trouble and puzzles to solve alongside some suggestions how to approach these puzzles. I hope to have demonstrated that for naturalism in ethics to make sense, science must become more philosophical and empirical philosophy must embrace the scrutiny that it expects from science in its own arguments. I guess acknowledging the challenge is the first step to the solution.

6. Appendices

6.1. Appendix 1 – Results of the exemplary study

All of the following results are to be found in Haidt et al. (1993) and were gathered with the methods laid out in chapter 2.4.1.

General effects of city and SES

In the case of the *Interference* question, an effect of SES ($F(1,174)=55.2$, $p<.001$) but not of city was revealed, filtering those answers that did not perceive the harmless-bothering acts in the desired way did not affect the significance of this finding (figure A1) .

In the case of the *Universal* question, an effect of SES but not of city was revealed in the unfiltered data ($F(1, 174) = 55.24$, $p < .001$). After filtering the data the effect of SES remained stable and a significant effect of city was revealed between Philadelphia and the Brazilian cities ($F(2, 127) = 4.58$, $p<.05$) (figure A2).

Table 1
Percentage of Adults Who Said the Actor Should Be Stopped or Punished

	Recife		Porto Alegre		Philadelphia		
Story	Low SES	High SES	Low SES	High SES	Low SES	High SES	Total
Moral							
Swings	77	80	87	93	100	100	89
Convention							
Uniform	77	36	60	40	83	62	60
Hands	30	30	47	33	53	13	34
Harmless-offensive							
Flag	63	23	53	17	50	0	34
Promise	57	7	23	7	20	3	20
Dog	57	40	50	33	80	10	45
Kissing	68	53	70	50	87	57	64
Chicken	79	50	87	63	80	27	64
M harmless-offensive	63	35	57	34	63	19	45
M when filtered[a]	65	47	60	41	78	16	51

Note. SES = socioeconomic status.
[a] Cases were removed when not explicitly declared to be harmless and offensive.

Figure A1: Positive answers to *Interference* question in percent (from Haidt et al. 1993)

Table 2
Percentage of Adults Who Universalized Their Judgment

Story	Recife		Porto Alegre		Philadelphia		Total
	Low SES	High SES	Low SES	High SES	Low SES	High SES	
Moral							
Swings	83	50	87	60	87	67	72
Convention							
Uniform	40	14	20	13	10	7	17
Hands	37	7	50	3	23	3	21
Harmless-offensive							
Flag	50	24	67	13	50	3	35
Promise	87	28	53	23	40	20	42
Dog	60	13	60	17	57	7	36
Kissing	67	20	53	33	80	17	45
Chicken	87	43	87	57	87	23	64
M harmless-offensive	70	26	64	29	63	14	44
M when filtered[a]	76	33	79	28	60	10	47

Note. SES = socioeconomic status.
[a] Cases were removed when not explicitly declared to be harmless and offensive.

Figure A2: Positive answers to *Universal* question in percent (from Haidt et al. 1993)

The authors conclude:

> Taken together, the results of the Interference and Universal probes support the first four research predictions. The majority of high-SES Philadelphians took a permissive stance toward the harmless-offensive stories (Prediction 1). In Recife, the majority of low-SES subjects took a moralizing stance (Prediction 4). There was a large and consistent effect of social class (Prediction 3), in which high SES-groups were more permissive than low SES-groups. The Philadelphia college students were consistently the most permissive group on the Interference and Universal probes; however, the overall effect of city was significant only in the filtered analysis of the Universal probe, so Prediction 2 (main effect of westernization) received only weak support.
>
> (Haidt et al., 1993,p. 619)

Permissive vs. moralizing groups

For further analysis, the different participant groups were categorized according to their general tendency to judge harmless-offensive acts: When a subject considered interference necessary and universalized this judgment, the subject was taken to show a *fully moralized* response. When neither universalization nor interference was endorsed, the answers were coded as a fully permissive response. The two remaining options occurred much more seldom in the case of the harmless-offensive scenarios.

When this distinction is applied to the different groups of participants, a clear image emerges:

- On the one hand side, there are four groups (college students in all three cities and Philadelphia high SES-children) in which an absolute majority of the responses to harmless-offensive acts were fully permissive (P-h-A 72%, PA-h-A 58%, R-h-A 50%, P-h-C 55%).
- On the other hand, in all low-SES groups as well as the high-SES children group from Recife, the majority of answers was fully moralizing (45%-92%, Mean:61% - percentage of highly permissive answers was between 3% and 25%).
- Only the high-SES children in Porto Alegre showed a more or less even distribution between fully permissive (32%), fully moralized (31%), and personal-moral (interference, but no universalization - 27%).

Distinctions among story types

Further analysis was performed to see whether city or SES had an influence on the distinction between prototypical moral situations (swing) and prototypical conventional situations (uniform and one other story) or on the distinction between prototypical moral situations (swing) and the harmless-offensive scenarios. The maximum distinction was in both cases considered to be the case when the swing story was moralized but none of the contrasting stories were. The minimum distinction was the case when all stories were moralized. Between these two endpoints, a rating scale with 100 points (100 for maximum, 0 for minimum) was introduced. The percentage of the 100 point-scores for both distinctions can be found in the figure below.

The effect of city on the moral conventional distinction was significant in adults ($F(2,171)=3.85$, $p<.05$) as well as in children. There was no effect of SES.

In the case of the moral-harmless distinction, there is a significant effect of city ($F(2,171)=3.31$, $p<.05$) as well as of SES ($F(1,171)=6.51$, $p<.01$).

The authors conclude:

> In sum, this analysis confirms the conclusions of the previous sections and supports all five research predictions. Both of the Philadelphia high-SES groups made large distinctions between the harmful story (Swings) and the harmless-offensive stories (Prediction 1); the low-SES Recife subjects made small or nonsignificant distinctions (Prediction 4); and the moral-harmless distinction was affected by city (Prediction 2) and SES (Prediction 3) in the predicted ways.).

Table 5
Percentage of Maximum Possible Distinction Between Story Types on the Universal Probe

	Recife		Porto Alegre		Philadelphia		
Story	Low SES	High SES	Low SES	High SES	Low SES	High SES	Total
			Moral–conventional distinction				
Adults	45	40	52	52	70	61	53
Children	17	25	40	57	72	57	44
			Moral–harmless distinction				
Adults	13	24	23	31	24	51	27
Children	8	12	11	33	22	38	20

Note. SES = socioeconomic status.

Figure A3: Distinctions among story types (from Haidt et al. 1993)

Justifications

The codings of the answers to the justification question can be seen in figure A4 below.

The following effects could be observed among adults:

- High-SES groups were more likely to use autonomy, the precursor of Haidt's harm domain , $F(1,174)=58.04$, $p<.001$, less likely to use community, the precursor of the authority domain, $F(1,174 = 20.37$, $p<.001$, and less likely to use norm statements, $F(1,174)=34.50$, $p<.001$.
- The effect of city was effective only for autonomy, $F(2,174)=3.87$, $p<.05$ which was used more often in Philadelphia than in Recife ($p<.05$) The interaction of SES and city was signifi-

cant for autonomy, F(2,174)=5.72, p<.01; divinity, F(2,174)=3.36, p<.05; and norm statement, F(2, 174)=3.48, p<.05.

Table 6
Percentage of Justifications Referring to Each of Shweder's Moral Codes

	Recife		Porto Alegre		Philadelphia		
Code	Low SES	High SES	Low SES	High SES	Low SES	High SES	Total
			Adults				
Ethics of autonomy	15	37	23	39	16	59	31
Ethics of community	29	21	32	22	31	11	24
Ethics of divinity	7	12	14	14	16	9	12
Norm statement	21	4	11	6	17	4	11
Uncodable	27	27	20	19	19	17	22
			Children				
Ethics of autonomy	15	13	12	26	26	27	20
Ethics of community	30	37	45	43	25	27	34
Ethics of divinity	7	5	4	2	6	7	5
Norm statement	23	22	13	7	15	7	14
Uncodable	25	23	27	23	28	32	26

Note. Adult data reflect responses to five harmless-offensive stories, including the Chicken story. Child data reflect responses to four stories.

Figure A4: Codings of justifications - "Ethics of autonomy" is the precursor of the harm domain, "Ethics of Divinity" is the precursor of the purity domain, "Ethics of community" is the precursor of the authority domain. (from Haidt et al. 1993)

Predicitve power of harm vs. affect

In each of the four permissive groups, the concordance of *Harm* with *Universal* (mean concordance = 66%) was higher than the concordance of *Bother* with *Universal* (mean concordance = 55%). In each of the seven moralizing groups, the concordance of *Bother* with *Universal* (mean concordance = 70%) was higher than the concordance of *Harm* with *Universal* (mean concordance = 56%).

6.2. Appendix 2 – Stimuli of psychophysical study

6.2.1. Dilemma Stimuli

1. Tax Fraud – Adaptation ill spouse (Adapted from Greene, 2008)

You are the owner of a small business trying to make ends meet. Your spouse is sick and needs an expensive drug treatment in order to save her. However, this treatment is not at all sure to work, but you are desperate to try anything. It occurs to you that you could get more money back in your tax return by pretending that some of your personal expenses are business expenses.
For example, you could pretend that the stereo in your bedroom is being used in the lounge at the office, or that your dinners out with your wife and friends are dinners with clients. You are well aware that this would be tax fraud and the money you are taking from the community would be desperately needed for communal expenses like schools or municipal hospitals.

You are faced with a choice between two options:

Option A: Even though it means committing tax fraud - you decide to scramble together the money for your spouse's treatment.

Option B: Even though you risk not being able to scramble together the money for your spouse's treatment - you decide to refrain from committing tax fraud.

2. Transplant –adaptation priority list law (Adapted from Greene, 2008)

You are a doctor. You have two patients who are both critically ill waiting for the same organ transplant operation. Both patients will die if nothing is done immediately.
Patient A has been waiting for many years and has suffered a long time due to his illness. You feel he deserved the organ although his illness has weakened him so much that his chance of survival after surgery are very low.
Patient B's chance of surviving the operation is much higher, since he has been waiting for only a year and endured far less suffering from his defective organ.

You are faced with a choice between two options:

Option A: Even though Patient B has a much higher chance of surviving surgery - you decide to give the organ to Person A because he has waited much longer for the organ.

Option B: Even though Person A has waited much longer - you decide to give the organ to Patient B because he has a much higher chance of surviving surgery.

3. Footbridge—adaption of question (Adapted from Greene, 2008)

A runaway trolley is heading down the tracks toward five workmen who will be killed if the trolley proceeds on its present course. You are on a footbridge over the tracks, in between the approaching trolley and the five workmen. Next to you on this footbridge is a stranger who happens to be very large.
The only way to save the lives of the five workmen is to push this stranger off the bridge and onto the tracks below where his large body will stop the trolley. The stranger will die if you do this, but the five workmen will be saved.

You are faced with a choice between two options:

Option A: Even though it means pushing the stranger in front of the trolley - you decide to save the five workmen.

Option B: Even though it means letting the five workmen die - you decide to leave the stranger untouched.

4. Modified Lifeboat – (Adapted from Greene, 2008)

You are on a cruise ship when there is a fire on board, and the ship has to be abandoned. The lifeboats are carrying many more people than they were designed to carry. The lifeboat you're in is sitting dangerously low in the water—a few inches lower and it will sink.

The seas start to get rough, and the boat begins to fill with water. If nothing is done it will sink before the rescue boats arrive and everyone on board will die. However, there is one injured person who is unlikely to survive in any case. If you throw that person overboard the boat will stay afloat and the remaining 30 passengers will be saved.

You are faced with a choice between two options:

Option A: Even though it means pushing the injured person overboard - you decide to save the remaining passengers

Option B: Even though it means that all passengers will drown - Youdecide not to push the injured person overboard

5. Lost Wallet—modified for Charity goal short of 50 (Adapted from Greene, 2008)

You are a good law abiding citizen who is raising money for a charity. The deadline for meeting your target sum is at midnight and you are just short of 50 euros. If you meet your goal, the money collected would be doubled by being matched by sponsor.
You are walking down the street when you come across a wallet lying on the ground. You open the wallet and find that it contains several hundred dollars in cash as well the owner's driver's license. From the credit cards and other items in the wallet it's very clear that the wallet's owner is wealthy. You, on the other hand, just need 50 euros for your charity goal. You consider sending the wallet back to the owner without the cash, keeping the cash for yourself.
You are faced with a choice between two options:

Option A: Even though it means stealing the money - you keep the money you found in the wallet in order to double the money for your charity

Option B: Even though it means being able to double the money for your charity - you refrain from stealing the money

6. Crying Baby (Adapted from Greene, 2008)

Enemy soldiers have taken over your village. They have orders to kill all remaining civilians. You and some of your townspeople have sought refuge in the cellar of a large house. Outside you hear the voices of soldiers who have come to search the house for valuables.
Your baby begins to cry loudly. You cover his mouth to block the sound. If you remove your hand from his mouth his crying will summon the attention of the soldiers who will kill you, your child, and the others hiding out in the cellar. To save yourself and the others you must smother your child to death.

You are faced with a choice between two options:

Option A: Even though it means killing the baby - you decide to save yourself and the other villagers.

Option B: Even though it means letting all villagers die - you decide to leave the baby untouched.

7. Sophie's Choice (Adapted from Greene, 2008)

It is wartime and you and your two children, ages eight and five, are living in a territory that has been occupied by the enemy. At the enemy's headquarters is a doctor who performs painful experiments on humans that inevitably lead to death.
He intends to perform experiments on one of your children, but he will allow you to choose which of your children will be experimented upon. You have twenty-four hours to bring one of your children

to his laboratory. If you refuse to bring one of your children to his laboratory he will find them both and experiment on both of them.

You are faced with a choice between two options:

Option A: Even though it means personally bringing one of them to the laboratory - you decide to save one of your children.

Option B: Even though it means that both of them will be brought to the laboratory – you decide to leave your children untouched.

8. Airplane

You are the person in charge as minister of defense. You have just been told that terrorists have hi-jacked a commercial aircraft with 350 passengers. After a distress call from the pilots, the line goes dead. Now the plane it is flying slowly at a low altitude towards a well populated area with several major cities. You are informed by various experts that there is strong evidence that the hijackers plan to crash the plane on the nearby populated area. Military aircraft is up in the air waiting on your decision on whether or not to shoot the plane.

You are faced with a choice between two options:

Option A: Even though it means shooting down the airplane and killing all passengers – you decide to avoid a likely catastrophe

Option B: Even though it means risking a very likely catastrophe – you decide not to shoot down the plane and leave the passengers' lives untouched

9. Trial phase 2

You are the lead of a team of researchers who have reached very promising results on a possible cure for a terrible children disease which so far always ends fatal. So far you have conducted your research on animals, and the next phase of your trial would include administering the treatment on patients. There is a promising chance that it might work and a terrible disease could be stopped, however because it the first trial phase on humans this could come at the expense of several of your infant trial patients having lifelong painful side effects. On the other hand, these patients who already suffer from the disease will die anyway if you do nothing.

You are faced with a choice between two options:

Option A: Even though it means risking several infant trial patients suffering from life long side effects – you decide to test the treatment in order to stop the terrible disease

Option B: Even though it means letting go all hope for stopping the disease and save their lives - you decide to spare the infant trial patients the risk from suffering lifelong painful side effects

10. Hostage taker

You are on command over the team handling a hostage situation in which a hostage taker has a group of 30 innocent lives under his threat. He has threatened to kill them because his demands cannot be met. One of the snipers has clear shot at the hostage taker but this would entail killing one of the hostages that is in the way.

You are faced with a choice between two options:

Option A: Even though it means letting one hostage die - you decide to rescue the remaining hostages by having the hostage taker killed

Option B: Even though it means taking the risk of all hostages dying - you decide not to release the firing order and leave the hostage untouched

11. Standard Fumes (Adapted from Greene, 2008)

You are the late-night watchman in a hospital. Due to an accident in the building next door, there are deadly fumes rising up through the hospital's ventilation system. In a certain room of the hospital are three patients. In another room there is a single patient. If you do nothing the fumes will rise up into the room containing the three patients and cause their deaths.
The only way to avoid the deaths of these patients is to hit a certain switch, which will cause the fumes to bypass the room containing the three patients. As a result of doing this the fumes will enter the room containing the single patient, causing his death.

You are faced with a choice between two options:

Option A: Even though it means killing the one patient in the other room – you decide to hit the switch and avoid the deaths of the three patients

Option B: Even though it means killing the three patients - you decide not to hit the switch and leave the one patient in the other room untouched.

6.2.2. Nondilemmatic/"normal" Stimuli

12. Five-for-Seven Trolley (Adapted from Greene, 2008)

You are at the wheel of a runaway trolley quickly approaching a fork in the tracks. On the tracks extending to the left is a group of five railway workmen. On the tracks extending to the right is a group of seven railway workmen.
If you do nothing the trolley will proceed to the left, causing the deaths of the five workmen. The only way to save these workmen is to hit a switch on your dashboard that will cause the trolley to proceed to the right, causing the deaths of the seven workmen on the other side.

You are faced with a choice between two options:

Option A: Even though it means letting the seven workmen die - you decide to hit the switch and avoid the deaths of the five workmen

Option B: Even though it means letting the five workmen die - you decide to do nothing, leaving the health of the seven workmen untouched

13. Guarded Speedboat (Adapted from Greene, 2008)

While on vacation on a remote island, you are fishing from a seaside dock. You observe a group of tourists board a small boat and set sail for a nearby island. Soon after their departure you hear over the radio that there is a violent storm brewing, a storm that is sure to intercept them.
The only way that you can ensure their safety is to warn them by borrowing a nearby speedboat. The speedboat belongs to a miserly tycoon who has hired a fiercely loyal guard to make sure that no one uses his boat without permission. To get to the speedboat you will have to lie to the guard.

You are faced with a choice between two options:

Option A: Even though it means lying to the guard – you decide to convince him to let you borrow the speedboat and warn the tourists about the storm

Option B: Even though it means risking not warning the tourists about the storm - you decide to remain honest to the guard

14. Stock Tip (Adapted from Greene, 2008)

You are a management consultant working on a case for a large corporate client. You have access to confidential information that would be very useful to investors. You have a friend who plays the stock market. You owe this friend a sizable sum of money.
By providing her with certain confidential information you could help her make a lot of money, considerably more than you owe her. If you did this, she would insist on canceling your debt. Releasing information in this way is strictly forbidden by national and international law.

You are faced with a choice between two options:

Option A: Even though it means taking part in massive market manipulations through insider trading - you decide to release this information to your friend so that she will cancel your debt.

Option B: Even though it means continuing to owe a lot of money to your friend - you decide to refrain from taking part in market manipulation by keeping the information to yourself.

15. Subway (adapted from Sommer et al., 2010)

You want to board the subway train that is about to leave and that runs only every 15 min. Next to you on the platform is an old man with a bag of groceries. As you are getting on the train, he accidentally drops the bag on the floor.

You are faced with a choice between two options:

Option A: Even though it means you will miss the train – you decide to help the man

Option B: Even though it means not helping the man – you decide to take the train

16. Clothes (adapted from Sommer et al., 2010)

At a department store you discover your dream clothes. On the way to the cash register you remember a report on child labor which you have recently seen on TV. The brand name of the clothes you want to buy was mentioned there, too.

You are faced with a choice between two options:

Option A: Even though this means financing child labor - you decide to buy the clothes

Option B: Even though this means not buying the clothes - you decide to refrain from financing child labor

17. Club (adapted from Sommer et al., 2010)

You are at a club that is really packed tonight. After you finally get your drink you realize that the barkeeper has given you 10D too much in return. In order to give the money back you would have to make your way back through the crowd to the counter.

You are faced with a choice between two options:

Option A: Even though it means having to get back to the counter - you give back the 10D

Option B: Even though it means stealing the 10D - you decide not to go back to the counter

18. Appartment (adapted from Sommer et al., 2010)

You could move into an apartment that you really like. However, the landlord does not permit pets but you own a cat. The landlord lives 100km from the apartment and would probably never find out that you have a cat.

You are faced with a choice between two options:

Option A: Even though it means lying to the landlord - you decide not to tell about the cat and get the apartment

Option B: Even though it means not getting the apartment – you decide to be honest to the landlord

19. Radiator (adapted from Sommer et al., 2010)

You want to sell your old car. You know that the car's radiator would need to be exchanged pretty soon. A man who does not notice the problem wants to buy the car right away and offers to pay in cash.

You are faced with a choice between two options:

Option A: Even though it means risking receiving a lower price for the car - you decide to remain honest and mention the radiator.

Option B: Even though it means being dishonest – You decide to keep quiet about the radiator and make the sale.

20. Crutches (adapted from Sommer et al., 2010)

You are very tired and you are sitting on the bus home. At the next stop a woman with crutches gets on the bus. All seats in the bus are taken.

You are faced with a choice between two options:

Option A: Even though it means risking having to stand for the rest of the trip - You decide to offer my seat to the woman.

Option B: Even though it means risking that the woman on crutches will not find a seat - You decide to refrain from offering my seat.

21. Donation (adapted from Greene, 2008)

You are at home one day when the mail arrives. You receive a letter from a reputable international aid organization. The letter asks you to make a donation of two hundred dollars to their organization.
The letter explains that a two hundred-dollar donation will allow this organization to provide needed medical attention to some poor people in another part of the world.

You are faced with a choice between two options:

Option A: Even though it means losing the money – you decide to make the donation

Option B: Even though it means not helping the poor people – you decide to save the money.

6.3. Appendix 3: Results of psychophysical study – further psychophysical functions

A total of 208 (104 female) subjects participated in the online questionnaire through Qualtrics, a third-party online survey administration company. Subjects were compensated with "survey cash," credits that could be converted into monetary compensation after individuals participated in a certain number of research studies, including our own. Reliability measures (reliability questions and minimum time limit) were taken to ensure accurate subject participation. Subjects began by reading a general description of the test and were asked to acknowledge the test's nature and context. Specifically, they were asked to keep in mind that when asked whether a situation is morally permissible, to consider what is morally acceptable, not what they would actually do in that situation. Subjects were asked to complete the test without interruption, read through all hypothetical scenarios and answer questions about each one.

The composition of the questionnaire is explained in section 4.3.3 of this thesis, the stimuli are featured in Appendix 2.In addition to the results discussed in 4.3.3, the following observations were made:

Function "graded ratings above 50 vs. positive binaries":

The relation "percentage of graded judgments above 50 vs. percentage of positive binary judgments" can be described via a linear regression of 1.026*x – 4.47 with a correlation coefficient of .867.

In regard to differences between the same function for dilemma cases and normal cases taken on their own, the same observations as for the "mean graded vs positive binaries" relation can be made: while the slope of the linear regression for the "normal" condition stays more or less the same ((y = 1.19 *x + 14.6) , the slope of the "dilemma" condition becomes less step (y = 0.77*x + 10).

Interestingly, the correlation between the two variables is slightly weaker than the one of the "mean graded judgment vs. positive binary judgments" in all conditions: the difference if all cases are considered is at about .04, if only dilemmas are considered at about .05 and if only normal cases are considered at about .02.

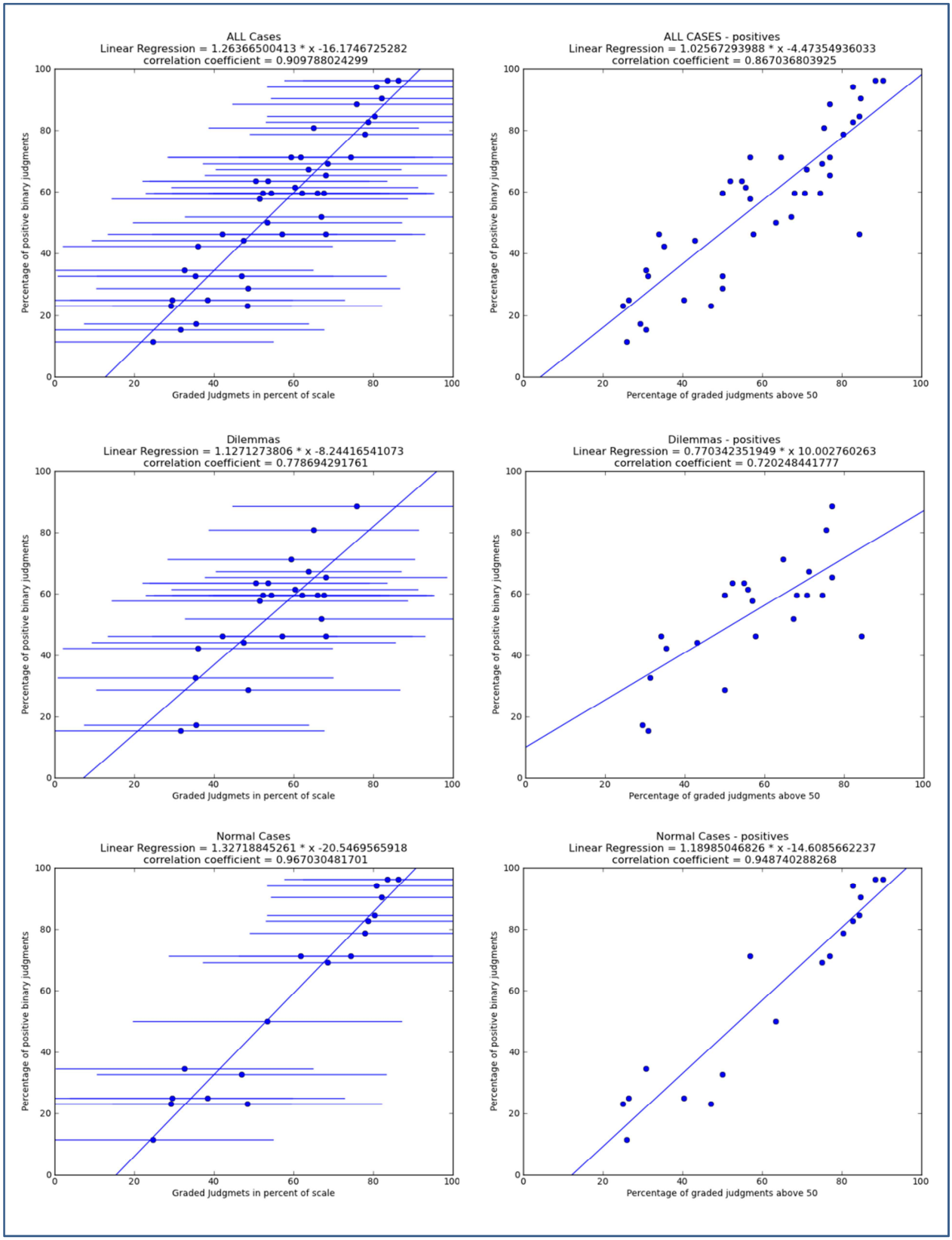


Figure A5: Psychometric functions "mean graded judgment vs. percentage of positive binary judgments" and "percentage of graded judgments above 50 vs. percentage of positive binary judgments" for all cases, dilemma cases and "normal" cases

Concerning answer patterns, a look at the distributions of answers in the single cases (all histograms can be found in Appendix 5) shows a clear picture: In normal cases, there is normally one clear peak at one end of the spectrum, namely for the option rated as more permissible. In dilemmas, we find a significantly reduced tendency for peaks, with clear peaks only in the *wallet* and the *footbridge* case. This tendency can be described via the relations "graded above 90 vs. binary positives" and "graded below 10 vs. binary positives". For the dilemma cases, the former had a linear regression of 0.315*x+3.67 with a correlation coefficient of .505, while the latter had a linear regression of -0.516*x-43.39 with a correlation coefficient of -.763. For the normal cases, these functions had much steeper slopes (.74*x -9.6 , -.485*x + 45.957)) and much higher correlation coefficients (.966 and -.908).

The values for all cases lie in between, as one might expect: 0.628* x -8.808 with a correlation coefficient of 0.8126 for the "graded above 90 vs. binary positives" relation and -0.485 *x + 43.54 with a correlation coefficient of - 0.835 for the "graded below 10 vs. binary positives" relation.

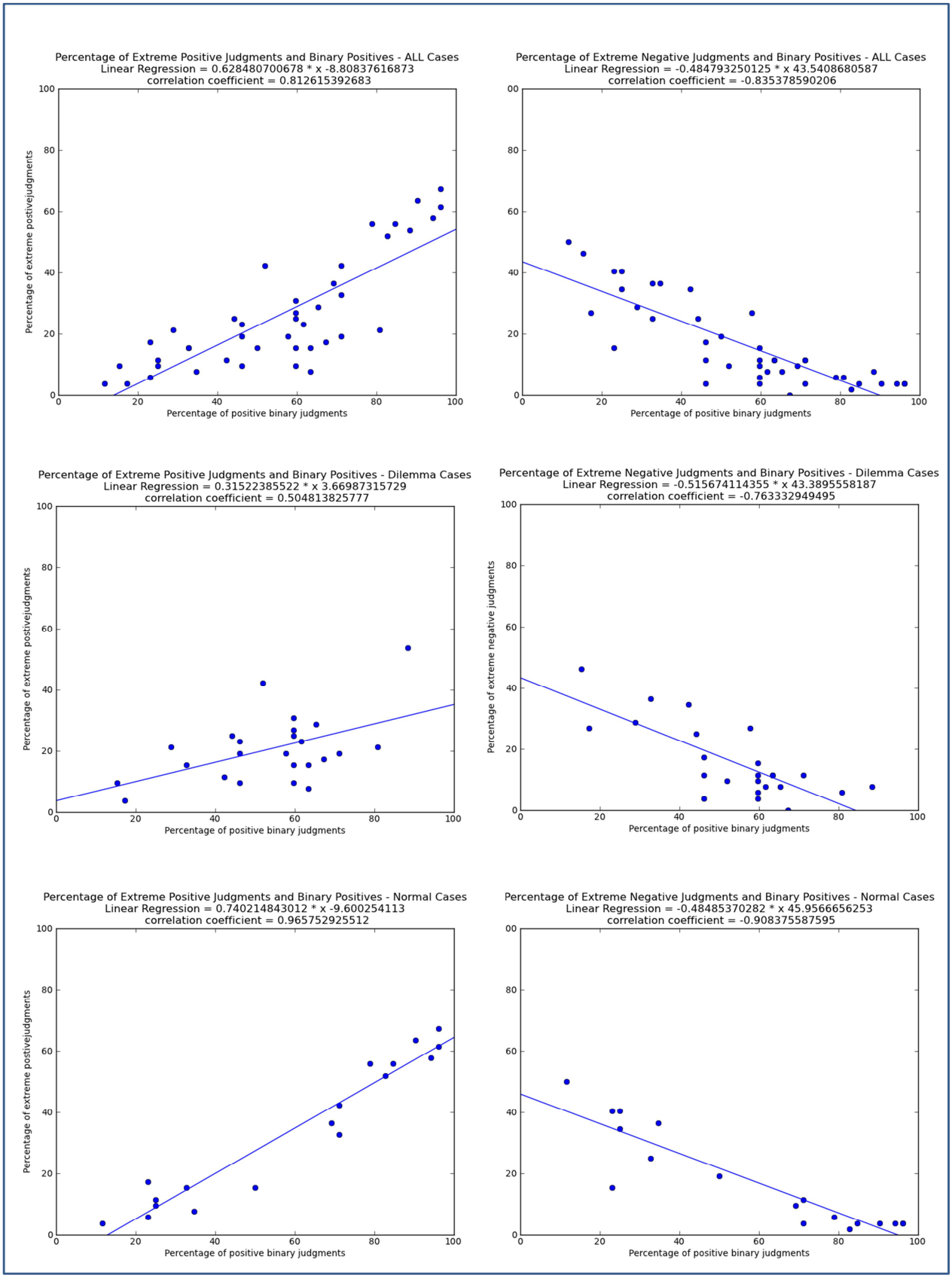


Figure A6: Psychometric functions "percentage of positive binary judgments vs. percentage of graded judgment above 90" and "percentage of positive binary judgments vs. percentage of graded judgment below 10" for all cases, dilemma cases, and "normal" cases

6.4. Appendix 4 – Simulations of computational ordering

6.4.1. Deriving Predictions via simulation of Kantian and Utilitarian Picture of moral judgment

In order to derive actual predictions concerning the primacy of either binary or utilitarian judgment, we prepared two simple simulations of both scenarios that I will call according to their respective source of inspiration the "Kantian" and the "Utilitarian" simulation:

In the "Utilitarian" simulation, 60 fictive scenarios with a random *actual* graded value and a number of fictive judgments grouped around these values in form of a normal distribution were generated. Values above 100 or below 0 were capped to 0 or 100. Additionally, binary judgments were generated that were derived from graded judgments with the additional step of turning any judgment over 50 into a confirmative judgment and any judgment below 50 into a negative judgment.

In the "Kantian" simulation, 60 fictive scenarios were attributed an *actual* probability of people regarding the action in the scenario as permissible. Furthermore, an *actual imperfect duty value* between 0 and 50 was derived for each scenario for generation of graded judgments. Participants' binary judgments were generated by combining the outcome-probability derived for the scenario with a random number generator. Graded judgments were generated by first creating a binary value and then adding or subtracting a value that was generated randomly within a normal probability distribution around the actual imperfect duty value. Values that crossed the 0 or 100 threshold or the 50-points barrier were capped and set to either 0, 100 or 50.

In order to assess the fit of the models ideally, we decided to take into account the actual results for deciding which standard deviation to apply for the probability distributions around the "actual" values. We therefore simulated predictions of both models for standard deviation values between .05 and .4 in order to assess a posteriori which sigma value to take for judging the fit of the model. These values were 0.3 for the utilitarian prediction and 0.2 for the Kantian prediction.

The models made predictions that can be exemplified by the following predictions (see the scripts in Python in figures A8 and A9):

6.4.2. Kantian Predictions:

Concerning the mean graded judgment vs. percentage of positive binary judgments, 10,000 runs of the Kantian simulation predicted a linearly distributed cloud of points with a mean correlation coefficient of 0.76 (sd = 0.013) around a linear regression of 1.23 (sd = 0.017)*x-12.75 (sd=0.68). Concerning the relationship between the percentage of graded values of over 50 and the percentage of positive judgments, we find a cloud of points with a mean correlation coefficient of 0.9 (s=0.03) and a

linear regression of a significantly shallower slope (0.99 (s=0.03)*x + 7.81 (s = 1.01). Furthermore the points tend to drift away from each other with rising y-values (see example graph in figure A7).

Looking at the distributions of judgments for single cases, we found that in general two peaks could be found in the distribution of graded judgments: At one of the two ends of the spectrum as well as in the middle (see example histograms in figure A7). Depending on the binary result of the case, the peaks come in different strengths. The tendency of the distributions to form peaks at the endpoints of the moral judgment scale can be described through the functions binary positives vs. graded judgments above 90 (lin. reg.: y = 0.28 (sd=0.05)*x + 0.7(s=1.62), correlation coefficient at .52(sd=.09)) and binary positives vs. graded judgment below 10 (lin. reg. is y = - 0.24 (sd=0.07)*x + 24.8 (sd=4.66), correlation coefficient at -0.49 (sd=.08))

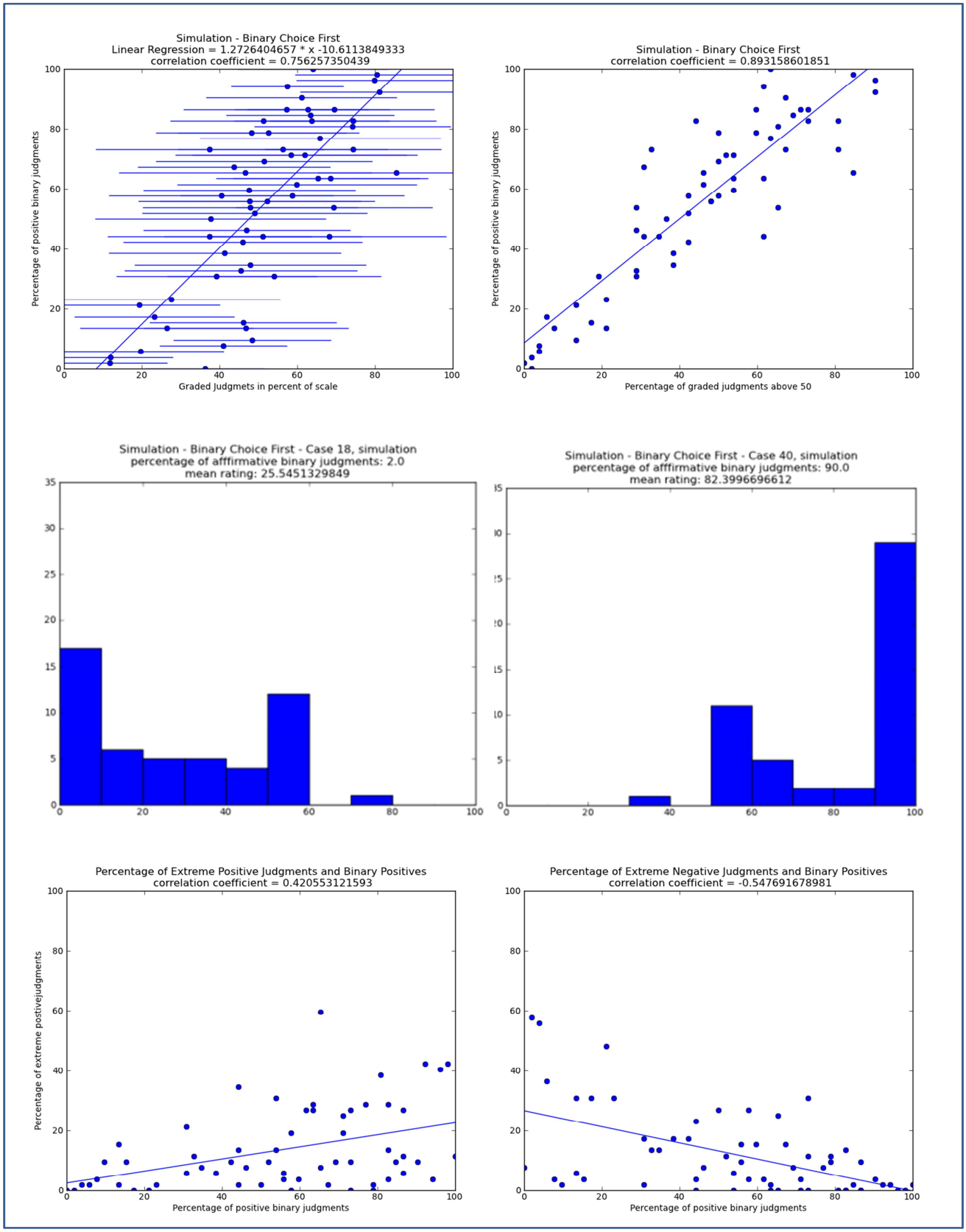


Figure A7: Exemplary simulated psychometric functions and distributions of judgments, Kantian simulation (binary choice first)

6.4.3. Utilitarian Predictions:

Concerning the form of the point cloud in the *positive binary vs. mean graded* judgments relation, we find that the Utilitarian simulation predicts a much higher correlation coefficient of .97(sd =.01) compared to the Kantian prediction. The linear regression takes the values of y = 1.25 (sd=0.02)*x – 12.17 (sd=1.68) and is in its mean values very close to the Kantian prediction, though with much smaller standard deviation.

Concerning the *binary positives vs. graded above 50* relation, the mean correlation coefficient of .96 (sd=.02) is substantially higher than in the Kantian prediction, though marginally lower than in the Utilitarian *mean graded vs. positive binary* relation. However, in 10,000 trials, it was higher than in the mean graded vs. positive binary relation with a probability of merely 12%. The slope is with y = 0.96 (sd= .0001) *x + 2.21 (sd=1.01) significantly shallower than in the Kantian prediction. Furthermore, the tendency of points to drift away from each other with rising y-values is absent.

Concerning the distributions of graded judgments, we find that there is normally only one peak involved which can be found at one of the ends of the scale. However, just like in the Kantian simulation, the salience of the peak depends very much on how close the mean graded judgment is to the ends of the scale. A look on the relations concerning the tendency to extreme judgment, we see that the slopes of the linear regressions are much steeper and the correlation coefficients much higher: *Graded above 90 vs. binary positives* is at y = .57 (sd=.01)*x – 11.3 (sd=.19) with a correlation coefficient of .88 (s=.02), *Graded below 10 vs. binary positives* is at y = -0.56 (sd=.06)*x + 45.5 (sd=.01).

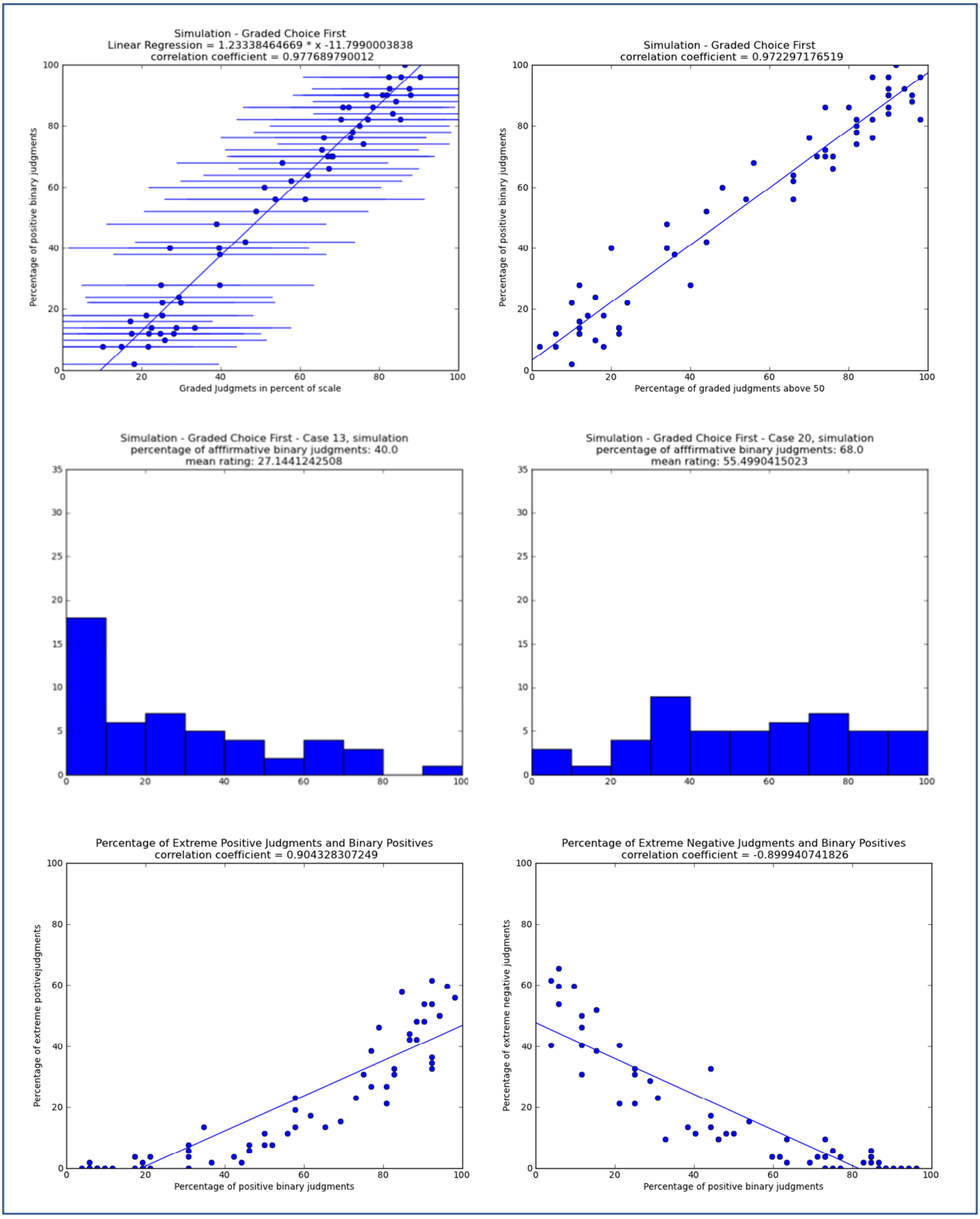


Figure A7: Exemplary simulated psychometric functions and distributions of judgments, Utilitarian simulation (graded choice first)

```
def simGradedFirst(NumberCases = 60, NumberTrials = 52, sigma = 0.3):
    import random
    import pylab

    ntrue = [] # Empty list of "true" values for simulated cases
    for i in range(NumberCases):
        ntrue.append(random.random())# List is filled with values for 60 cases
    Binary = []
    Graded = [] # Empty lists for binary and graded judgments
    for i in range(NumberCases):
        Binary.append([])
        Graded.append([]) # For each simulated case, a list is added to the judgment-lists

    for i in range(NumberCases):
        for j in range(NumberTrials):
            graded_judgment = ntrue[i] + random.normalvariate(0.0, sigma)  # For each case and for each simulated graded judgment,
                                                                           # a value is generated with a probability distribution
                                                                           # with mu = 0 and sigma = 0.3 around ntrue
            graded_judgment = min(max(ns,0),1) # the values are capped at the end of the judgment spectrum
            Choice [i].append((ntrue[i] + random.normalvariate(0.0, sigma)) > 0.5)  # Binary judgments are derived the same way
                                                                                    # plus a Boolean for 'x > 0.5?'
            Graded [i].append (graded_judgment)

    return Graded, Choice
```

Figure A8: Script for simulation of the Utilitarian model in Python

```
def simBinaryFirst(NumberCases = 60, NumberTrials = 52, sigma = 0.3):
    import random
    import pylab

    ntrue = [] # Empty list for binary answer distributions for simulated cases
    nvar = [] # Empty list for imperfect duties variables
    for i in range(NumberCases):
        ntrue.append(random.random()) # the lists are filled
        nvar.append(random.random()/2.0)

    Choice = []
    Graded = []# Empty lists for binary and graded judgments
    for i in range(NumberCases):
        Choice.append([])
        Graded.append([])# For each simulated case, a list is added to the judgment-lists

    for i in range(NumberCases):
        for j in range(NumberTrials):
            Choice[i].append(random.random()<= ntrue[i])# For each case and for each simulated graded judgment,
                                                       # a value is generated with a positive judgment proba-
                                                       # bility of p = ntrue
            graded_judgment = nvar[i] + random.normalvariate(0,sigma)   # For graded judgments an imperfect duty-value is
                                                                        # generated with a probability distribution
                                                                        # with mu = 0 and sigma = 0.3 around nvar
            graded_judgment = min(max(ns,0),0.5)                        # nvar is capped at 0 and 0.5
            if (random.random()<= ntrue[i]) == 1:
                graded_judgment = 0.5 + graded_judgment # A binary judgment is derived and the impefect duty value added
            Graded[i].append(graded_judgment)

    return Graded, Choice
```

A9: Script for simulation of the Utilitarian model in Python

6.4.4. Comparison of predictions and actual results

In order to compare the overall fits of the model, we determined four indicators of fit into the comparison: On the one hand the relation *mean graded vs. binary positives*, on the other the *binary vs. graded above 50*, relations *graded below 10 vs. binary* and *graded above 90 vs. binary*.

To assess the fit of both simulations, we optimized the standard deviation value of the judgment distributions per case to give each simulation the best possible fit to the four functions of the actual results. The procedure here was to first pick the 10 percent of standard deviation values that achieved the best predictions for the function in regard to all cases and then to pick from these the standard deviation value that offered the best results in the normal and dilemma fits combined. If the fits with two different standard deviation values were of similar closeness to the actual results for different functions,, the one that offered the better fit for the "mean graded vs. binary positives" function was picked. According to this procedure, the best fits were achieved for the Kantian simulation at a standard deviation value of 0.2, while the best fits for the Utilitarian model were achieved at 0.3.

The numerical assessment of the model fit was made in form of a Monte Carlo Simulation that counted the fits of both models' predictions in 1000 trials. If the y-value of the linear regression of the simulated data points was within 10 points above or below the actual value at x=0 and x= 100, and if the correlation coefficient of that relation was within a 0.1 tolerance below or above the actual one, the particular prediction of the model was counted as a fit.

The Monte Carlo Simulation revealed a dramatically better fit of the Utilitarian simulation in respect to all indicators of fit (see figure A10 and A11). The better fit of the Utilitarian model has different reasons for the different indicators. While the fit of the Kantian prediction to the *binary positives vs. graded above 50* and *binary positives vs. mean graded judgment* indicators was low because of the rather low correlation coefficient as well as the stronger variation of linear regressions between single predictions, the fit to the indicators that related to the distribution of graded judgments per case was low mainly because of a significant divergence of the linear regressions (see figure A10).

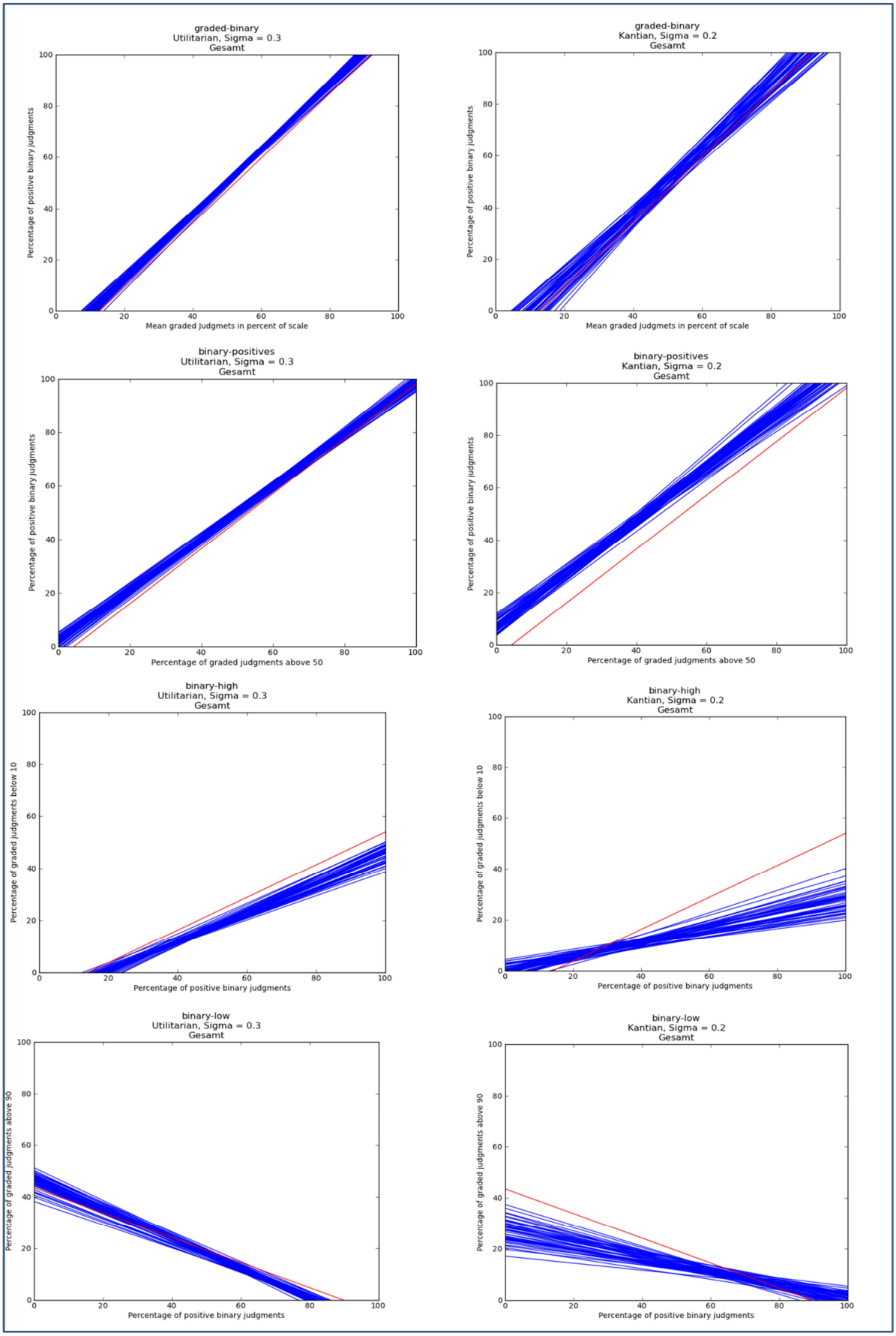


Figure A10: Actual psychometric functions (red) compared with 50 simulated functions from the Kantian (right) and Utilitarian (left) model.

These results were to be expected from the example predictions derived from the models, as on one side the histograms revealed significant differences in distribution of graded judgments per case while on the other significantly different predictions concerning the correlation coefficient of the psychometric functions were made.

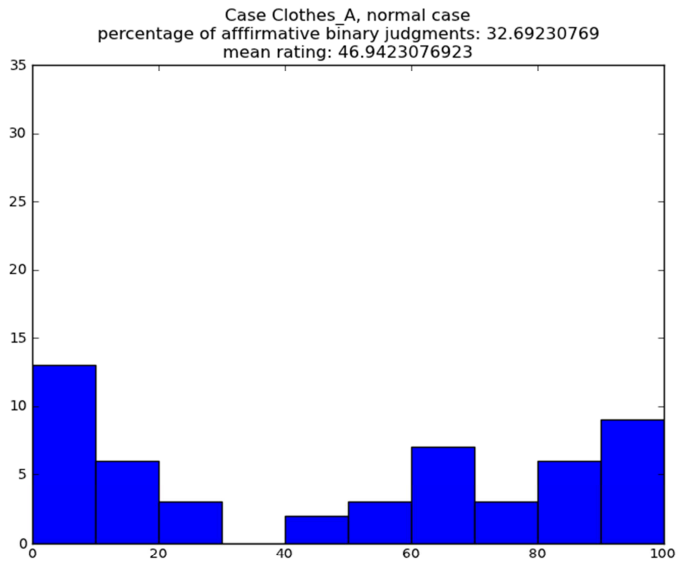
While in respect to the the normal cases the Utilitarian prediction continues to offer better predictions, even though only for the *binary positives vs. mean graded judgment* and the *graded judgments below 10 vs. positive binary* function, the dilemma-cases offer a radically different outlook. In regard to the dilemma situations, both models struggle to make reasonable predictions, with the Kantian model ahead in its fit to the *binary positives vs. mean graded judgment* and *graded judgments above 90 vs. positive binary* indicators and the Utilitarian model ahead in the *graded judgments below 10 vs. positive binary* indicator.

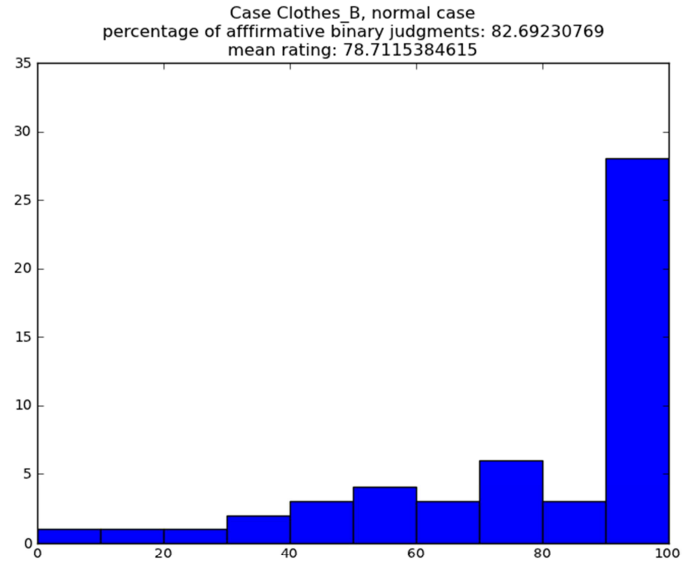
	Dilemma Cases	All Cases	Normal Cases
Kantian Hits - *Mean*	500	193	2
Kantian Hits - *Low*	0	40	0
Kantian Hits - *High*	21	2	0
Kantian Hits - *Positives*	483	0	0
Utilitarian Hits - *Mean*	0	999	896
Utilitarian Hits - *Low*	0	381	0
Utilitarian Hits - *High*	167	989	889
Utilitarian Hits - *Positives*	0	635	1

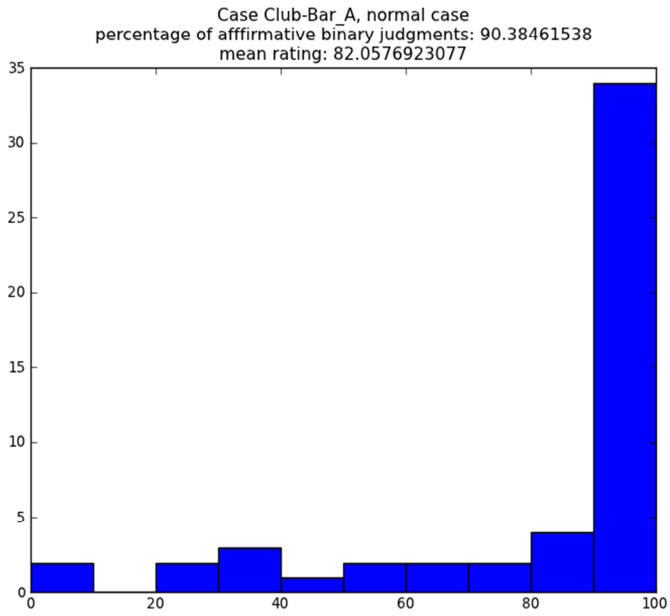
Figure A11: Correct predictions of psychometric functions in 1000 trials for different simulations

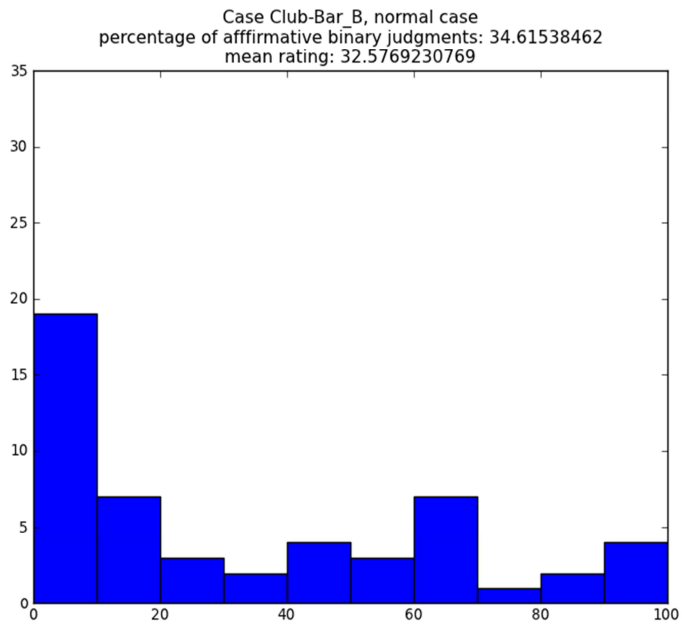
6.1. Appendix 5 – Results of own psychophysical study – Histograms

6.1.1. Histograms of normal cases

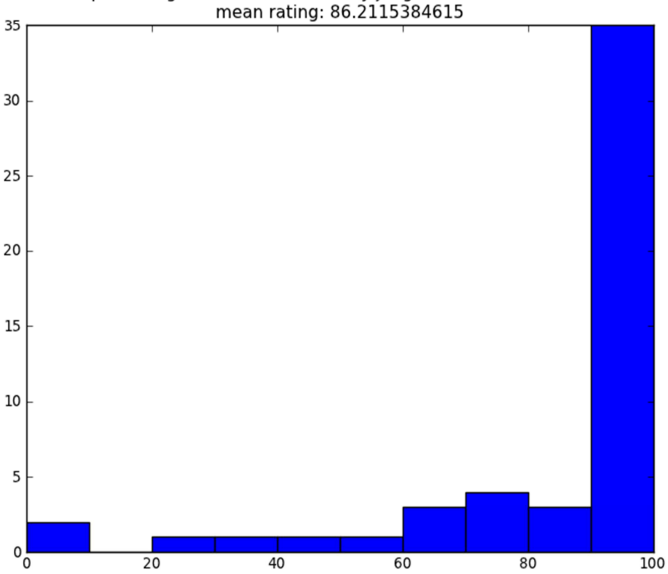




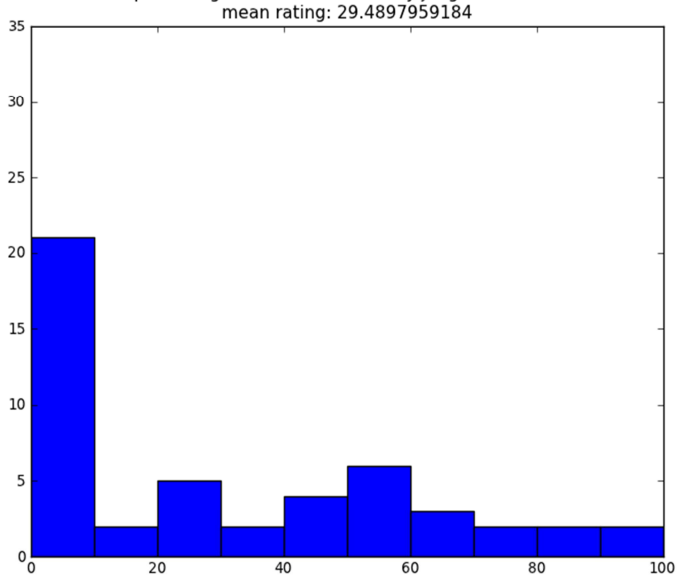




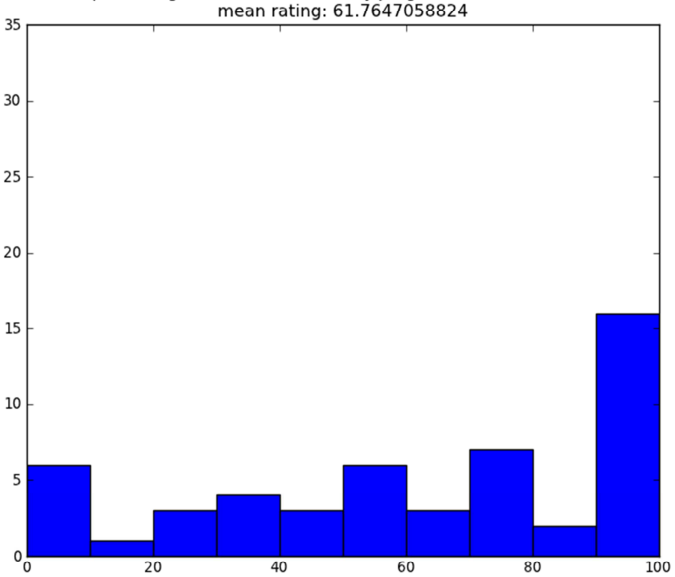
Case Crutches_A, normal case
percentage of afffirmative binary judgments: 96.15384615
mean rating: 86.2115384615



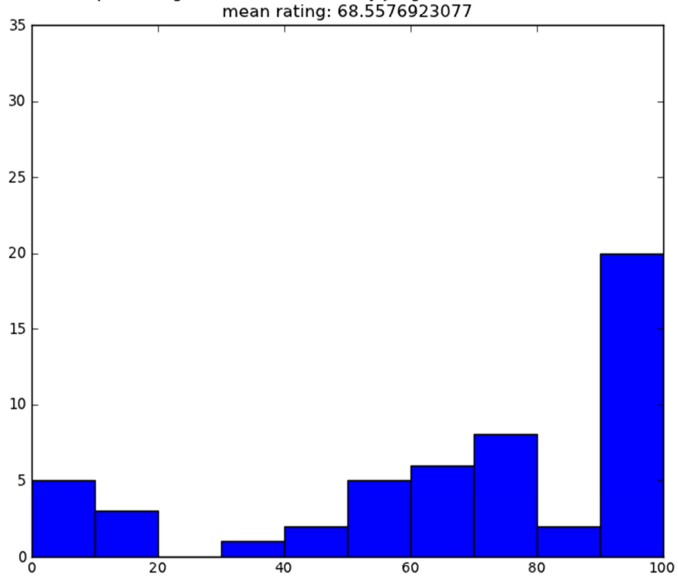
Case Crutches_B, normal case
percentage of afffirmative binary judgments: 25.0
mean rating: 29.4897959184



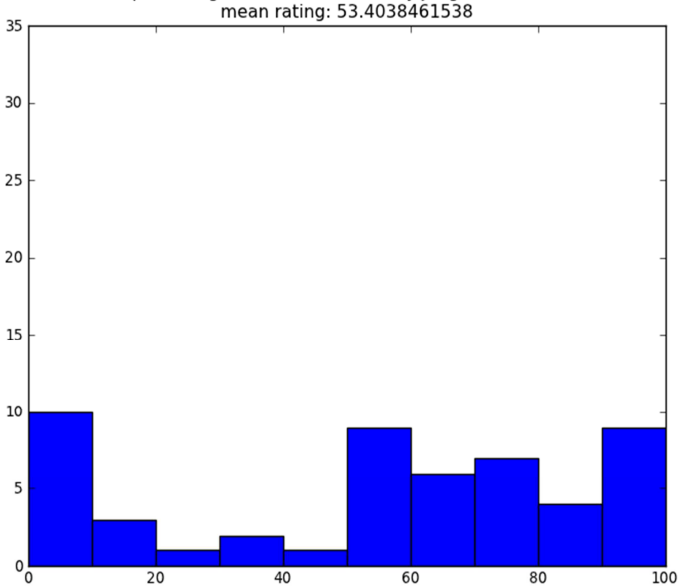
Case Donation_A, normal case
percentage of afffirmative binary judgments: 71.15384615
mean rating: 61.7647058824



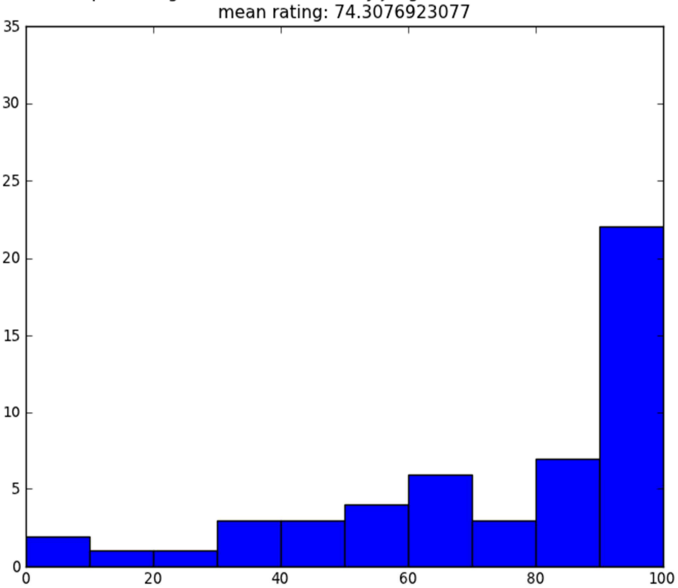
Case Donation_B, normal case
percentage of afffirmative binary judgments: 69.23076923
mean rating: 68.5576923077

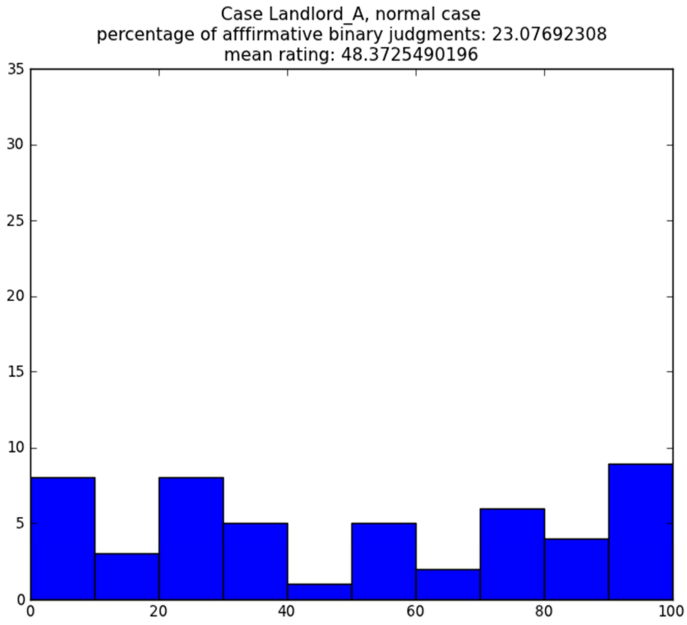


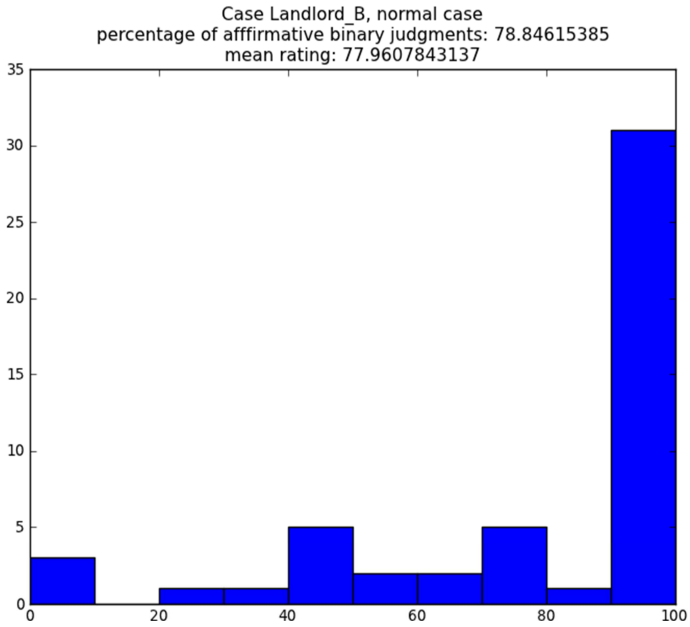
Case Guarded_Speedboat_B, normal case
percentage of afffirmative binary judgments: 50.0
mean rating: 53.4038461538

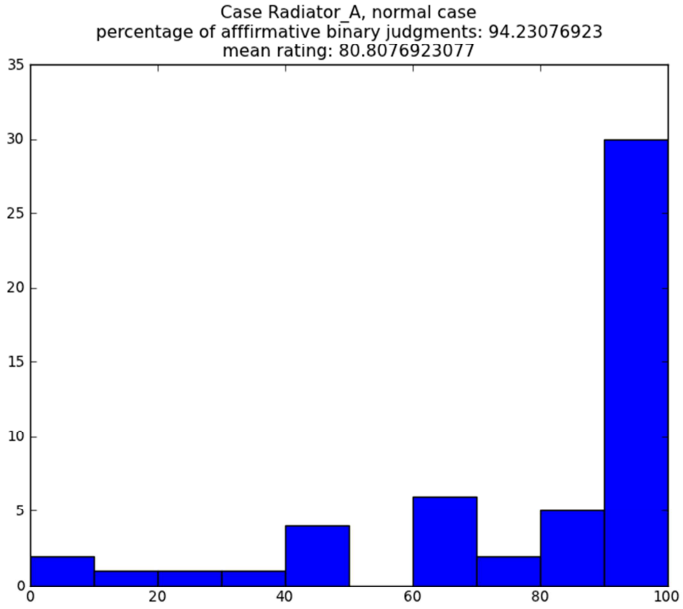


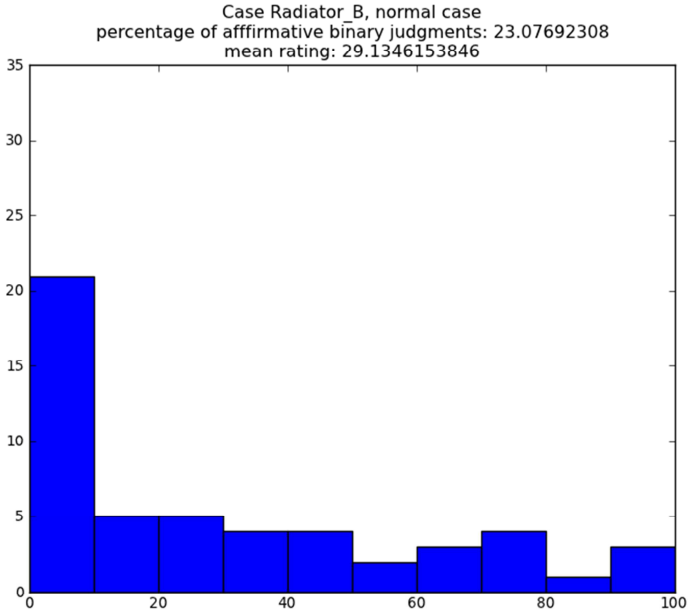
Case Guarded_Speedboat_A, normal case
percentage of afffirmative binary judgments: 71.15384615
mean rating: 74.3076923077

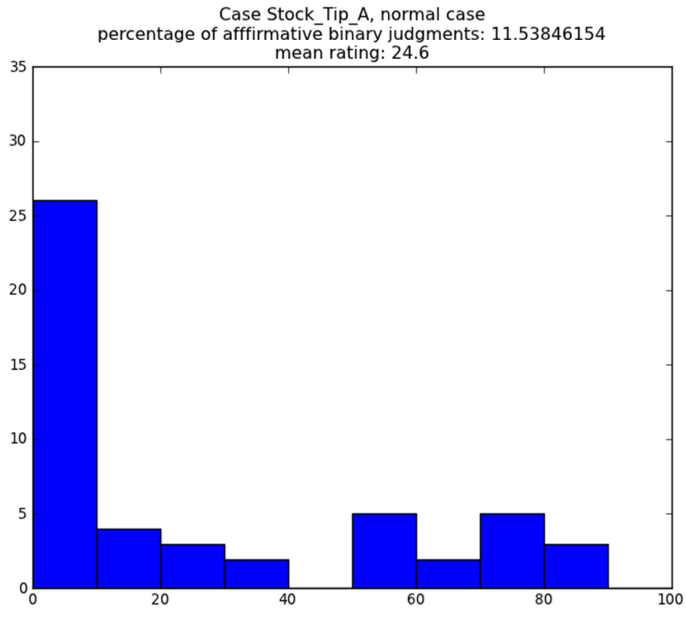


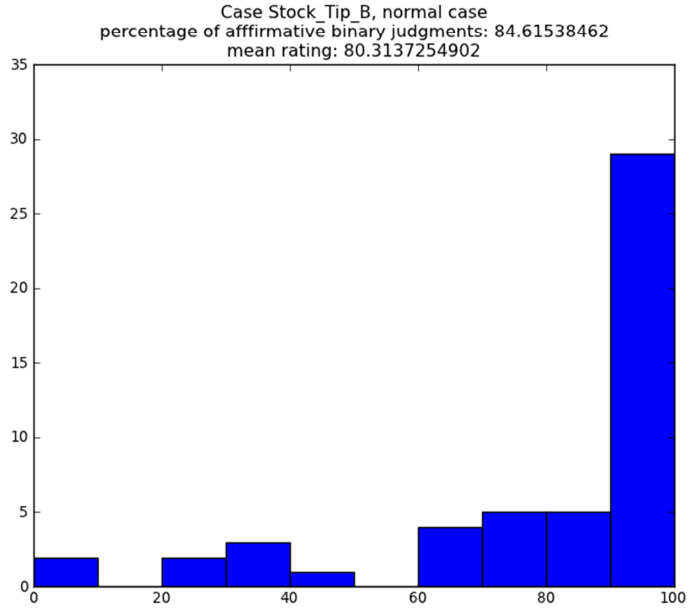

Case Landlord_A, normal case
percentage of afffirmative binary judgments: 23.07692308
mean rating: 48.3725490196

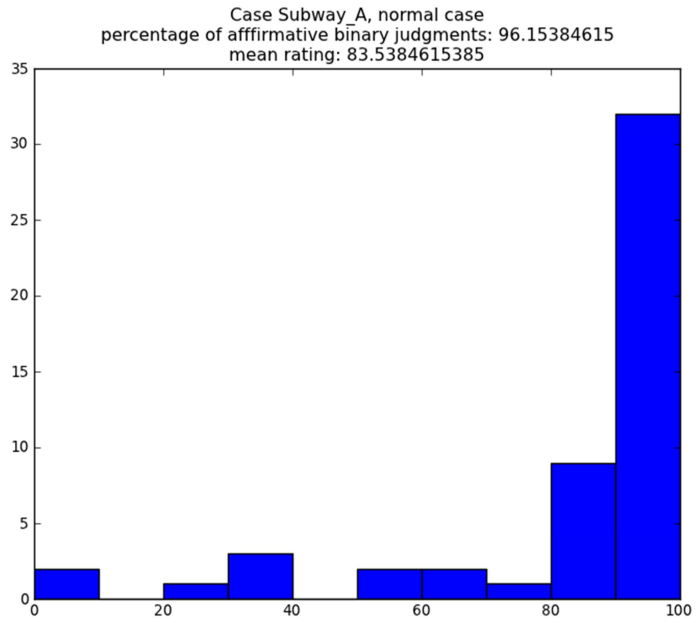

Case Landlord_B, normal case
percentage of afffirmative binary judgments: 78.84615385
mean rating: 77.9607843137

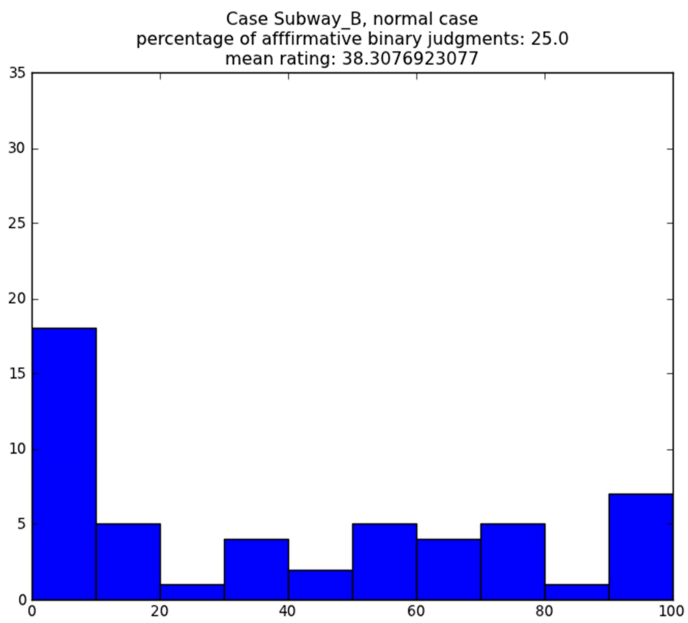

Case Radiator_A, normal case
percentage of afffirmative binary judgments: 94.23076923
mean rating: 80.8076923077


Case Radiator_B, normal case
percentage of afffirmative binary judgments: 23.07692308
mean rating: 29.1346153846

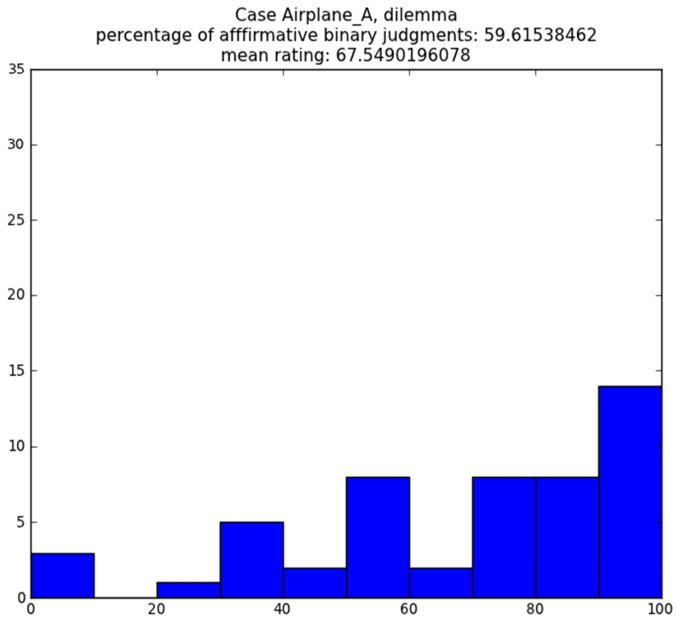

Case Stock_Tip_A, normal case
percentage of afffirmative binary judgments: 11.53846154
mean rating: 24.6

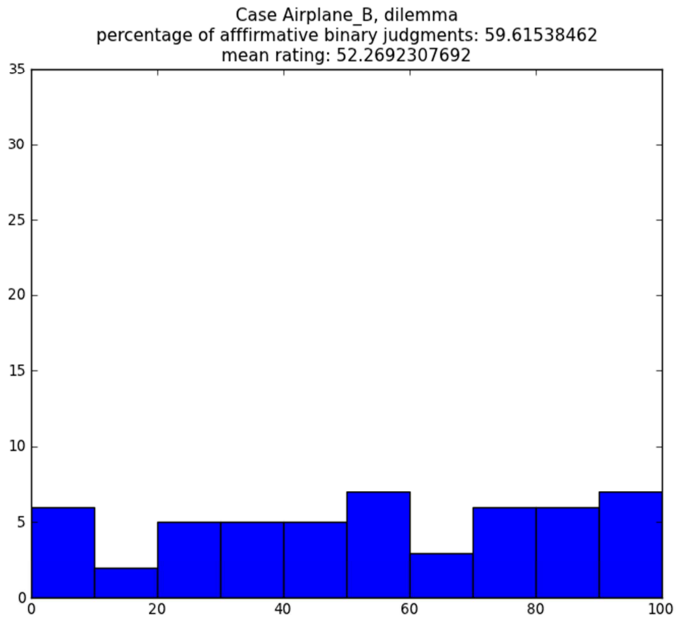

Case Stock_Tip_B, normal case
percentage of afffirmative binary judgments: 84.61538462
mean rating: 80.3137254902

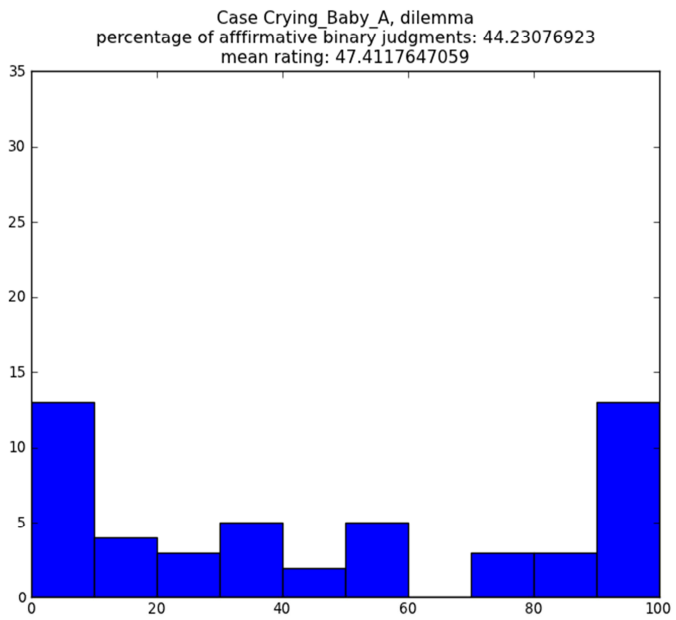

Case Subway_A, normal case
percentage of afffirmative binary judgments: 96.15384615
mean rating: 83.5384615385
0
5
10
15
20
25
30
35
0
20
40
60
80
100

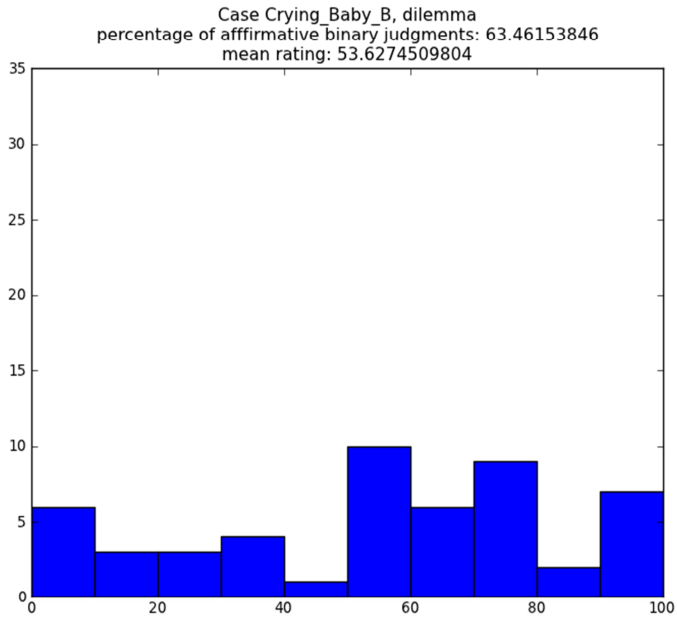

Case Subway_B, normal case
percentage of afffirmative binary judgments: 25.0
mean rating: 38.3076923077
0
5
10
15
20
25
30
35
0
20
40
60
80
100

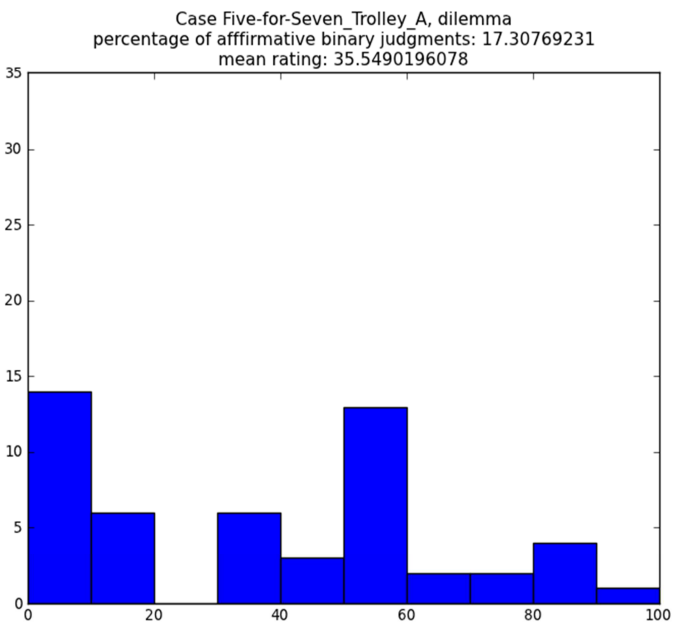
6.1.2. Histograms of dilemma cases

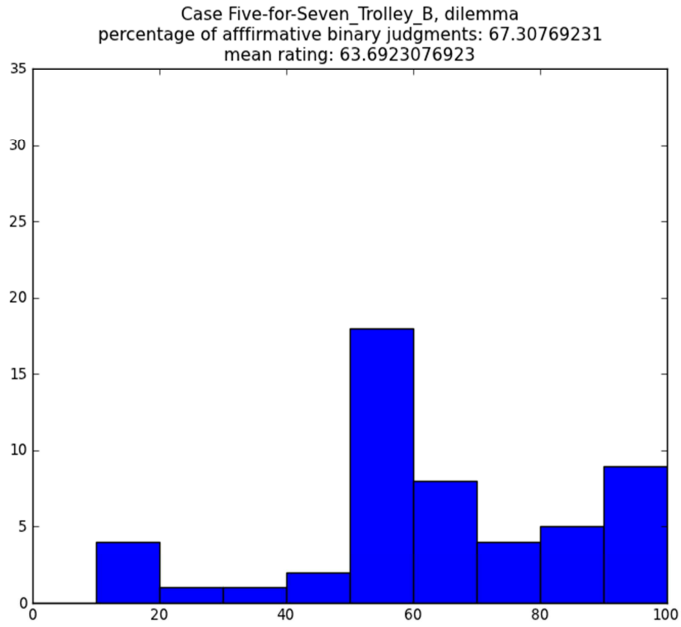


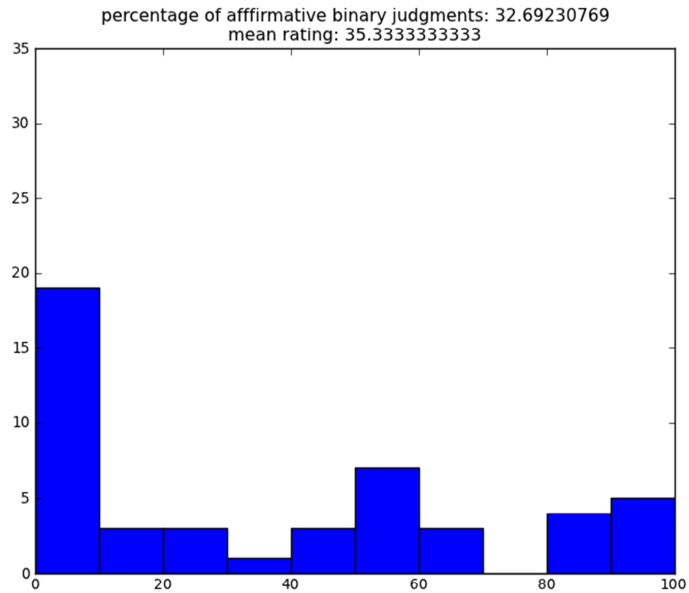


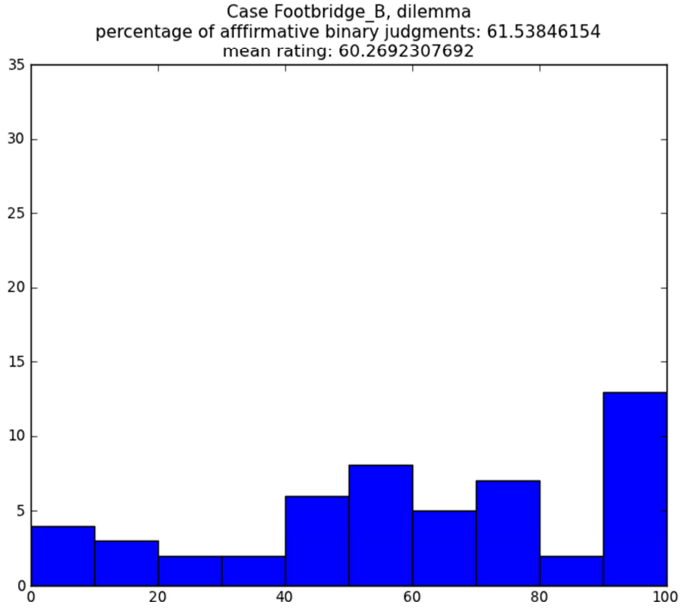


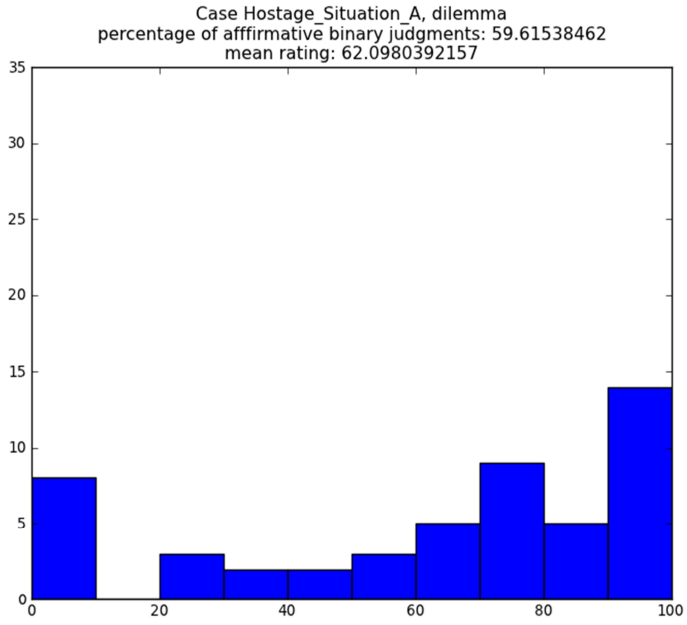


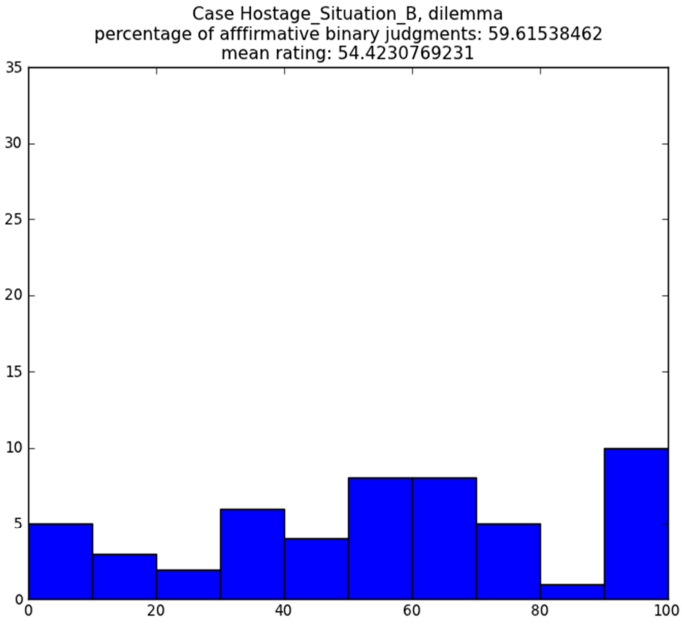

Case Five-for-Seven_Trolley_A, dilemma
percentage of afffirmative binary judgments: 17.30769231
mean rating: 35.5490196078


Case Five-for-Seven_Trolley_B, dilemma
percentage of afffirmative binary judgments: 67.30769231
mean rating: 63.6923076923

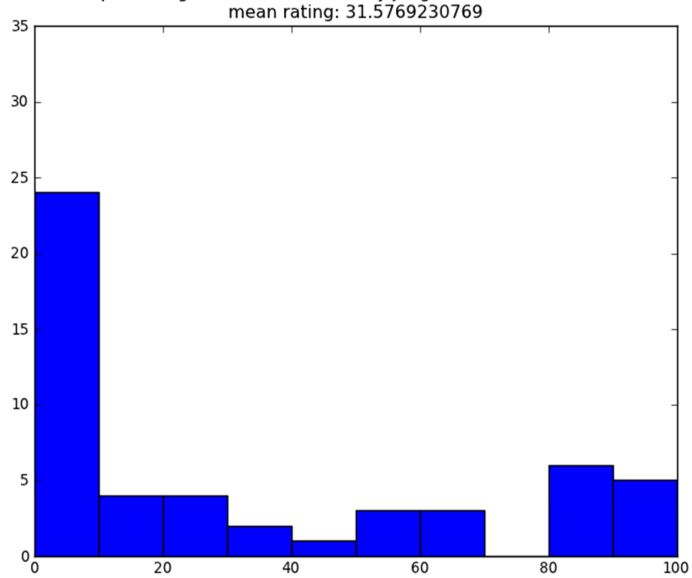

percentage of afffirmative binary judgments: 32.69230769
mean rating: 35.3333333333


Case Footbridge_B, dilemma
percentage of afffirmative binary judgments: 61.53846154
mean rating: 60.2692307692

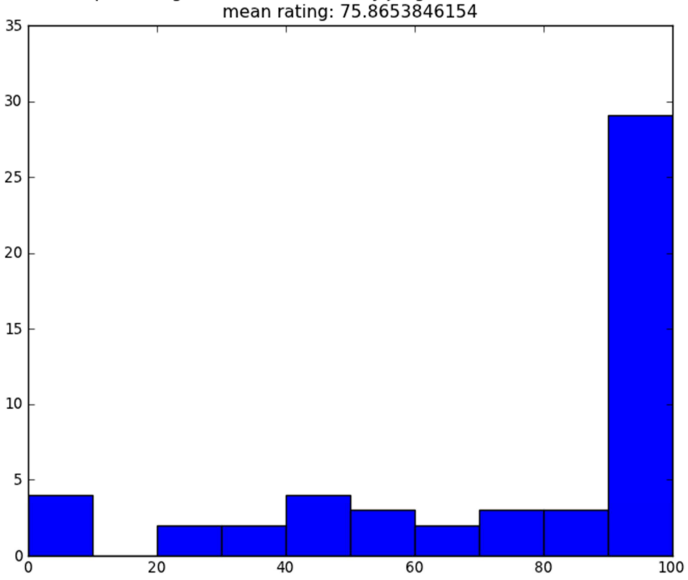

Case Hostage_Situation_A, dilemma
percentage of afffirmative binary judgments: 59.61538462
mean rating: 62.0980392157


Case Hostage_Situation_B, dilemma
percentage of afffirmative binary judgments: 59.61538462
mean rating: 54.4230769231

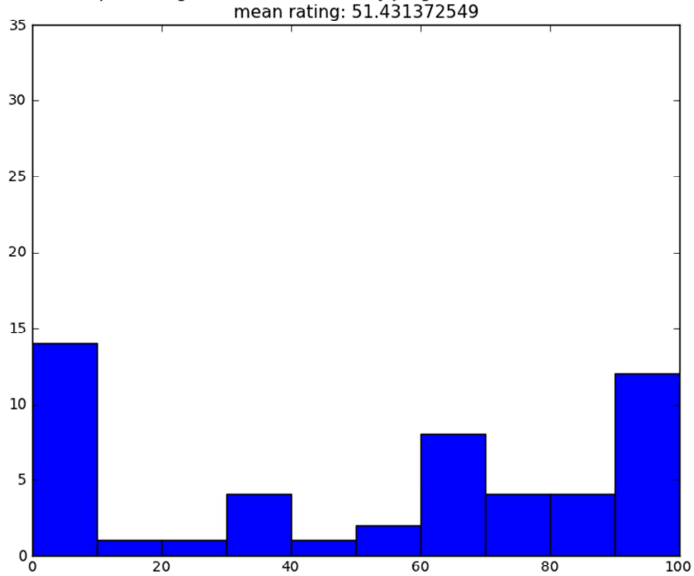
Case Lost_Wallet_A, dilemma
percentage of afffirmative binary judgments: 15.38461538
mean rating: 31.5769230769



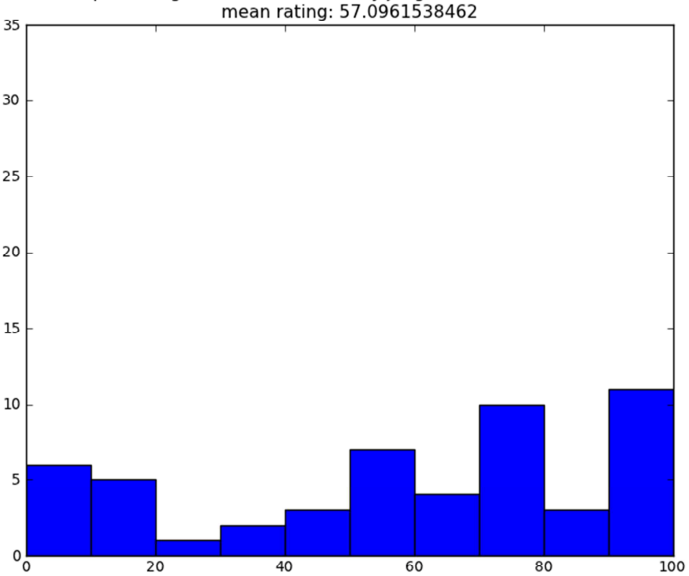
Case Lost_Wallet_B, dilemma
percentage of afffirmative binary judgments: 88.46153846
mean rating: 75.8653846154



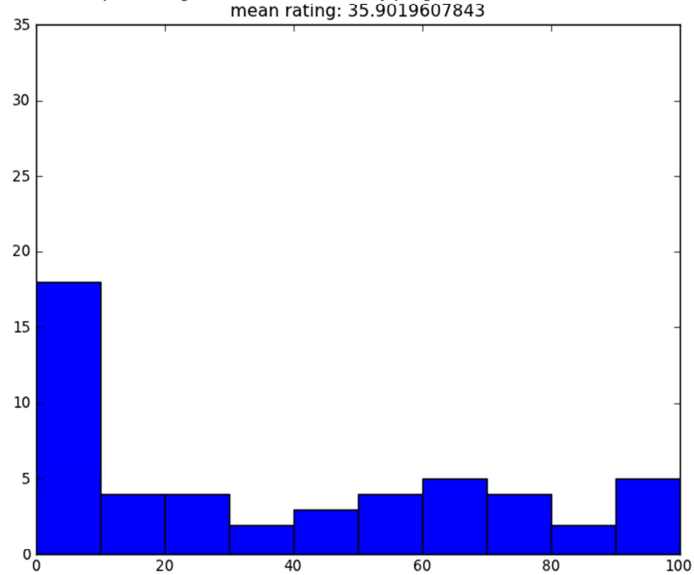
Case Modified_Lifeboat_A, dilemma
percentage of afffirmative binary judgments: 57.69230769
mean rating: 51.431372549



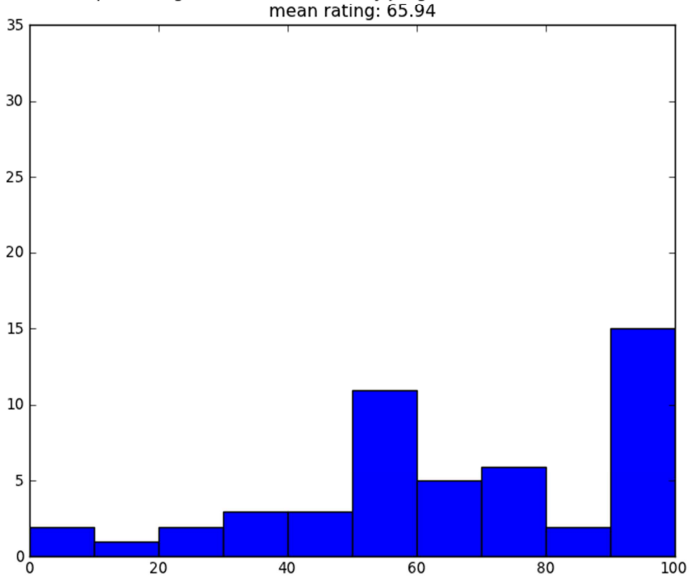
Case Modified_Lifeboat_B, dilemma
percentage of afffirmative binary judgments: 46.15384615
mean rating: 57.0961538462

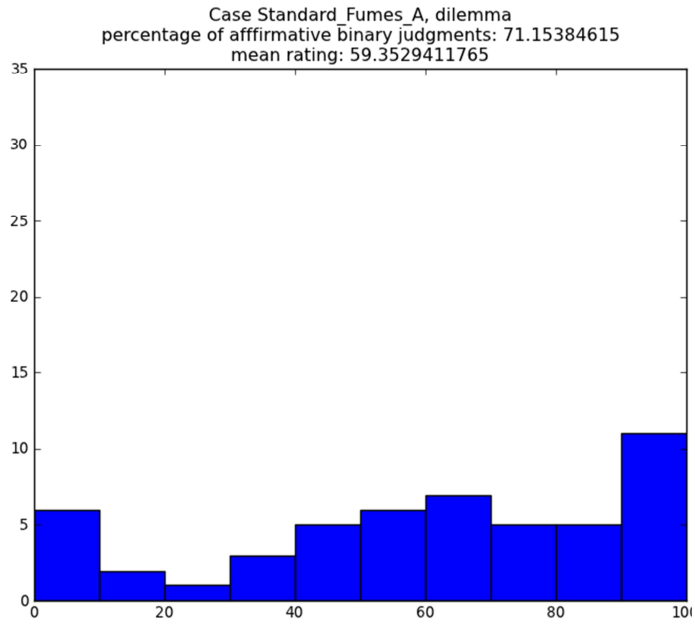


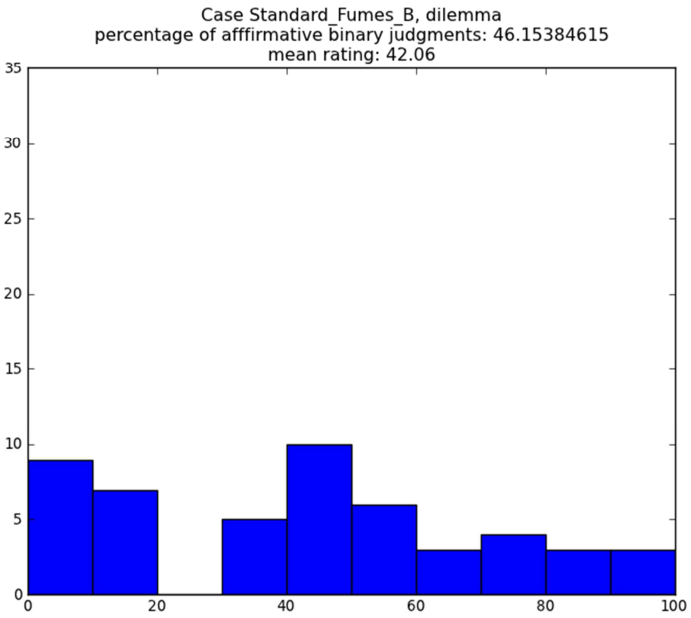
Case Sophie_Choice_A, dilemma
percentage of afffirmative binary judgments: 42.30769231
mean rating: 35.9019607843

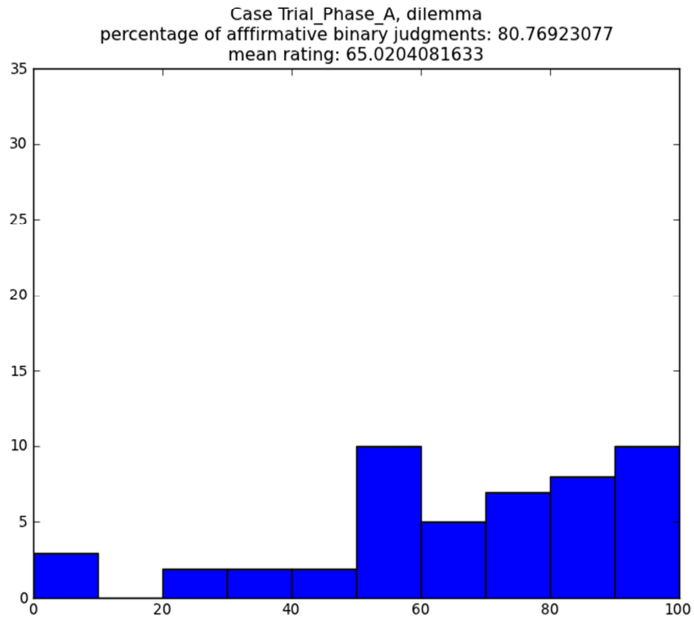


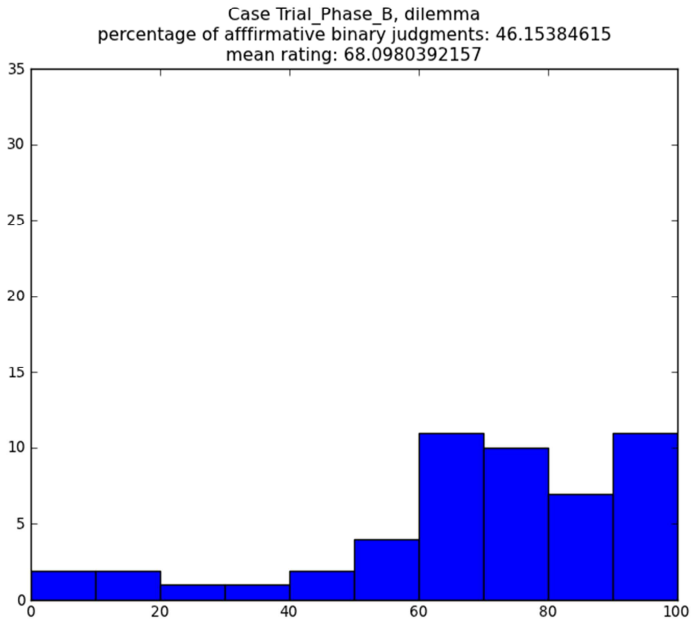
Case Sophie_Choice_B, dilemma
percentage of afffirmative binary judgments: 59.61538462
mean rating: 65.94

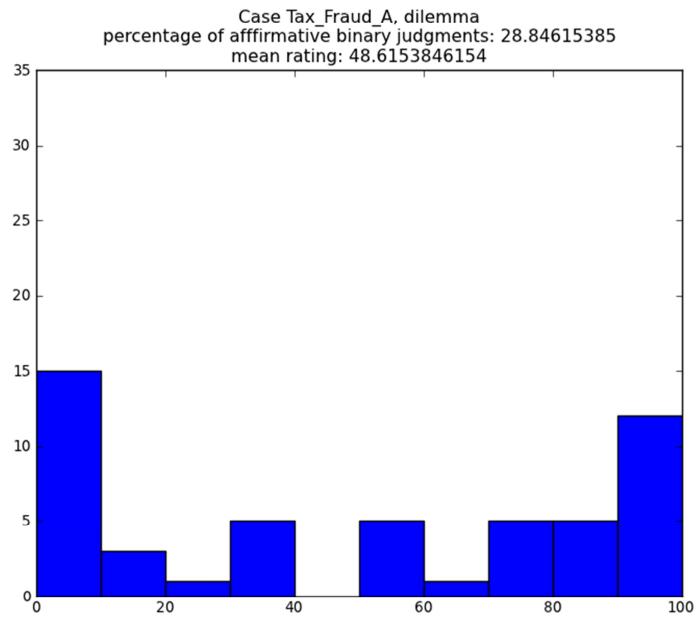


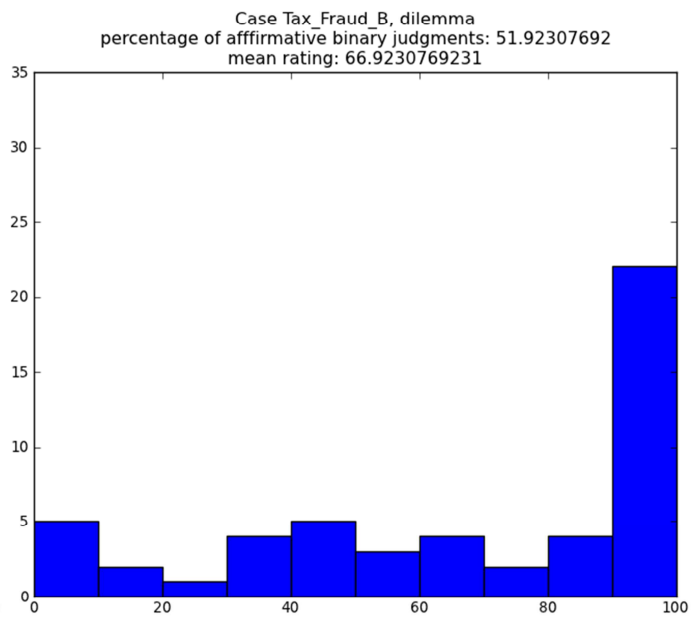

Case Standard_Fumes_A, dilemma
percentage of afffirmative binary judgments: 71.15384615
mean rating: 59.3529411765


Case Standard_Fumes_B, dilemma
percentage of afffirmative binary judgments: 46.15384615
mean rating: 42.06


Case Trial_Phase_A, dilemma
percentage of afffirmative binary judgments: 80.76923077
mean rating: 65.0204081633


Case Trial_Phase_B, dilemma
percentage of afffirmative binary judgments: 46.15384615
mean rating: 68.0980392157


Case Tax_Fraud_A, dilemma
percentage of afffirmative binary judgments: 28.84615385
mean rating: 48.6153846154


Case Tax_Fraud_B, dilemma
percentage of afffirmative binary judgments: 51.92307692
mean rating: 66.9230769231

7. **Bibliography**

Aristotle (1999): *Nicomachean Ethics.* Translated with Introduction, Notes, and Glossary by Terence Irwin. Hackett Publishing, Indianapolis

Bargh, J. A. (1994): The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In: Wyer, R. S., Jr. & Srull, T. K. (Ed): *Handbook of social cognition, Vol. 1: Basic processes; Vol. 2: Applications (2nd ed.)*. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc, 1994

Ben-Menahem, Y. (2006): *Conventionalism*, Cambridge: Cambridge University Press

Bentham, J. (1996): *An Introduction to the Principles of Morals and Legislation.* An authoriatative Edition by J.H. Burns and H.L.A. Hart with a New Introduction by F.Rosne and an Interpretive Essay by H.L.A. Hart. Oxford, OUP

Carnap, R. (1950): Empiricism, Semantics and Ontology. In: *Revue Intérnationale de Philosophie*, 4: 20–40. Reprinted in Carnap, R. 1956: *Meaning and Necessity: A Study in Semantic and Modal Logic*, Chicago: University of Chicago Press.

Carnap, R. (1967): Logical Foundations of Probability. University of Chicago Press; 2nd edition

Charland, L.C. (1997): Reconciling Cognitive and Perceptual Theories of Emotion: A Represenational Proposal. In: *Philosophy of Science*, Vol. 64, No.4, 1997, pp. 555 – 579

Christensen, J.F. & Gomila, A. (2012): Moral dilemmas in cognitive neuroscience of moral decision-making:A principled review. In: *Neuroscience and Biobehavioral Reviews* 36 1249–1264

Churchland, P.S. (2011): *Braintrust: What Neuroscience Tells Us about Morality* , Princeton 2011

Churchland, P.M. (1981): Eliminative Materialism and the Propositional Attitudes. In: Journal of Philosophy. pp. 67-90

Churchland, P.M. (1998): Toward a Cognitive Neurobiology of the Moral Virtues. In: *Topoi* 17, pp. 83–96

Cosmides, L. & Tooby, J. (1994): Origins of domain-specificity: The evolution of functional organization. In L. Hirschfeld & S. Gelman (Eds.): *Mapping the Mind: Domain-specificity in cognition and culture*. New York: Cambridge University Press.

Damasio, A. (1994): *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam Publishing

Damasio, A. (1996): The somatic Marker hypothesis and the possible functions of the prefrontal cortex. In: *Phil. Tran. R. Soc. Lond*. B 351, pp. 1413 – 1420

Deonna, J. & Teroni, F. (2012): *The Emotions: A Philosophical Introduction.* Taylor & Francis 2012

Duhem, P. M. M., 1954 (1906), *The Aim and Structure of Physical Theory*, P. P. Wiener (tr.), Princeton: Princeton University Press.

Dworkin, R. (1977): *Taking Rights Seriously.* Harvard University Press

Evans, J. St. B. T. (1982): *The Psychology of Deductive Reasoning*. London. Routlege & Kegan Paul

Fechner, G. (1966): *Elements of psychophysics; translated by Helmut E. Adler* ; edited by Davis H.Howes, Edwin G. Boring ; with an introduction by Edwin G. Boring. New York, Holt, Rinehart and Winston

Frege, G. (1892): Über Sinn und Bedeutung. In: *Zeitschrift für Philosophie und philosophische Kritik*, NF 100, S. 25-50.

Frege, G. (1918): Thoughts. In his *Logical Investigations*. Oxford: Blackwell, 1977.

Galotti, K. M. (1989). Approaches to studying formal and everyday reasoning. In: *Psychological Bulletin*, 105, 331-351.

Gazzaniga, M.S. & Le Doux, J.E. (1978): *The Integrated Mind*. New York: Plenum Pr.

Gettier, E. (1963): Is Justified True Belief Knowledge?, in: *Analysis* 23 , pp. 121–123

Greene, J. (2007): The Secret Joke of Kant's Soul. In: Sinnott-Armstrong W: *Moral Psychology, Vol. 3: The Neuroscience of Morality: Emotion, Disease, and Development.*

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. In: *Neuron*, 44, 389–400.

Greene JD, Morelli SA, Lowenberg K, Nystrom LE, Cohen JD. (2008): Cognitive Load Selectively Interferes with Utilitarian Moral Judgment. In: *Cognition* 107, pp. 1144-1154.

Goldman, A. (2007): Philosophical Intuitions: their target, their source, and their epistemic status. In: *Grazer Philosophische Studien 74*, pp. 1–26.

Haidt, J. (2001): The Emotional Dog and Its Rational Tail: A Social Intuitionist. Approach to Moral Judgment. In: *Psychological Review* 108.4

Haidt, J., & Joseph, C. (2007): The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. In: P. Carruthers, S. Laurence, and S. Stich (Eds.): *The Innate Mind*, Vol. 3., pp.367

Haidt, J., & Joseph, C. (2004): Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues. In: *Daedalus*, pp. 55-66, Special issue on human nature.

Haidt, Koller & Dias (1993). Affect, culture, and morality, or is it wrong to eat your dog? In: *Journal of Personality and Social Psychology* 65, pp. 613-628

Haidt, J. (2003). The moral emotions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.),. Handbook of affective sciences. Oxford: Oxford University Press.

Harman, G (1975): Moral relativism defended. In: *The Philosophical Review*, Vol. 84, No. 1, pp. 3-22.

Harman, G. & Thompson, J.J. (1996): *Moral Relativism and Moral Objectivity.* Wiley - Blackwell

Hume, D. (1975): *A Treatise of Human Nature*, ed. L.A. Selby-Bigge, rev. P.H. Nidditch, Oxford: Clarendon Press, 1975

Holyoak, K.J. & Nisbett, R.E. (1988): Induction. In: R.J. Sternberg & E.E. Smith (Eds.): *The Psychology of Human Thought.* Cambridge, Engalnd. CUP

International Commission on Microbiological Specifications for Foods (1998): *Microorganisms in Foods 6: Microbial Ecology of Food Commodities,* Originally published by Chapman & Hall

James, W. (1884): What is an Emotion? In: *Mind*, 9, 188-205.

Johnson, R. N. (2008): Was Kant a Virtue Ethicist?. In: *Kant's Ethics of Virtue*, M. Betzler, ed., Berlin: DeGruyter, 61-76

Kahane, G.(2012): On the Wrong Track: Process and Content in Moral Psychology. In:*Mind & Language*, Volume 27, Issue 5, pages 519–545, November 2012

Kahnemann, D. (2012): *Thinking, Fast and Slow*, Penguin

Kant, I. (1902ff): *Grundlegung zur Metaphysik der Sitten*, zitiert nach der Akademieausgabe, Berlin, Band 4

Kant, I. (1993); translated by James W. Ellington [1785] (1993). *Grounding for the Metaphysics of Morals* , 3rd ed. Hackett. p. 30

Kant, I. (1996): Practical Philosophy (M. Gregor, Trans.). Cambridge University Press, Cambridge, UK.

Kaufmann, D. & Kleinknecht, J. (2013): Experimentelle Philosophie. In: Bonk: *Lexikon der Erkenntnistheorie*, WBG

Kennett , J. & Fine, C. (2009): Will the real moral judgment please stand up? In: *Ethic Theory Moral Prac* 12, pp. 77–96

Kingdom, F.A.A. & Prins, N.: *Psychophysics – A Practical Introduction*. Academic Press, 2009

Knobe, J. (2003): Intentional Action in Folk Psychology: An Experimental Investigation. In: *Philosophical Psychology*, 16, 309-324.

Kohlberg, L. (1969). Stage and sequence: The cognitive-developmental approach to socialization. In D. A. Goslin (Ed.), *Handbook of socialization theory and research*. Chicago: Rand McNally. Maddy, P. (2007): *Second Philosophy*, Oxford University Press

Metcalfe, J., & Mischel, W. (1999). A hot/cool system analysis of delay of gratification: Dynamics of willpower. In: *Psychological Review*, 106, pp. 3-19.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. In: *Trends in Cognitive Sciences*, 11, pp. 143–152.

Mikhail, J. (2008): Moral Cognition and Computational Theory. In: Walter Sinnott-Armstrong, ed., *Moral Psychology*, Vol. 3: The Neuroscience of Morality (Cambridge: MIT Press), pp81-92

Mill, John Stuart (1972). U*tilitarianism, On Liberty, and Considerations on Representative Government*. Edited by H.B Acton. London, J.M Dent &Sons

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

Norcross, A. (2006): The Scalar Approach to Utilitarianism. In: West, Henry R..; *Blackwell Guide to Mill's Utilitarianism*, Blackwell

Nussbaum, M (2003): Emotions as Judgments of Value and Importance. In: R. Solomon (ed.) *Thinking about Feeling*. New York: OUP

Nussbaum, M. C. (1993). Non-relative virtues: an Aristotelian approach. In M. C. Nussbaum & A. Sen (Eds.), *The quality of life*, pp. 242-269. New York: OUP

Parfit, D. (1984): *Reasons and Persons*, Oxford: Oxford University Press.

Prinz, J. (2003): Emotions Embodied. In: R. Solomon (ed.) *Thinking about Feeling*. New York: OUP

Prinz, J. (2006): The Emotional Basis of Moral Judgements. In: *Philosophical Explorations*, Vol. 9, No. 1

Prinz, J. (2007): *The Emotional Construction of Morals*, OUP

Putnam, H. (1981): *Reason, Truth and History*, Cambridge

Putnam, H. (1982): Three Kinds of Scientific Realism. In: *Philosophical Quarterly*, 32: 195–200.

Putnam, H. (1990): *Realism with a Human Face*. Cambridge, Mass.: Harvard University Press

Putnam, H. (2004): *The Collapse of the Fact-Value-Dychotomy*, Harvard University Press

Quine, W.V.O. (1961): The two Dogmas of Empirircism. In Quine, W.V.O.: *From a Logical Point of View.* Harvard University Press, 1953; second, revised, edition 1961

Quine, W.V.O. (1953): On What There Is. In: In Quine, W.V.O.: *From a Logical Point of View.* Harvard University Press

Quine, W.V.O. (1969): Epistemology Naturalized. In: Quine, W.V.O.: *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Quine, W.V.O. (1970): *Philosophy of Logic*. Englewood Cliffs: Prentice Hall.

Ramsey, F.P. (1927) : Facts and Propositions. In: *Proceedings of the Aristotelian Society*, 7 (Supplementary): 153–170.

Rawls, J. (1971): *A Theory of Justice*. Cambridge, Massachusetts: Belknap Press of Harvard University Press

Roberts, R (2013): *Emotions in the Moral Life*, CUP

Rozin, P., Haidt, J., Fincher, K. (2009): From oral to moral. In: *Science*, Vol 323 (5918), pp. 1179-1180. doi: 10.1126/science.1170492

Rozin P., Lowery L., Imada S., Haidt J. (1999): The CAD triad hypothesis: a mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). In: *J Pers Soc Psychol.* (4), pp.574-86

Russel, S.J. & Norvis, P. (2010): *Artificial Intelligence: A modern approach*, Prentice Hall

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

Sauer, H. (2012): Morally irrelevant factors: What's left of the dual process-model of moral cognition?. In: *Philosophical Psychology* 25 (6):783-811

Searle, John (1984), *Minds, Brains and Science: The 1984 Reith Lectures*, Harvard University Press

Sellars, W. (1997): *Empiricism and the Philosophy of Mind*, edited by Robert Brandom, Harvard University Press.; Cambridge, MA

Sherman, N. (1998): *Making a Necessity of Virtue*. New York: Cambridge U. P.

Shweder, R. A., Much, N. C, Mahapatra, M., & Park, L. (1997): *The "Big Three" of morality (autonomy, community, divinity) and the "Big Three" explanations of suffering. In A. Brandt & P. Rozin (Eds.), Morality and health (pp. 119-169). New York: Routledge*

Sinnott-Armstrong, W., Young, L., & Cushman, F. (2010): Moral intuitions. In J. M. Doris & The Moral Psychology Research Group (Eds.): *The moral psychology handbook* (pp.246–272). Oxford, England: Oxford University Press.

Sommer, M., Rothmayr, C., Döhnel, K., Meinhardt, J., Schwerdtner, J., Sodian, B., & Hajak, G. (2010): How should I decide? The neural correlates of everyday moral reasoning. In: *Neuropsychologia,* 48, pp. 2018-2026.

Sosa, (2006): Experimental philosophy and philosophical intuition. In: *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, Vol. 132, No. 1, Selected Papers from the American Philosophical Association, Pacific Division, 2006 Meeting (Jan., 2007), pp. 99-107

Stein, R., & Nemeroff, C. J. (1995). Moral overtones of food: Judgments of others based on what they eat. In: *Personality and Social Psychology Bulletin*, 21, pp. 480-490.

Tannenbaum, D.; Uhlmann, E. L.; Diermeier, D.(2011): Moral signals, public outrage, and immaterial harms. In: *Journal of Experimental Social Psychology* 47 (6), pp. 1249–1254.

Timmons, M. (2008): Toward a Sentimentalist Deontology. In: *Moral Psychology*, vol. 3, Walter-Sinnott-Armstrong (ed.), Oxford University Press, 2008., pp. 93 – 104

Turiel, E. (1983): *The development of Social Knowledge – Morality and convention*. Cambridge University Press

van Fraassen, B. C. (1980): *The Scientific Image*, Oxford: Oxford University Press.

Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford Handbook of Thinking and Reasoning* (pp. 364-389). New York: Oxford University Press.

Weinberg, J., Nichols, S. and Stich, S. (2001) ;Normativity and Epistemic Intuitions. In: *Philosophical Topics*, 29, pp. 429–460.

Wheatley, T. & Haidt, J. (2005): Hypnotic Disgust Makes Moral Judgments More Severe. In: *Psychological Science*, Vol 16, No 10

Williams, B. (1973): A Critique of Utilitarianism. In J. J. C. Smart and B. Williams: *Utilitarianism: For & Against*. Cambridge: Cambridge University Press, pp. 77–150.

Williamson, T. (2004): Philosophical 'Intuitions' and Scepticism about Judgement. In: *Dialectica* Vol. 58, No 1, pp. 109-153

Wilson, M. (2002): Six views of embodied cognition. In: *Psychonomic Bulletin & Review* 9 (4), pp. 625-636

Wittgenstein, L. (2003): *Philosophische Untersuchungen*, Suhrkamp

Young, L.; Bechara, A.; Tranel, D.; Damasio, H.; Hauser, M.; Damasio, A. (2010): Damage to Ventromedial Prefrontal Cortex Impairs Judgment of Harmful Intent. In: *Neuron* 65 (6), pp. 845–851.

Young, l. & Saxe, R. (2011): When ignorance is no excuse: Different roles for intent across moral domains, in: *Cognition* 120 (2), p.202-214

.

8. **Publications**

Kaufmann, D. & Kleinknecht, J.: „Experimentelle Philosophie“ in Bonk, T. (Hrsg.): *Lexikon der Erkenntnistheorie*, WBG, 2013

9. Eidesstattliche Versicherung / Affidavit

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation „How can we know what is “moral”? - Philosophical commitments in empirical research on moral judgment “ selbstständig angefertigt habe, mich außer der angegebenen keiner weiteren Hilfsmittel bedient habe und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

I hereby confirm that the dissertation „How can we know what is “moral”? - Philosophical commitments in empirical research on moral judgment “ is the result of my own work and that I have only used sources or materials listed and specified in the dissertation.

Munich, September 18, 2015

David Kaufmann

10. **Contributions**

I conceived all of the hypotheses elaborated in chapter 1 as well as the conceptual analysis of the exemplary theory of moral judgment and the exemplary supporting empirical study in chapter 2. I originated all arguments figured in the discussion of evidence for my hypotheses in chapter 3. I elaborated the example of the subtler way of arguing interdisciplinarily presented in chapter 4.2 . In collaboration with Stefan Glasauer, I conceived the method of the psychophysical study referred to in chapter 4. 3 and supervised the data collection by Gloria Benson. The ensuing data analysis was performed by me in collaboration with Stefan Glasauer.

Munich, September 18, 2015

David Kaufmann

Prof. Dr. Stephan Sellmaier (Supervisor)