ARE TROLLEY DILEMMA JUDGEMENT MECHANISMS EVOLUTIONARY ADAPTATIONS?

Lara Pourabdolrahim Seresht Ardebili

Graduate School of
Systemic Neurosciences
LMU Munich

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

May 2018

Supervisor
Prof. Dr. Stephan Sellmaier
Research Center for Neurophilosophy and Ethics of Neurosciences
Graduate School of Systemic Neurosciences

First Reviewer:	Prof. Dr. Stephan Sellmaier
Second Reviewer:	Prof. Dr. Dr. Hannes Leitgeb
External Reviewer:	Assoc. Prof. Dr. Stephan Schleim

Date of Submission:	02/05/2018
Date of Defense :	28/09/2018

Contents

1 Introduction

Over the last few decades, something of a 'morbid' trend has arisen in moral psychology and moral philosophy: The use of Trolley Dilemmas. Trolley Dilemmas are thought experiments in which a train is heading towards five people. The train will hit and kill those five people, unless someone is sacrificed in order to stop the train. Trolley Dilemmas are enormously useful, in a variety of ways.

1.1 Brief History of Trolley Dilemmas

The first Trolley Dilemma, a version of the so-called side-track scenario, was used (amongst an abundance of other hypothetical scenarios) by ethicist Philippa Foot to evoke intuitions[1] about when it is permissible to kill a person to save another one in the context of a discussion about abortions.[2] Its wording was as follows: "[. . .] the driver of a runaway tram [...] can only steer [the tram] from one narrow track on to another; five men are working on one track and one man on the other; anyone on the track he enters is bound to be killed." (Foot 1967, p. 7). She assumes that steering the train onto the track with the one person would be deemed permissible or even obligatory by most people (Foot 1967, pp. 7, 10). Judith Jarvis Thomson and Frances Myrna Kamm followed her in her use of train-track life/death scenarios and added more variants to the pool of Trolley Dilemmas, such as the Loop Case (Thomson 1985, pp. 1402/1403), where the track with the one man loops back to the track with the five. This dilemma reads as follows: "Let us now imagine that the five on the straight track are thin, but thick enough so that although all five will be killed if the trolley goes straight, the bodies of the five will stop it, and it will therefore not reach the one. On the other hand, the one on the right-hand track is fat, so fat that his body will by itself stop the trolley, and the trolley will therefore not reach the five. May the agent turn the trolley?" (Thomson 1985, 1403). Thomson then adds: "Some people feel more discomfort at the idea of turning the trolley in the loop variant than in the original Bystander at the Switch"(Thomson 1985, 1403). All three Ethicists use their

[1] Like Peter van Inwagen (Inwagen 1997) and Ernest Sosa (Sosa 2007), I will take intuitions to be (conscious) inclinations to believe, or, as Williamson rephrased Peter van Inwagen ((van Inwagen 1997, p. 309), cited after (Williamson 2004, p. 126)): "in some cases, the tendencies to make certain beliefs attractive to us". Hence, if I intuit that Person One should turn the trolley to the track with the one person, I am inclined to believe that Person One should turn the trolley to the track with the one person. The issue of whether a moral statement such as "one should turn the trolley to the track with the one person" bears truth value, hence, can be true or false (as Sosa holds for the contents of "propositional intuitions" (Sosa 2007, p. 52), is a matter for a different discussion. For a discussion of this matter, see (Stratton-Lake and Zalta 2016). For a discussion of different accounts of "intuition", see (Pust 2017) and (Kauppinen 2015). Sosa defends a somewhat narrower definition of what he calls "propositional intuitions" than the one I will apply: "S intuits that p if and only if S's attraction to assent to <p> is explained rationally by two things in combination: (a) that S understands it well enough, (b) that <p> is true." (Sosa 2007, p 52)

[2] This is a context in which she placed the scenario; I do not believe that abortion is a matter of one person's rights against another's.

(or, as in Thomson's case, their and other people's) opinions about the permissibility to kill the one and then seek to suggest normative principles that explain those opinions and can predict intuitions for further cases (and, if they are established, guidelines for further actions): [3] They look for moral principles that are coherent with all or most of those opinions and, ideally, seem to transport some moral relevance. To provide an example for coherent principles without moral relevance: If we came up with a principle like "people wearing brown shirts may be sacrificed in any case to save other people" this may not be an appropriate moral principle, even if it was coherent with all our intuitions (we had until then only deemed sacrificing people with brown shirts permissible in all scenarios and this is the only obvious feature that discerns the people whom we may sacrifice from the ones we may not). This is because wearing brown (where the color does not refer to anything morally relevant such as a political opinion) is usually not seen as a morally relevant feature (unlike, for instance, using someone merely as a means to achieve something but not as an end; see (Kant 1785, p. 38)).

The requirement for those principles to reflect properties that most people would accept as morally relevant will be of some importance when we come to a different usage of Trolley Dilemmas:

In 2001, neuroscientists entered the field of "trolleyology". Joshua Greene et al. (Greene et al. 2001) used fMRI to compare how their subjects' brains processed two different types of dilemmas: the classical side-track dilemma and a scenario going by the name "footbridge dilemma". If confronted with the former, many people agree that it is permissible to save the five by sacrificing the one on the side track (see, for instance, (Cushman et al. 2007) and (Foot 1967, pp. 7, 10). Confronted with the latter, in which a large person is shoved in front of the trolley to stop it, many people find it impermissible to sacrifice the large person, including "Everybody to whom I have put this case" according to Thomson (Thomson 1985 p. 1409), who invented this scenario. Greene et al. 2001 hypothesized that most people find it more permissible to turn the trolley onto the side track and kill one than to shove the person in front of the trolley and kill them not (only) because of the moral principles at work in their decision-making, but because the "footbridge dilemma" elicited negative emotions which then led the subjects to decide against the negatively connotated action. They predicted and found "that brain areas associated with emotion would be more active during contemplation of dilemmas such as the footbridge dilemma as compared to during contemplation of dilemmas such as the [side track] trolley dilemma" (Greene et al. 2001, p. 2106) and longer reaction times for those who had to override their

[3] See also "reflective equilibrium", for instance, Rawls 2007. A reflective equilibrium also incorporates the following: "A rule is amended if yields an inference we are unwilling to accept; an inference is rejected if it violates a rule we are unwilling to amend."(Goodman (1965), pp. 66-67; cited after: STEPHEN STICH 1988: REFLECTIVE EQUILIBRIUM, ANALYTIC EPISTEMOLOGY AND THE PROBLEM OF COGNITIVE DIVERSITY, p. 394). In our case, principles we are unwilling to accept are rejected; intuitions are rejected if they violate a principle we are unwilling to amend.

emotions and declared shoving the man from the footbridge permissible (Greene et al. 2001, p. 2106).

In 2008, Greene published a (late) follow-up article: In “The Secret Joke of Kant’s Soul”, he provocatively argued that people who deemed shoving the person in the footbridge dilemma impermissible (according to him, a “deontological judgement”) were not making morally rational judgements based on reflection and that deontological philosophy in general was “to a large extent” just “an exercise in moral rationalization” (Greene 2008, p. 36) because they were simply defending their emotionally based decisions.

Both articles, alongside a storm of protest (see, for instance, (Kahane et al. 2015)and (Kahane 2015)) triggered a revival of Trolley Dilemma-inspired philosophy and a rise of psychological Trolley Dilemma experiments and empirically informed moral philosophy. Or maybe the time was just ripe: Other people started to research the topic around the same time. John Mikhail was concerned with Trolley Dilemmas as early as 2000, when he published his dissertation “Rawls’ Linguistic Analogy. A Study of the ‘Generative Grammar’ Model of Moral Theory Described by John Rawls in “A Theory of Justice” (Mikhail 2000) which leaned heavily on Trolley Dilemma experiments. In 2012 he wrote a book entitled “Elements of Moral Cognition” which will be one of the main sources of this dissertation and which was referred to as “John Mikhail’s important book” and treated alongside Hume in John M. Doris’s, Edouard Machery’s and Stephen Stich’s essay “Can Psychologists Tell Us Anything About Morality?” (Doris, Machery, and Stich 2017) and is highly praised in a draft by Stich and Doris about “Moral Disagreement, Moral Realism and Moral Grammar” (Stich and Doris 2016). Susan Dwyer suggested an empirical account of moral principles even earlier, in 1999, with her article “Moral Competence” (Dwyer 1999). Trolley Dilemma research numbers finally took off explosively after Cushman et al.’s huge empirical Trolley Dilemma online study was published in 2007 (Cushman et al. 2007). We will consider the content of the latter three shortly.

The importance of Trolley Dilemmas has lately received another boost as self-driving vehicles reached the late development stages and have to be programmed to make the choice between killing one (the driver/passenger) to save many (potential run-overs), or letting the latter be killed and saving the driver/passenger.(See, e.g., Powell, Cheng, and Waldmann 2014) Now such dilemmas have suddenly become real.

Their significance in psychological and philosophical literature has since been unabated. As Waldmann puts it in his article “Moral Judgment”: “Trolley dilemmas have become the Drosophila for testing alternative philosophical and psychological theories of moral judgments in harm-based moral dilemmas” (Waldmann, Nagel, and Wiegmann 2012, p. 377).

1.2 A few preliminaries

One reason why I am attracted to Trolley Dilemmas is because, depending on how the person is sacrificed, people generally judge the sacrifice more or less permissible. To give one example, people find it more permissible (and more people find it permissible) to divert the train to a side track on which a person is standing and who will be hit and killed by it than to shove a person from a bridge in front of the train to stop it from killing the five people on the main track (see, for instance, Cushman et al., 2007). This is startling because in each case one person gets killed in order to save five. People judge different ways of killing the one person differently, and that tells us a lot about their moral judgement, e.g. that it is influenced by more than just the ratio between saved and sacrificed lives. Another reason Trolley Dilemmas are useful is that, as they are thought experiments, we can isolate parameters even more easily than we can in "real life" experiments.

Furthermore, a majority of people seems to judge Trolley Dilemmas according to a pattern of moral principles such as that actively harming someone is worse than letting someone get harmed; hence, they tend to judge harming the one person in Trolley Dilemmas worse if someone has to act in order to sacrifice them than if someone has to refrain from acting to sacrifice them (e.g. if the train is rushing towards one person anyway and the five persons are on the other track). This net of moral principles, as well as the fact that people seem to judge according to similar principles in very different cultures, (again, see Cushman et al., 2007) as well as a hint by John Rawls (Rawls, 1971) [4] motivated some people to devise the so-called Linguistic Analogy (see, e.g., Dwyer et al., 2009). This analogy is drawn between Noam Chomsky's Universal Grammar and the net of principles that seems to underlie moral judgements such as Trolley Dilemma judgements: In its strong form, its proponents hold that those principles are part of a universal, innate endowment [5] just as Chomsky hypothesized some kinds of linguistic principles (the Universal Grammar) to be universal and innate (e.g. Dwyer, 2006, p. 6), (J Mikhail, 2007, p. 143). And where innateness is at stake, Evolutionary Psychology is not far away, which in its strongest form suggests that all cognitive mechanisms are either products or by-products of evolutionary adaptations to Pleistocene environments. [6] By cognitive mechanisms, I mean the 'processing units' that 'compute' mental processes, [7] including "perceptual and linguistic processes, as well as conceptual, reasoning, and problem-solving ones" (Siegler, 1989, p. 354). Later, when I write about Mechanistic Modularity, I am

[4] "A conception of justice characterizes our moral sensibility when the everyday judgments we do make are in accordance with its principles. [...] A useful comparison here is with the problem of describing the sense of grammaticalness that we have for the sentences of our native language." (Rawls, 1971), referring to (Chomsky, 1965, pp. 3–9); (see also Mikhail, 2000)

[5] I will discuss the notions of universality and innateness at length in the Universality and the Poverty of Stimulus chapters.

[6] I will discuss Evolutionary Psychology and its proponents' claims in the chapter about Modularity and Evolutionary Psychology.

[7] On Marr's algorithmic level of computation, see (Marr, 1982).

going to use the word “mechanism” more specifically as informational processing unit responsible for one complex of behaviour. It is a processing unit that computes one evolutionarily advantageous function, but does not necessarily do this from environmental input to behavioural output and not necessarily informationally isolated. If an individual has this unit, it has an evolutionary advantage. On a related note, I am going to use the word “system” for an informational unit comprised of several sub-units, the mechanisms. As I aim to show in the chapter on Mechanistic Modularity and Evolutionary Psychology, many proponents of the Linguistic Analogy hypothesize that Trolley Type Dilemmas are not only innate, but also computed in a modular way, [8] and modularity is, in the tradition of Evolutionary Psychologist reasoning, closely related to evolutionary adaptedness.

In this dissertation, I will investigate whether the cognitive mechanism that computes Trolley Dilemmas is an evolutionary adaptation (some would call it a *moral* mechanism or, if they think it is an adaptation, a “moral faculty” (e.g., Hauser, 2006) and I sometimes I will use the term 'moral' for this from now on). [9] This is significant because it tells us, depending on whether and what part of the mechanism is an adaptation and hence part of our 'default' development, how susceptible we are or could be to environmental influences, hence how fixed our moral beliefs are, and whether there are any kinds of moralities that are harder for us to adopt than others. And, viewed in the bigger picture of Evolutionary Psychology, it shows us whether morality is one of the realms that fits into its picture of a mind full of adapted mechanisms. If Trolley Dilemma judgement mechanisms are not adaptations, the picture of a mind full of pre-set, adapted mechanisms becomes less comprehensive, lacking part of the best evidence for an adapted 'moral faculty'. My method in examining whether Trolley Dilemma judgement mechanisms are Evolutionary Psychologist adaptations or at least congenital (even if they are not an adaptation to an ancestral problem) will be to examine several points that proponents of the Linguistic Analogy have made to show that they are innate and to test whether they were right in claiming so. A large part of my work consists of taking up argumentations for the innateness of Trolley Dilemma judgement mechanisms and clarifying them in a way that they can be tested empirically (see, for instance, my Universality and Poverty of Stimulus chapters). Once the arguments are clear enough, I will review experiments to examine their empirical premises and, hence, their soundness. This classical philosophical work on concepts may seem rather exegetic but is necessary to test whether the proponents of the Linguistic Analogy are right with their innateness claims. A further reason for me to

[8] I will discuss notions of modularity in the chapters about Mechanistic and Functional Modularity.

[9] I will not discuss whether Trolley Dilemmas are the right way to test for our moral intuitions, and I will not define the term 'morality' here. If you do not agree with calling Trolley Dilemma judgements moral judgements, or with calling the mechanism producing them a moral mechanism, I do not disagree with you and you are welcome to mentally replace 'moral' with 'Trolley Dilemma' when I talk about the respective mechanisms and judgements. For the sake of brevity and also in line with the conventions of Linguistic Analogy proponents who use them exemplarily for moral judgements (e.g. J Mikhail, 2007), I will use the word 'moral' for them at times.

put a strong focus on elaborating those arguments (e.g. showing the intermediate steps between 'it is universal, hence it is innate') is to provide a framework with a 'checklist' of properties that adapted or genetically transmitted cognitive mechanisms of all kinds should have. Hence this dissertation can also be used as list of conceptually precise arguments for innateness that clarifies which empirical evidence is necessary to show whether a mechanism, especially one that deals with social behaviors, is genetically inheritable. I hope that this will be helpful for others who do similar research in the future.

Next, I will discuss a few choices I have made when writing this thesis. Although the issue of how Trolley Dilemma Judgements should be interpreted has been rather controversial (for a review of all principles in question see (Bruers and Braeckman, 2014)), I will concentrate on the following three principles (or slightly more complex versions of them, see the generativity part of the Poverty of Stimulus chapter) that seem to underlie Trolley Dilemma judgement patterns:

"(a) harm caused by action is worse than an equivalent harm caused by omission;

(b) harms caused as a means to some greater good are worse than equivalent harms caused as a foreseen side-effect of an action" [10] (Dwyer et al., 2009, p. 497), based on (Cushman et al., 2006) and

(c) harms caused by a combination of personal force (direct muscle impact, as in pushing someone with a pole) and spatial closeness (as in standing next to someone) are worse than harms caused by a combination of impersonal force (as in dropping someone through a switch-operated trap door) and spatial distance. (Greene, 2014, p. 23)

1.3 The significance of the Doctrine of Double-Effect

Principle b), in more elaborated form also known as the "Doctrine of Double-Effect", is particularly interesting because of its long tradition and because it is one of the most discussed principles in deontological ethics. A search of the "Philosopher's Index" for "Double-Effect" yields 367 results (it is not always called a "Doctrine", but a search of "Doctrine of Double-Effect" yields 123 results and "Principle of Double-Effect" yields 96), while "Categorical imperative" yields 804 results (searched on 24/01/2018). It has been applied to problems like collateral damage in war, harmful pain medication and self-sacrifice to save others (McIntyre 2014) and it is still considered relevant in fields like Medical Ethics (see, for instance, (Watt 2017)).

Some forms of the Doctrine have been around for a long time. Dana Kay Nelkin and Samuel C. Rickless (Nelkin and Rickless 2015, p. 405)believe that the distinction between intended and foreseen actions in moral contexts goes back to the Talmud (following a suggestion of Fischer, Ravizza, and Copp (Fischer, Ravizza, and Copp 1993)) and is

[10] Or a similar but more complex principle, as we shall see later.

reflected in Jewish law thereafter. The distinction comes up in Thomas of Aquinas' Summa Theologiae (Aquinas 1265, p. 41,42/ II–II, q. 64, a. 7.) as well, when he argues that homicide in self-defense is licit.

There, not only does he introduce the difference between intended and foreseen actions, but his example is also an action that has a good and a bad effect, namely defending oneself and (possibly) killing the aggressor, hence a Double-Effect. (See, for instance, (Cavanaugh 1997)). We will return to the Catholic tradition of the Doctrine of Double-Effect in the chapter on Poverty of Stimulus.

I will limit myself to these three principles because they are the only ones that have been tested systematically with respect to cultural background, religion etc., found to be universal and because they are the main principles to which proponents of the Linguistic Analogy refer. There is an incomparably large body of research about these very principles from the Cushman/Hauser lab, the most influential of which was the major online study by Cushman et al. (Cushman et al. 2007).

Although I will confine myself to testing whether the mechanisms that produce judgement patterns along the lines of these principles are genetically inherited, my methodology is applicable to all candidate principles. One of the main competing theories is Michael Waldmann and Alex Wiegmann's "Double Contrast Theory" (Waldmann and Wiegmann 2010). This theory claims that the salient feature for many Trolley Dilemma judgement patterns is not the Means/Side-Effect distinction, but the harm that is done to the manipulated entity in contrast to the fate the manipulated entity would have had in absence of manipulation; hence, if you push someone from a bridge to stop a train that would otherwise hit five people, the person you would have pushed would have been unharmed and well if you had not pushed them. The harm you did by pushing them is worse than the fate this person would have had without your intervention and the action is rather impermissible. If you, however, redirect a train that is about to hit five persons onto a side track with only one person, you manipulate the train whose passengers stay unharmed in both cases; hence, the action is permissible. The explanation according to the Means/Side-Effect distinction would explain the same judgement pattern as in the bridge case: someone is harmed as a means to stop the train whereas in the sidetrack case someone is harmed as a foreseeable side-effect of redirecting the train on the side-track; the person on the track is not necessary to stop the train. Although Waldmann and Wiegmann empirically show that people seem to judge according to this principle, this does not invalidate the Means/Side-Effect distinction, as Cushman et al. 2007 present two dilemmas (Ned/Oscar case, see Mechanistic Modularity chapter) that differ along the Means/Side-Effect lines, but not along Waldmann and Wiegmann's Double Contrast lines. Subjects judge this dilemma pair unequally, too; hence, the Means/Side-Effect distinction is the best explanation for at least some of the empirical data, maybe in some cases in conjunction with Waldmann and Wiegmann's theory, which is why I will adhere to this.

In a large part of this thesis, I will discuss some properties that hint at a cognitive mechanism's development as an evolutionary adaptation. The properties I will discuss are Mechanistic Modularity, Functional Modularity, Universality, Similar Stages of Development and Development Despite Poverty of Stimulus.

I will not discuss whether Trolley Dilemma judgement mechanisms have a particular place in the brain, although localization is usually a classical feature that people cite for assuming that a mechanism is modular (Fodor 1983). I will stay on the computational level of analysis, hence I will not write about 'hardware'; in other words, this thesis will stay at the level of cognitive psychology and will not rely on neuroscientific data because proponents both of the Linguistic Analogy and of Evolutionary Psychology argue mainly on this level and I answer them; for a discussion of neuroscientific data with regard to the innateness of moral mechanisms, see (Jacob and Dupoux 2007) and (Prinz 2007a).

After setting out the reasons for my methodological decisions, I will now follow up with a rough overview of the argumentational structure of this thesis.

1.4 Argumentational structure and content

Many Evolutionary Psychologists believe that “the mind is modular”, and this assumption is closely related to their argumentation that many cognitive mechanisms are innate. To test their argumentation, I will first disentangle their concept of “module”. I will show, in chapter 2, p. 15, that Evolutionary Psychologists refer to two entirely different kinds of entities when they speak of “modules”. The first kind, which I will call “Mechanistic Modules”, are determined by their isolated informational processing, the second kind by their function, hence their name “Functional Modules”. This conceptual groundwork enables me to make explicit how the two different kinds of modularity, according to Evolutionary Psychologists, are related to an evolutionary development of said modules. After I have determined what properties they would need to have in order to be favored by evolution, I will resort to empirical data to show that the Trolley Dilemma judgment mechanism is unlikely to fulfil any of those properties, first for “Mechanistic Modules” and then for “Functional Modules”. The property that makes the latter evolutionarily eligible turns out to be, roughly, the fit of currently displayed behavior to ancestral problems, hence, how fit this behavior is to solve ancestral problems in a way that promotes reproduction of one's own genes. I will show that it is problematic to test this empirically for any social behavior, including Trolley Dilemma judgements, and will then proceed to empirically test other indications that a behavior has developed as an evolutionary adaptation: Its early development that follows a typical trajectory, its universal prevalence and the lack of sufficient input for learning it empirically. Finally, I will briefly sketch a counter-proposal as to how the three principles could have been acquired using inputs every learner is likely to receive by way of domain-general learning mechanisms (hence, learning mechanisms that are suitable to copy any behavioral pattern your environment displays), without this being an evolutionary adaptation. This way of learning avoids the issues raised by innateness proponents who deemed the input too sparse for empirical learning, and thereby shifts the burden of the argument: The argument says that the Trolley Dilemma judgement pattern cannot be learned empirically due to the sparse input and hence the learner must, at least partially, resort to domain-specific information. I will show that there is at least one possible way to acquire the typical judgement pattern using domain-general learning mechanisms, so why should the learner need innate information?

After I have demonstrated that none of the commonly cited arguments for an evolutionarily adapted Trolley Dilemma judgement mechanism hold, I come to the conclusion that it is very improbable that such a mechanism is at work in humans.

Now that I have sketched the argumentation structure, I will return to the contents of each section as well and examine them in more depth.

’The relation between Evolutionary Psychology and different kinds of modularity: ’Mechanistic Modularity’ and ’Functional Modularity’ – A differentiation’ is an introduction to the main section about Modularity. I will briefly introduce the paradigm of Evolutionary Psychology and sketch its relation to modularity.

In the ’Mechanistic Modularity and Evolutionary Adaptations’ chapter, I will present the structure of my argument: I will first show how, according to Evolutionary Psychologists, a module is likely to process information if it has developed as an evolutionary adaptation in the chapters about the Evolvability and the Practicality argument and then examine whether Trolley Dilemma judgement mechanisms process information accordingly.

In the chapter ’Evolvability argument (watchmaker argument) and the module’s properties’ chapter, I will elaborate how the ’Evolvability’ or ’Watchmaker argument’ – often cited but not explained by Evolutionary Psychologists – is applicable to cognitive mechanisms. It states that a module needs to successfully serve its evolutionary function to be selected for and hence no partial modules will develop stepwise over generations until they are functional because those partial modules are dead ends that will not be selected for. Rather, modules that are less complex, serve only one function, have developed in one step and are functional at the very moment they developed will be evolutionarily successful. It follows that such mechanisms compute in a relatively informationally secluded manner so they did not interfere with already existing mechanisms and disrupt their functioning when they first developed, and that evolutionarily newer mechanism may depend on the output of more ancient ones, but not the other way round.

In the chapter ’Practicality argument (specific breakdown pattern argument) and the module’s properties’, I will elaborate said argument, which is equally often cited. It states that evolutionarily selected modular mechanisms should process information relatively independently of each other, hence compute most information inside the module and only exchange sparse information between modules, and that they should not rely on many other modules' outputs so that the breakdown of only those few modules can impair their functioning. A system as a whole with such modules will be less vulnerable to damage in single modules as the other modules will continue to function. This makes it more evolutionarily 'electable'. I will then show in the chapter 'Are Trolley Dilemma Judgement Mechanisms Mechanistic Modules?' that Trolley Dilemma judgement mechanisms do not process information as the above arguments would predict, which weakens the case that they have developed as evolutionary adaptations.

There is, however, a second notion of evolutionarily advantageous modularity that might apply to Trolley Dilemma mechanisms: Functional Modularity. This notion does not concern the way a mechanism processes information, but instead the fit of adapted modular mechanisms to evolutionarily salient problems: Does some behavior answer evolutionary challenges in the Pleistocene? If so, it might have developed as a specialized mechanism and be selected for because it solved those challenges.

In 'Functional Modularity and Evolutionary Adaptations', I will give a very brief introduction to the coming sub-section on this topic and then go on to give a broad introduction to the paradigm of Evolutionary Psychology in the chapter 'Are Trolley Dilemma Judgement Mechanisms Functional Modules? - Evolutionary Psychologists' main evidence re-evaluated', in which I will also elaborate on how Mechanistic and Functional Modularity differ and proceed with ways to test a cognitive mechanism for Functional Modularity. Because 'being a functional module' basically means having developed as an adaptation, the entire section 'Are Trolley Dilemma Judgement Mechanisms Functional Modules? - Evolutionary Psychologists' main evidence re-evaluated' is dedicated to doing just this: Re-evaluating Evolutionary Psychologists' main evidence. I will first introduce the concept of 'Reconstructive Engineering', in which we consider an ancestral challenge and look for modern behaviors that might have been adaptations to this challenge; after this, I will introduce 'Reverse Engineering' in which we consider a complex of modern behavior and establish whether it is an adaptation to an ancestral challenge. I will show ways to empirically test hypotheses in both 'Reconstructive' and Reverse Engineering and will show that those ways are often problematic, especially in the realm of social cognition, and therefore not very apt for testing Trolley Dilemma judgement mechanisms.

But there are other ways of testing for innateness: Seeing whether the principles underlying Trolley Dilemma judgements have developed in a certain way that suggests they follow a 'pre-programmed' trajectory or whether they have developed very early in childhood, before they could have been learned; seeing whether they are universal and hence not contingent on education, and whether they could have been learned in the form people uphold them without any inherited domain-specific information that complements their environmental information. I will assess this in the chapters "Development", "Universality", and "Poverty of Stimulus". I will especially concentrate on Poverty of Stimulus as this argument is one of the most prominent ones in the discussion about the innateness of

Trolley Dilemma principles (and, by the way, an argument adapted from Chomsky by proponents of the Linguistic Analogy).

In the Poverty of Stimulus Chapter, I will first introduce the reader to Poverty of Stimulus arguments. Poverty of Stimulus arguments, roughly expressed, state that learners cannot have learned something on a solely empirical basis because the environmental input is not sufficient and hence they must be endowed with some specialized inherited mechanism that provided them with the missing information. I will then assess the relationship between being an adaptation, being genetically inherited and being an evolutionary psychological adaptation to show why the inheritedness of a mechanism in the Poverty of Stimulus sense can be evidence for the mechanism's evolutionary adaptedness in the sense of Evolutionary Psychology: Inheritability is a necessary precondition for something to be an evolutionary adaptation.

I will then go on to show how Trolley Dilemma principles are generative in analogy to linguistic generativity and which version of the Doctrine of Double-Effect can best be operationalized and tested in Trolley Dilemmas to indicate what the 'grammatical' system of Trolley Dilemma principles might look like and what its rules are.

This prepares the ground for the application of the Logical Problem of Language Acquisition to Trolley Dilemma principles. Here is a very simple version of this:

Without explicit instructions about grammar or negative evidence (as in: "What you said was grammatically wrong"), children will overgeneralize linguistic rules if they learn them empirically; hence they will never be able to learn the right kind of grammar. As they usually do not get those necessary kinds of feedback, they cannot have learnt language solely empirically. Most people, however, have learnt language; hence they must have had some inheritable information that helped them avoid overgeneralizations.

Some proponents of the Linguistic Analogy write that explicit instructions or negative evidence are necessary to learn Trolley Dilemma principles and hereby draw an analogy to the Logical Problem of Language acquisition.

I will answer several of the Linguistic Analogy proponents' arguments relating to the Logical Problem of Language Acquisition one by one, and show that most of these do not apply. Finally, I will conclude the thesis with a way of acquiring Trolley Dilemma principles without running into any of those problems. My account will be a language learning mechanism proposed by Steven Pinker and slightly extended by Fiona Cowie, applied to Trolley Dilemma principles and combined with Jesse Prinz's theory about moral sentimentalism.

2 The relation between Evolutionary Psychology and different kinds of modularity: ”Mechanistic Modularity” and ”Functional Modularity” – A differentiation

Many proponents of the Linguistic Analogy claim that the Trolley Dilemma judgement mechanism is modular. By claiming this, they support their thesis that the Trolley Dilemma judgement mechanism is genetically inheritable.

In the following section, I will show why Evolutionary Psychologists regard modular mechanisms as more likely to be genetically inheritable by demonstrating how modularity and Evolutionary Psychology are related, define modularity as is needed to make Evolutionary Psychological claims and assess whether empirical evidence supports such innateness claims.

The main claim of Evolutionary Psychology is that certain phenotypical traits have been advantageous in the Pleistocene environments either for the procreation of the individual showing them or for the procreation of their kin with similar genes. Hence, having those traits caused an easier transmission of the genes of the individual that showed them or similar genes by their relatives. But, unlike in 'classical' evolutionary theory, these traits are extended to the realm of cognition: [11] According to Evolutionary Psychology, there are genetic properties that cause the development of genetically inheritable cognitive features or abilities. Those cognitive features or abilities promote the survival of the individual in question or the survival of individuals that share similar genes (relatives of the individual).

One example many Evolutionary Psychologists use to illustrate this view is the fear of spiders or snakes (see, for instance, (Buss 1995, p. 8), based on (Marks 1987).) According to Evolutionary Psychology, the distribution of phobia shows that humans are much more prone to developing phobia against spiders than, for example, against cars. This seems to be the case even though in many contemporary environments the probability of having a car accident and thereby being severely harmed is far higher than the probability of being severely harmed by a poisonous spider (but see Grossi 2014 for counter-evidence). Advocates of Evolutionary Psychology assume that this is because humans who developed the cognitive feature of being afraid of spiders or the cognitive feature of easily [12] acquiring fear of snakes (Leda Cosmides and Tooby 1994 p. 106) had survival advantages in

[11] For the sake of simplicity, in this text I will skip the issue of whether and how the brain's structure is genetically determined, a structure that I assume cognitive mechanisms to supervene on. Instead, I will concentrate on the cognitive structure: If the rearrangement of genes leads to a change in cognitive structures, this is however probably so because brain structures have been changed. Therefore, I will henceforth discuss the evolution of cognitive mechanisms instead of discussing the evolution of a particular manner of brain organization that underlies cognitive mechanisms.

[12] As compared to other fears.

an ancestral environment which lead to a propagation of their set of genes, including those genes responsible for their fear of spiders or their easily acquiring fear of spiders. Dispositions to develop certain kinds of behavior (avoiding spiders/snakes) based on certain cognitive traits (fear of snakes/ability to easily acquire fear of snakes) promote survival fitness (or procreational fitness or procreational/survival fitness of individuals with similar sets of genes, relatives) and are therefore passed on to larger numbers of offspring as compared to genes for cognitive endowment that do not include the disposition to develop those kinds of behavior. In addition to these presuppositions, a great majority of upholders of Evolutionary Psychology believe in some or other version of massive modularity. [13] To believe in massive modularity is to believe that the mind is entirely composed of cognitive modules. Claims of Evolutionary Psychologists for the mind to be modular can be found in, but are not limited to, the following sources: (Barrett and Kurzban, 2006, p. 629), (Barrett and Kurzban, 2006, p. 630), (Carruthers, 2006, p. 18), (Cosmides and Tooby 1997, "The Modular Nature of Human Intelligence", p. 81) (Cosmides and Tooby, 1997, p. 85), (Leda Cosmides and Tooby, 1994, p. 105), (Pinker, 1997, p. 21), (Pinker, 1997, p. 27), (Sperber, 2002, p. 1), (Sperber, 2002, p. 3). For a more comprehensive list, including the authors' actual quotes, please refer to the appendix.

There is no logical or nomological (compare: (Gamez, 2009, p. 215)) *necessity* that connects this basic idea of Evolutionary Psychology with the idea that the mind is entirely composed of modules, i.e. that every cognitive task is computed in a modular fashion. However, thinking that massive modularity is true gives us, as I will elaborate in the next chapter, some reasons to believe that Evolutionary Psychology is true and this is why belief in Evolutionary Psychology and massive modularity co-occur in scientific practice so often.

'Modularity' is a highly heterogeneous term and has, depending on its realm of application, many notions. If Evolutionary Psychology is true, many cognitive mechanisms very likely have developed modularly in two senses. For those, we can claim a probabilistic nomological connection to being Evolutionary Psychologist adaptations: If Evolutionary Psychology is true and if certain laws that seem to underlie evolution, computation and physics are true, then it is very probable that two kinds of cognitive modules have developed in this world, and this is probable for several reasons. I will thoroughly define those two senses in the next two chapters on Mechanistic and Functional Modularity. Suffice here to say that Mechanistic Modularity is a modular way of computing information, related to Fodorean modules (Fodor, 1983), while Functional Modularity is the fit between a behavior and a Pleistocene survival/procreation problem, where the behavior clearly solves that ancestral problem.

[13] I will give a detailed account of the different relevant versions of modularity concepts at the beginning of the chapter about Functional Modularity. For now, I have assembled citations from some of the most-cited proponents of Evolutionary Psychology regarding modularity in the appendix to show the strong connection between Evolutionary Psychology and beliefs in modularity.

If a certain mechanism is found to be non-modular in both ways, it is also improbable for this mechanism to have developed according to the principles of Evolutionary Psychology because, according to many proponents of Evolutionary Psychology, if a cognitive mechanism has developed as an evolutionary adaptation, it is likely to develop either as a functional or as a mechanistic module, and hence being an evolutionary adaptation is correlated to being a module.

I will also show that the mechanisms that compute the principles underlying Trolley Dilemma judgements are not modular in one of the two senses mentioned above and thereby argue that they probably have not developed in the way upholders of Evolutionary Psychology suggest. Moreover, if you take problems that have a structure analogous to the structure of Trolley Dilemmas to be paradigmatic for moral problems, moral mechanisms probably have not evolved as an adaptation to a problem in an ancient environment. [14]

I have introduced the first kind of modularity as “Mechanistic Modularity”. The second kind of modularity is Functional Modularity, and I will define the former in the next chapter and the latter in the chapter after that.

2.1 Mechanistic Modularity and Evolutionary Adaptations

As I have mentioned above, Evolutionary Psychologists are often proponents of some kind of modularity, but they often do not exactly define which kind of modularity they are referring to and the concept is heterogeneous. The so-called ’evolvability argument’ and the ’specific breakdown argument’ do not refer to Functional Modularity as I will define in the next chapter because they clearly concern cognitive architecture and mechanisms and their informational and processing structure, similar to Fodorean Modules,(Fodor 1983); they do not concern, in essence, matches between problems in ancestral environments and behavior, as would be classical adapted modules in the Evolutionary Psychological sense.

However, many Evolutionary Psychologists use them as arguments about why the mind is supposedly massively modular. [15] If Evolutionary Psychologists are not referring to Functional Modularity when they use the term ”modularity” in those arguments, they are likely referring to some kind of mechanistic (processing) modularity. Unfortunately, I was not able to find those arguments spelled out in the literature. So, my first task was to spell them out myself. The aim of this chapter is to map out the connection between Mechanistic Modularity and the reasons Evolutionary Psychologists habitually give about why cognition is likely to be built modularly, but which cannot refer to functional. I will show that many of those reasons refer instead to information processing properties and hence to properties of Mechanistic Modularity. But what is the relation between

[14] Adaptations, according to Evolutionary Psychology, take so long to develop that human cognitive systems have adapted to an ancient hunter-gatherer environment as opposed to the different environments many people live in now which, in evolutionary terms, have existed very briefly

[15] For quotes, please see respective beginnings of the ’evolvability argument’ and ’specific breakdown argument’ sections.

mechanistically modular cognitive architecture (hence, the way information is processed) and a development of those modules as evolutionary adaptations? The main connection I have found in the literature were two arguments, the so-called "watchmaker argument" and what I termed, based on Sarva 2003 (Sarva, 2003, p. 225), the "practicality argument".

The "watchmaker argument" provides us with a phylogenetic story how our cognitive mechanisms evolved. According to this argument, complex cognitive systems are more likely to have developed modularly, one module at a time, because it is unlikely that a complex cognitive system developed all at once (the leap would be too big for one evolutionary step) und similarly unlikely that only unfinished components of that complex cognitive system developed with every evolutionary step, because those components would be useless as parts of an unfinished system and, therefore, their bearers would not have any evolutionary advantages. If, however, one module developed at a time, each module would more likely be propagated to the next generation because it would already have an additional adaptational advantage; it would already be "useful". The "practicality argument" is about information processing and specific breakdowns: It holds that if cognitive mechanisms are relatively isolated from each other (in terms of information exchange) then if one mechanism breaks down, the others are more likely to keep on working than if much information is exchanged between those mechanisms. The single mechanisms are more autarkic this way. This will prove to be an evolutionary advantage: If an individual gets injured or sick and one of the cognitive mechanisms fails, they will still be able to maintain the other modular cognitive functions and therefore have an evolutionary advantage over individuals who cannot maintain any cognitive functions after their cognition has been impaired because the whole system breaks down. The former will at least be able to maintain some of their evolutionarily advantageous behaviors while the latter probably will not be able to survive. I have dedicated a whole subchapter to these two arguments where I recapitulate these arguments in more detail [page numbers].

As it is common among Evolutionary Psychologists 'nowadays' not to uphold Jerry Fodor's very strict conception of modularity, [16] I have taken those arguments as a basis to find out what kind of assumptions they have to make about the computational architecture of the mind that goes with the assumption of massive modularity. I will show in which sense a mechanism should be mechanistically modular if it has developed as an evolutionary adaptation. The context prompting me to show that is the following:

1. Claim: The Trolley Dilemma judgement mechanism is modular and innate according to proponents of the Linguistic Analogy. If you find a cognitive mechanism to be mechanistically modular, you have one more reason to assume that a mechanism has developed as a module if you agree with the "watchmaker" and the "practicality"

[16] Including domain specificity, mandatoriness, limited central (top down) access, fastness, informational encapsulation, shallow inputs, specific breakdown patterns and characteristic development pace (Fodor 1983).

arguments.

2. To date, there has been only one prominent account of breaking Trolley Dilemma computation mechanisms down into smaller computational steps [17] (Mikhail 2007b, Mikhail 2011).

3. The mechanisms that process those computational steps (from 2.) are not only used for Trolley Dilemma judgements, but for different kinds of computations in completely different realms: among others, reasoning, grammar, comparison of numbers, causality.

4. Hence they are not mechanistically modular in the sense they should be if they have developed as evolutionary adaptations.

5. The proponents of the Linguistic Analogy can no longer use the Mechanistic Modularity of the Trolley Dilemma judgement mechanisms as arguments for an evolutionary development and, hence, inheritability of those mechanisms.

2.1.1 Evolvability argument (watchmaker argument) and the module's properties

In a nutshell, Mechanistic Modularity and Functional Modularity concern different properties of mechanisms. One and the same mechanism might be mechanistically and functionally modular at the same time, but it might also only be modular in one of the two senses. Mechanistic Modularity concerns the way in which information is computed. There are different notions of Mechanistic Modularity, and different authors within Evolutionary Psychology do not agree about which properties are necessary and sufficient for a mechanism to be modular (compare Seok 2006, Barrett 2006, Carruthers 2006), but the notion more or less revolves around the amount of information [18] the mechanism computes inside the module as compared to the amount of information that is computed between two or more of such mechanisms (compare the notion of "subsystem" in (Haugeland 1995, pp. 211–219)). I will first expound the reasons why Evolutionary Psychology (EP) is connected to Mechanistic Modularity and then present the notion of Mechanistic Modularity that is presupposed if this connection is to be maintained. There are two main reasons to be found in the literature why Mechanistic Modularity is connected to EP. One of them is the "evolvability" argument, and I will call the second one, based on (Sarva 2003 p. 225), the "practicality" argument.

[17] This account explains how people make Trolley Dilemma judgments that show patterns that conform with certain principles, e.g. the three principles mentioned in the introduction.

[18] As information is something of a loaded term in philosophy, I will commit myself and define its measure as Kolmogorov complexity (Adriaans 2013). However, the kind of information does not play a crucial role for the broad theory as I assume that the ratios between inside-module and outside-module information flow roughly converge for different kinds of information in most cases (although possibly not in all) and those ratios are the key issue for our theory, not the absolute amount of information.

In his paper "The case for massively modular models of mind" (Carruthers 2006, p. 8), in which Carruthers unsurprisingly makes a case for massively modular models of the mind, he briefly condenses both of the arguments, referring to (Simon 1962):

> "The first argument derives from Simon (1962), and concerns the design of complex systems quite generally, and in biology in particular. According to this line of thought, we should expect such systems to be constructed hierarchically out of dissociable subsystems, in such a way that the whole assembly could be built up gradually, adding subsystem to subsystem; and in such a way that the functionality of the whole should be buffered, to some extent, from damage to the parts."

The evolvability argument was first put forward by Simon, 1962, albeit in a broader context without a clear reference to mental structures (Simon 1962). (Sarva 2003, p. 225) puts it this way: "[Modular structures] are [...] more evolvable, since each structure is more simple than a single all-encompassing structure would be, they are closer in the reach of evolution." But why are 'simpler' structures 'closer in the reach of evolution' than more complex structures? One main reason is the occurrence of stable intermediate states. Simon describes this using the analogy of two watchmakers (Simon 1962, pp. 470–472).

One of them assembles watches by putting together a whole watch in one work step and arranges 1000 pieces of the watch to one big whole. Whenever he is forced to disrupt his work and lay the pieces down he has already assembled, he has to start over arranging the pieces because everything falls apart.

The second watchmaker, however, builds his (equally complex) watches by putting together parts consisting of 10 pieces each. After that, he pieces them together to building blocks, each with 10 parts consisting of 10 pieces. The 10 building blocks consisting of 10 parts consisting of 10 pieces are then arranged to build a complete watch. This watch consists of 1000 elementary pieces just as the other watchmaker's watches. But there is a difference to the building process. Whenever he is forced to lay down his work, the second watchmaker only loses whatever part of one assembly of 10 elements he has already put together. Instead of losing 999 steps of his work process in the worst case, he only loses 9 steps if everything goes wrong. Obviously, there are a lot more intermediary stable states to his partly assembled watches than to the ones (namely, zero) produced by the other watchmaker. [19]

Simon says:

> "The model assumes that parts are entering the volume at a constant rate, but that there is a constant probability, *p*, that the part will be dispersed before another is added, unless the assembly reaches a stable state." (Simon 1962, p. 471).

[19] The last two paragraphs are based on (Simon, 1962)

But how can this model be applied to biological organisms? Simon applies it to the evolution of cells from molecules and of multi-celled organisms from single-celled organisms.

"The time required for the evolution of a complex form from simple elements depends critically on the numbers and distribution of potential intermediate stable states." (Simon 1962 p. 471). Assemblies of simpler organisms on one (hierarchical) level are more likely to develop (or develop more quickly) than bigger, more complex mechanisms that could do similar things as an assembly of simpler mechanisms, but would need to evolve at once.

And what can be applied to those biological forms might also be applicable to cognitive mechanisms. Imagine that a new cognitive mechanism is about to evolve: A cognitive mechanism has developed in a living organism (e.g. a human being) that, due to divergent genetic endowment, differs from every cognitive mechanism the creature's ancestors were equipped with. This mechanism has developed by chance, meaning that it has not developed to serve any particular purposes and its features are more or less random.

Now, there are different ways for a new cognitive mechanism to evolve.

A) The mechanism might evolve as an extension of an already existing mechanism, making it more complex (imagine a mechanism with some randomly added features). It would look like this:

   (o), where ( ) is the already existing mechanism and o is its internal add-on.

B) It might evolve as piece of a more complex mechanism 'yet to come' and in addition to already existing mechanisms, but not intermingling with them in terms of computation: It would be part of a more complex mechanism that it might evolve into, assembled with other parts, in future generations. Imagine part of a mechanism that is not functional yet (does not yield any procreational advantages) but might become so if more pieces are added: It would look like this:

   ( )(, where ( ) is the already existing mechanism and (is its incomplete external add-on. To individuate this more complex mechanism as something that is 'complete' could be done in terms of its function: It is complete whenever it leads to the organism equipped with it, or the organisms equipped with genes that make it more probable for their offspring to be equipped with it, to have procreational (including survival) advantages. When the second mechanism would be complete, both mechanisms would look like this: ( ) ( ) and both would be able to promote manifold and complex behaviors that promote survival and procreation.

C) It might evolve as one of those complete mechanisms ( ) described in the last paragraph that have developed in a way leading them to promote the computation of highly diverse survival and procreational 'tasks'. But, unlike the mechanism above, it would have developed in 'one step', not evolving part-by-part over several generations but occurring as a complete, 'functioning' mechanism in one individual:

The parental mechanism would have ( ) and the offspring in the next generation ( ) ( ).

D) It might evolve as a more or less computationally independent subsystem (module) that is able to compute something that leads to the organism carrying it being more easily able to transmit its genes. It would be less complex than what I illustrated as ( ) and would promote a simpler and narrower bandwidth of behaviors. Assembled with other such modules, it could be a building block of a more complex mechanism that can be 'applied' to even more complex survival or procreation problems (parallel to single cells that would have developed more easily and more quickly than an entire organism, are more or less independently functioning subsystems and can, if assembled, build up more complex organisms that are able to carry out far more complex tasks such as walking or even thinking as opposed to e.g. the metabolism a single cell can perform). The new mechanism would look like this:

o.

And the entire system would look like:

ooooo.

This system, as you can see, consists of assemblies of smaller, less complex mechanisms that can only compute simpler tasks each (or solve one evolutionary problem in a way that promotes procreation of the organism which has it). The assembly ooooo would be able to fulfil similarly complex tasks as ( )( ).

There are several reasons connected to the watchmaker story why the mechanism in Possibility D is the one with the biggest chances to evolve and 'stick around', being the basis for a more complex cognitive mechanism to develop as an assembly of modular subsystems. However, we need to make a presupposition if this line of argument is to be sound. The following assumption holds for all four possibilities of developing a new mechanism as listed above: Cognitive mechanisms responsible for solving certain tasks can be passed on by way of genetic inheritance. If this was not the case, there would be no point in discussing any property of a cognitive mechanism which might influence the chances for procreation, because it would die out after the next generation. As this presupposition is one of the core assumptions of Evolutionary Psychology, it surely makes sense to accept it in this context, as the goal of this chapter is to argue what kind of mechanisms a theory about a biological evolution of those mechanisms would predict.

I will consider Possibility A first: There is an already existing cognitive mechanism that somehow promotes the transmission of the genes underlying its development, e.g. a cognitive mechanism that leads to the individual carrying it to be afraid of spiders and thereby promoting their chances of survival/procreation. Now this individual's offspring is equipped with the same mechanism, but the mechanism is somehow changed more or less randomly as the genes underlying this mechanism or the new part of this mechanism are ‘scrambled’ in a non-targeted fashion, i.e. not ordered in any way as to produce a mechanism that is able to fulfil a particular function.

There is a chance that the original cognitive mechanism performing a ‘function’ (being afraid of spiders) will be disturbed or will not perform the ‘task’ anymore at all:

As there are far more non-functional random structures than functional random structures and mechanisms can be easily disturbed by changes, it is more probable for the organism carrying the new mechanism to die or not procreate or promote a kin’s procreation than it was for their ancestor carrying the undisturbed mechanism. (Peck 1994)

This means it is less probable for the genes giving rise to an altered, originally functional mechanism to be passed on than for genes leaving the mechanism unaltered. Or, as the programmer would put it: In most cases, it would be wiser not to change an already functioning subroutine but to add a new one to implement a new function.

Expressed in terms of the watchmaker analogy, individuals carrying an altered original mechanism are analog to an unstable state: They tend to dissolve before they (including their cognitive mechanism) can be further altered (in a coming generation).

This leads, in terms of modularity, to the assumption that mechanisms that exchange great amounts of information with already existing mechanisms or alter the computation process of the already existing mechanisms are not very probable to evolve biologically.

Is the same true for Possibility B? If the new mechanism is something that might be used as part of a bigger or more complex mechanism that is yet to develop (by chance, of course) in the future and that might lead to strong advantages concerning the organism’s procreation or procreation of its kin, this could be, put very simply, a first step in the right direction. The right direction in this case would point towards a ’complete’ mechanism that would give rise to behavior fit to promote the procreation of the individual carrying the mechanism, or the individual’s relatives, and hence the distribution of the genes that underlie the mechanism. It would, for instance, be a direction towards a mechanism that leads to the organism being excellent (or at least better than organisms without the mechanism) at avoiding some death cause that is common in the individual’s environment or doubling its rate of producing offspring or something similar. However, this part of a potentially very useful mechanism (in terms of promoting the genes to be passed on) would unlikely be propagated widely because the organism carrying it will probably not be any more successful at solving the prospective task than any organism that does not carry the part of the mechanism: The part *as it is* does not enable the organism carrying it to be excellent at avoiding common death causes etc. because it is only a part, and a non-functional one, of a ’complete’ mechanism.

Given the presupposition that mechanisms that are only parts of bigger mechanisms don't serve any purpose and hence, according to our definition, are 'incomplete' and given that mechanisms that don't serve any purpose lead to the organism's consuming just as much or more energy than an organism that does not carry those mechanisms, they are evolutionarily disadvantageous 'at worst' and useless 'at best' (procreationally speaking) and hence do not lead to any evolutionary advantage.

This makes the state of the individual carrying the mechanism either *less stable* or *just as stable* as their ancestors' state: It is either less probable or equally probable for the individual [20] equipped with the genes giving rise to this mechanism to pass them on as compared to an otherwise equally equipped individual who does not carry those genes.

Possibility C in contrast could lead to an evolutionary advantage. In this case, chance operates in a way that the genes for the complex ('multi-purpose') mechanism that the mechanism in Possibility C would have been "the first step to" are assembled all at once in the 'right' way, hence, giving rise to a 'functional' mechanism, in one individual: A random mutation has given rise to a 'complete', complex mechanism. This would lead to enormous fitness advantages. The *random* evolution of such a complex mechanism however (e.g. one that would promote fear of snakes and also lead to a better ability to predict how objects move) is fairly improbable in the light of evolution.

This, again, is only valid on the basis of a few presuppositions: Firstly, that cognitive systems that perform complex (manifold) tasks are more complex. Secondly, that more complex cognitive systems need a bigger set of genes to be arranged in a particular way so as to be able to perform complex functions. Thirdly, that there are fewer gene conformations that give rise to cognitive mechanisms that are evolutionarily advantageous than there are genetic conformations that do not lead to any evolutionary advantage. [21]

In this case, if the watchmaker analogy is applied, the interruption would set in during the gestation process when genes reassemble to form the 'new' individual's genes: It is very improbable for the watchmaker to accidentally assemble the 1000 building blocks of the watch in a way that it builds a watch or any other functioning device by throwing them in a bucket and shaking it. [22]

To sum this up, if the presuppositions listed above hold, it is very improbable for a complex mechanism of this kind to evolve. [23]

[20] I use 'organism' and 'individual' interchangeably here.

[21] If this seems implausible to you, please bear in mind that I am merely reconstructing an argument referred to by Evolutionary Psychologists in the most beneficial form I can devise.

[22] And given the presupposition from Possibility 2, that individuals equipped with 'useless' mechanisms consume just as much or more energy, this mechanism is not only very unlikely to come into existence: If there were genes leading such complex mechanisms to evolve randomly, the nonfunctional 'trials' would not only not promote the transmission of the genes giving rise to them by e.g. evoking fear of spiders and enhancing object recognition skills, but also lead to the individual carrying them to have an equal or worse energy balance (also standing in the way of spreading the genes); so 'meta-genes' leading to genes assembling in a way they would randomly build up very complex systems would also be likely to die out. This argument, however, involves the acceptance of so many presuppositions I will not elaborate on it.

[23] If, however, something like this kind of mechanism would evolve by chance, this mechanism's computation would still be compatible with cognitive mechanistic modularity; the module that would

A mechanism as sketched in Possibility D, however, is far more likely to evolve in the light of evolution: In the case that a modular mechanism serving one function only and informationally encapsulated to a certain degree (does not interfere with already existing mechanisms) evolves, this mechanism would not be subject to all the other mechanisms' stability problems: All discussion of functionality in the following section relates to the aptness that may lead to the spreading of the genes underlying the respective mechanism. Agreeing with all presuppositions above, the mechanism in Possibility D would not interfere with a mechanism that was already existent in the individual's ancestors, and would therefore not possibly render the existent mechanism non-functional as does the mechanism in Possibility A; it would not be a useless burden to the system as is the partial and therefore not functioning mechanism in Possibility B; and there is a far larger possibility for it to develop than the very complex mechanism in Possibility C.

Put very simplistically, replacing evolutionary explanations by teleological explanations: What was the watch in our story are now properly functioning cognitive mechanisms able to solve manifold tasks.

We have already dismissed the alternatives of disturbing already existing functioning systems by extending them randomly (A) and of building one complex system responsible for everything by randomly assembling useless fragment after useless fragment (B), whereby it is improbable for evolution to not be disrupted while building it. In the latter case, the mechanism evolves piece by piece over several generations, in which case interruption means that every individual carrying genes for some piece of the mechanism dies or is unable to procreate. We have also dismissed Case C where an entire complex mechanism is assembled at once, because it is likely for randomness to set in in a way that the genes for the complex mechanism are not assembled all at once. Here, the interruption would set in during the gestation process when genes reassemble to form the 'new' individual's genes).

Instead, we concluded that if we follow the watchmaker argument, in all likeliness a complex mechanism (possibly the mechanism underlying all cognitive processes) evolves as result of a process of adding one functional building block after the other. If there are *functioning* subsystems (modules) instead of unstable intermediary states, the potential interruptions in the phylogenetic development of cognitive mechanisms, namely the death of an individual before they were able to procreate or promote a kin's procreation, become less probable. This process allows for 'dead ends' simply to die out: If a more complex mechanism (the watch, or, in this case, the brain) is assembled from pieces that function more or less independently of one another, it will not even matter if the new one does not function too well; even if individuals with the new, additional, non-functioning mechanism

eventually evolve would just be more complex than the ones that could be paralleled to cells or parts of watches and if it was able to "solve" various evolutionary "problems" this would make it less compatible with the notion of functional modularity.

die out, there are still the ones equipped with the mechanisms that have developed before the new, useless additional mutations. Hence, given that the presuppositions above hold, it is more probable for an organism to develop a simple, evolutionarily new cognitive mechanism than to develop a complex new cognitive mechanism. Given that it is necessary for those simple cognitive mechanisms to compute information in a mechanistically modular form to more reliably evolve, and given that being a complete cognitive mechanism serving a particular new function leads to evolutionary advantages, it is more probable for mechanistically modular cognitive mechanisms to have evolved in a trial-and-error, hence evolutionary, fashion than for any kind of other mechanism. What kind of modularity is needed for those kinds of stability?

The mechanism must not interfere with other already existing mechanisms that serve an evolutionary function:The outputs of the new modules should not change the existing systems' input in a way that disturbs or destroys their functionality. The established systems should also not eliminate any information from the new system while computing. Expressed in Fodorean terms this means: The phylogenetically older systems' input systems should be ***informationally encapsulated*** *as* ***related to the newly evolved system***: The question whether mechanisms are informationally encapsulated is, according to Fodor (Fodor 1983, p. 72), the question"[...] what access they have to information that is available to other systems." Or, put differently, as property of the newly evolved system: Its processing and information should be ***inaccessible*** to the phylogenetically older systems. If the system is to promote the transmission of the genes that underlie its development, it has also, by definition, to fulfil the criterion of ***Functional Modularity***, which I will further elaborate in Chapter 2.3, p. 56.

This connection between Evolutionary Psychology and Mechanistic Modularity is one of the weaker links as there are many presuppositions to be made in order for the watchmaker analogy to hold; nonetheless, it is widely cited. To name just a few, see (Sarva 2003, p. 225); (Carruthers 2006, p. 8). Sperber (Sperber 2002, p. 3) writes:

> "'To quote a theoretical biologist, "The fact that the morphological phenotype can be decomposed into basic organizational units, the homologues of comparative anatomy, has [...] been explained in terms of modularity. [...] The biological significance of these semi-autonomous units is their possible role as adaptive 'building blocks'." (Wagner 1995)";

Claudia Lorena Garcia (Garcia 2007, p. 63) writes:

> "'In a nutshell, the argument concludes that a mind that is massively composed of cognitive mechanisms that are cognitively modular (henceforth, c-modular) is more evolvable than a mind that is not c-modular (or that is scarcely c-modular), since a cognitive mechanism that is c-modular is likely to be biologically modular (henceforth, b-modular), and b-modular characters are

> more evolvable (e.g., Sperber 2002; Carruthers 2005). In evolutionary biology, the evolvability of a character in an organism is understood as the "organism's capacity to facilitate the generation of non-lethal selectable phenotypic variation from random mutation" with respect to that character (Gerhart and Kirschner 2003)."

There is, however, the even more powerful argument of 'practicality' to connect mechanistic modularity to Evolutionary Psychology, which I will elaborate on in the following paragraph.

To sum up: A cognitive apparatus that has developed evolutionarily by random mutations is likely to be composed of relatively autonomous mechanisms (let us call them mechanistic modules; they have the main attribute that they are informationally encapsulated (Seok, 2006)). I have come to this conclusion by analyzing the 'evolvability argument' for modularity put forward by many evolutionary psychologists. This states that a modular cognitive architecture would be evolutionarily advantageous because it is unlikely for complex cognitive mechanisms to develop all at once. Instead, most complex systems consist of aggregated simpler systems like modules. As I have argued above, this can be explained as follows: If mechanisms that already function are interfered with by changing them randomly, they are prone to produce evolutionarily dysfunctional computations. The organisms carrying the genes for those mechanisms then will be fairly unsuccessful in reproducing or in aiding kin to reproduce. If, however, new mechanisms are added in the course of evolution without changing the already functional existing ones, the organisms carrying them will still have the same evolutionarily functional mechanisms they have inherited from their forebears. And they might, additionally, be able to solve more problems than their ancestors who lacked those new mechanisms. This means that they are more likely to survive as compared to the organisms that carry the randomly changed mechanism (apart from the possibility of a slightly higher energy consumption that may be caused by the additional mechanism) because they are able to solve at least as many problems as their ancestors. Based on the 'evolvability argument', the following computational architecture seems to be optimal in terms of evolutionary development: Phylogenetically more ancient mechanisms do not draw much information from newer mechanisms, because they already existed when the new mechanism was added. In more technical terms: Older mechanisms are informationally encapsulated or cognitively impenetrable towards newer mechanisms. As Ron McClamrock puts it: "[...] there may be constraints on what information can get into the module and influence its working". (McClamrock 2006, p. 1) The newly evolved mechanism's database is likely to be inaccessible to the phylogenetically older systems.

Ancient mechanisms are also more likely to be independent of newer mechanisms' *output* and more ancient mechanisms probably do not retrieve much information from more recently developed mechanisms while the latter are processing (again, because they were already 'complete' when the new mechanism joined).

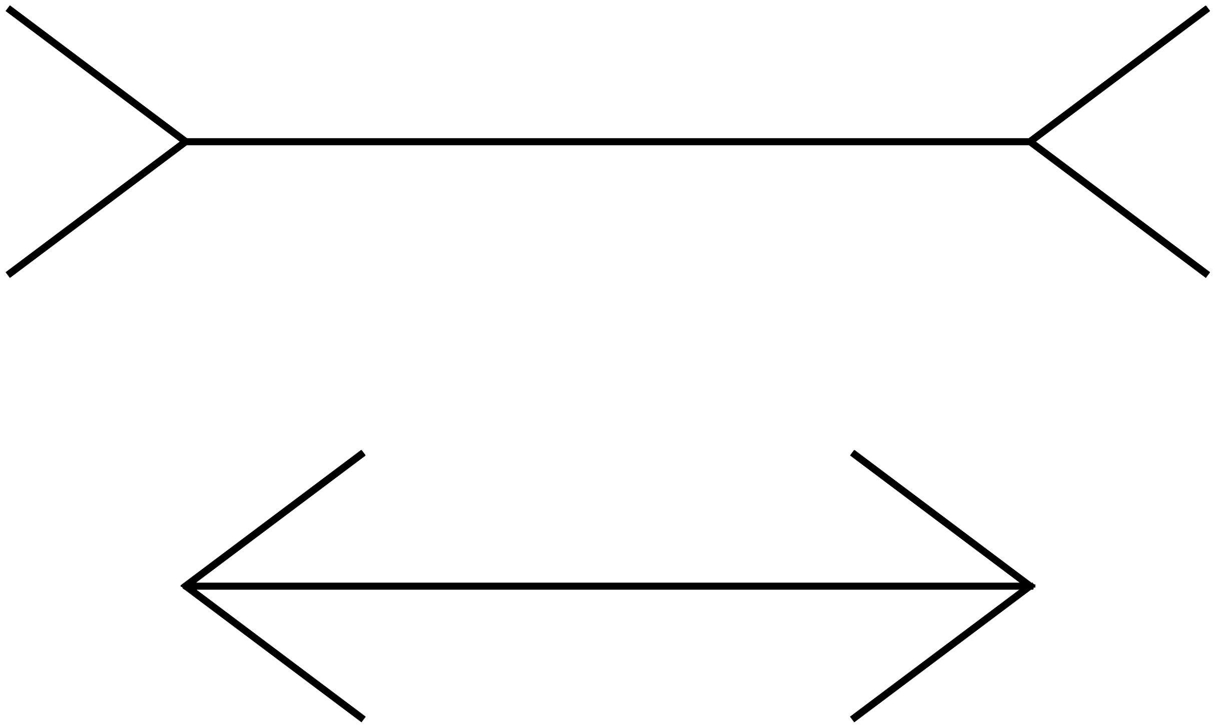


Figure 1: Müller-Lyer Illusion' (Müller-Lyer, 1889)

One classical example that has traditionally been used as evidence for informational encapsulation (Fodor 1983, p. 66) is the so-called 'Müller-Lyer Illusion' (Müller-Lyer 1889): Lines that are framed by arrows appear shorter if the arrows point outwards than lines that are framed by arrows that point inwards. Even if the tested persons know (e.g. because they have measured them) that two arrow-framed lines are the same length, they still perceive the one framed by outward-pointing arrows as shorter than the one framed by inward-pointing arrows. One possible explanation for this phenomenon is that the modules computing the visual information do not have access to the database or outputs (depending on where the acquired knowledge about the length of the lines is stored, whether it is 'incorporated' into the database or whether it is an output) of the modules computing the knowledge about the length. That is why their output - the perception of the length - stays unchanged despite the additional information – the knowledge about the actual length (Fodor 1983, p. 66), (Müller-Lyer 1889).

Now this independence of one module on the other, according to our 'evolvability', holds only for mechanisms in reverse order of development: An architecture where more ancient mechanisms draw little information from more recently developed mechanisms is evolutionarily more likely to persist and stay in the gene pool. Therefore, it might well be that all mechanisms work sequentially from ancient to newly developed and every mechanism draws its information and input only from more ancient mechanisms. In that case, every mechanism's processing would also be influenced solely by the respective mechanisms that are older than the particular mechanism is.

The 'specific breakdown pattern argument' or 'practicality argument' in turn leads to more general restrictions.

2.1.2 Practicality argument (specific breakdown pattern argument) and the module's properties

The next argument linking evolutionary development of cognitive functions to a modular cognitive architecture revolves around specific (cognitive) breakdowns. The "practicality argument", as I have named it based on Amol R. Sarva (Sarva 2003, p. 225), is spelled out by him in the following way:

> "'Modular structures are more practical because they are less dependent on the operation of *all* an agents [sic!] other capacities; disabling one part of a highly modular collection of capacities will not disable the rest."

Peter Carruthers (Carruthers 2006, p. 8) brings up a similar argument: "[...] the functionality of the whole should be buffered, to some extent, from damage to the parts."

The "practicality argument", reformulated, says that in the event that some cognitive mechanism is damaged, the less other structures depend on that mechanism, the less their functioning is impaired by the damage.

It is difficult to individuate such a mechanism and its function without having recourse to its architecture.

Because the way that such mechanisms are built up is the feature to be examined in this chapter, I will instead tentatively individuate single mechanisms by their function. Their function in this case is their evolutionary function: The function they have, according to Evolutionary Psychology, developed as a result of selection. This function has been selected for because it has proved useful in some way for promoting the genes it is based on. This only holds if the mechanism reacts advantageously to situations that have reliably been arising in ancestral environments. The timespan, according to Evolutionary Psychology, would not have been sufficient for genetic selection for mechanisms that solve recent problems arising in modern environments. Before I proceed to the conclusions we can draw from the Practicality Argument, I will make some theoretical remarks to introduce you to the way mental information processing has often been thought about in the tradition of Cognitive Psychology and Philosophy.

To illustrate the following argument, and in the tradition of Computational Theory, I will partially rely on an analogy, treating a cognitive mechanism as a mechanism that sequentially computes information, similar to a procedural computer program.

The rationale behind Computational Theory is the following: We are not mainly interested in brain anatomy, but in cognition, hence, what environmental stimuli result in which behavioral outcome and why, hence, inputs and outputs and how the processing in between happens. According to David Marr (Marr, 1982) there are three levels of computation: The functional, the algorithm and the implementation level. Function in this context means an operation that maps exactly one input on exactly one output (as in environmental input and behavioral output), e.g. f(x) = y. One function can be executed by infinitely many algorithms, the function f(x) = 2*x for example can be executed by the algorithm f(x) = 2*x+7-7, f(x) = 2*x+0*y, etc. One algorithm can be implemented in different ways, with implementation meaning the way a process is actually 'built in' matters; an algorithm may be implementable in a Turing machine, by electrical wiring, by a water circuit and by other devices as brains. As all finite computations can be implemented by a Turing machine and environmental stimuli and behavioral outcomes can be formulated in the form of such computations, too, it can be useful to identify the function and the algorithm underlying cognitive (mental) operations first and then proceed to investigate the implementation level, hence, to not talk about the brain, but about inputs, outputs and information processing (the algorithm) first (see also, e.g., (Thagard, 2012, chap. 3)).

In our case, the input (which can be either (transduced [24]) sensory information (Fodor, 1983, p. 45) or the output of a different module or the output of another, non-modular computational process) is processed by the 'program' which is either hardwired or also changeable by certain inputs (similar to a von Neumann architecture). That way, the input is mapped on an output. One cognitive structure [25] can depend on another cognitive structure by either

- depending on a certain output of the other structure or
- depending on information at an intermediary temporal stage of the computational process. This information can, depending on the notion of information, be contained in

    a) some databank the structure operates on,

    b) some programming sequence of the structure or

    c) an intermediary stage of the input data (at some point of them being processed, hence after and before some processing stages).

Now let us move forward with the practicality argument. A damaged cognitive mechanism might also provide information, but this information is probably not going to be correct. The relevant criterion for correctness in this argument is whether the mechanism provides information in a way that is advantageous for the promotion of the genes that have generated the mechanism.

[24] Changed in a way so that it can be handled by the cognitive device. For instance, optical information (or, formulated in broader physical terms, information about beams/photons) is changed in a way to make it computable for cognitive mechanisms (hence into a signal understandable by whatever they are implemented in, e.g. electrical current).

[25] By structure, I mean the following here: A correctly functioning mechanism, hence a mechanism whose computations yield genetically advantageous outputs.

An existing mechanism has, according to Evolutionary Psychology, likely evolved to process information in a way that the behavioral output under usual environmental conditions (that have been stable over a long period of time) mostly leads to the promotion of those genes. [26] If a mechanism like that is changed (become modified or 'damaged', if I may anticipate as this implies former functionality), it is very improbable for it to have more advantageous outputs 'survival-wise' than it had before. The mechanism it was before has been evolutionarily selected for. If such a mechanism is changed by some random influence ('damaged'), an evolutionarily advantageous path is disturbed in a random way.

Opposed to that, an (undamaged) mechanism that has been selected for has been selected out of random mutations in terms of its being evolutionarily more *advantageous*. It has been selected for out of many possible random mechanisms as one that promotes the genes it is based on. The deficient mechanism is more [27] like one of those randomly generated, often non-functioning mechanisms from among which the advantageous mechanism has been selected.

So, whatever is deficient about this mechanism, be it its database or its program or the way it is connected to other mechanisms, the output information that is retrieved from it by another mechanism is probably not advantageous in terms of propagation of the genes that produced it. And as all of those components may be damaged, we can generalize that not only output information but any kind of information retrieved from a deficient mechanism can be deficient. There are some ways that the wrong kind of information could lead to evolutionarily useful information when computed by a second mechanism afterwards: As mentioned before, the information could be deficient but still, by chance, yield an evolutionarily advantageous result. The mechanism could have received a deficient input which results in its own output being correct as the two deficits cancel each other out; the same holds for the mechanism in the 'computing chain' after it: The deficient mechanism could provide incorrect input information for the next mechanism in the computation chain. This sequential mechanism could also be faulty in a way it cancels out the deficits of the mechanisms before. Either way, two deficits that cancel each other out in this way seem to be highly improbable.

What are, then, the conditions for the 'practicality argument' to apply? If other mechanisms should be shielded from the deficient mechanism in order for them to still work, they should not retrieve any information from the deficient mechanism. If the other mechanisms should not be impaired by the loss or deficient functioning of one module, they must either not be dependent on other modules' outputs or information from intermediary

[26] For suggestions how information processing structure may depend on the genotype and, contingent on the environmental input, lead to pathways that promote evolutionarily advantageous behaviors, see Barrett 2012, p. 10734.

[27] Except for it formerly having been an advantageous mechanism that now has been changed, which makes it not completely random. This could be a factor that makes it more likely for it still to be at least partially evolutionarily advantageous.

processing stages or be able to bypass or repair this kind of information. A mechanism is independent of another mechanism's information if it does not necessarily need the outputs or intermediary stage information or program/database information of the other, potentially defective mechanisms in order to work correctly.

If we extend this conclusion to all cognitive mechanisms in an organism, as many mechanisms as possible should compute more or less independently of the other mechanisms or at least they should be able to work independently. Hence, a sequential set-up of *all* mechanisms as well as recursive computation between many different mechanisms becomes evolutionarily very improbable. Either the mechanisms each directly map a sensory input to a motor output or their outputs are brought together by one or more central mechanisms. We get a picture very similar to the "pipelike" structure H. Clark Barrett and Robert Kurzban talk about in their article "Modularity in Cognition: Framing the Debate" (Barrett and Kurzban, 2006, p. 631) and Barrett in "Enzymatic computation and Cognitive Modularity".(Barrett 2005, p. 265, see figure 1, p. 28). By means of this picture, Barrett illustrates the view of modularity he suggests Jerry Fodor has held originally:

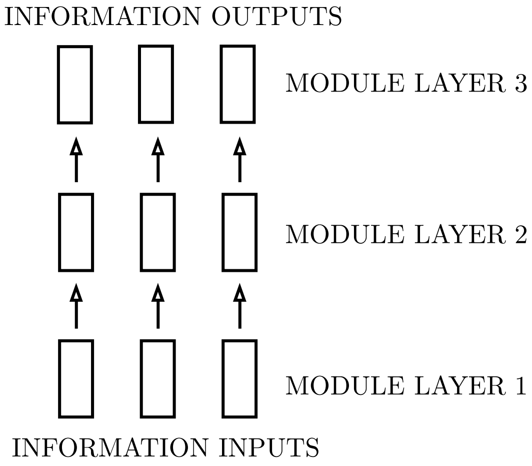
"This is the kind of architecture that Fodor (1983) originally proposed for input systems (and other peripheral systems, such as motor systems). Such architectures are composed of 'vertical' systems, or faculties. In vertical systems, information flows one way, in a bottom-up fashion; modules are arranged in layers, such that once information enters a device in one layer, it cannot subsequently enter another device in the same layer; and information is routed from one device to another. Consequently, different modules have access to different pools or sources of information. As Fodor puts it, they are nonoverlapping." (Barrett 2005, p. 264).

The less layers such an architecture has, the smaller the number of cognitive mechanisms that are dependent on each other. [28] The picture that presents itself is one of more or less functionally isolated or easily detachable cognitive mechanisms. This means that either mechanisms do not depend on each other's outputs and do not retrieve information from intermediary processes or each other's databases or that they can easily cease to do so if a mechanism is defective. This allows for one more advantage that is strongly interrelated with the 'practicality argument' insofar as it rests on the assumption that the rest of the functions is not disrupted by a non-functioning mechanism. Carliss Y. Baldwin and Kim B. Clark state it as follows: "Finally, modularity *in the design* of a complex system allows modules to be changed and improved over time without undercutting

[28] The same mode of argumentation can also be applied if we imagine cognitive mechanisms not as sequentially computing mechanisms but as mechanisms that communicate and compute via networks and compute via parallel activation. The more between-region (between-mechanism) connectivity a network has, the more one mechanism depends on the other mechanism. This, if we apply the practicality argument, then makes networks with clusters of more within-region connections more probable in evolutionary terms as compared to between-region-connections. A large ratio between within-region and outside-region connectivity is the definition of modularity in network theory (Barrett 2012, p. 10737). In this kind of paradigm, modularity can come in degrees.

the functionality of the system as a whole." (Modularity in the Design of Complex Engineering Systems", (Baldwin and Clark 2006, p. 180; emphasis in the original). So, a change of one cognitive mechanism (or, in Baldwin and Clark's case, system), even if evolutionarily disadvantageous, does not mean that the whole system (or organism) is going to collapse at once, which allows for new or changed mechanisms to build up step by step until an evolutionarily more advantageous state is reached. The organisms carrying the evolutionarily disadvantageous mechanisms, however, are per definition less likely to promote their genetic endowment which makes a further development (over following generations) of an evolutionarily advantageous mechanism out of the more disadvantageous mechanism less probable.

If we take the practicality argument from Evolutionary Psychology seriously, it follows that cognitive mechanisms are not only informationally encapsulated, which is defined by Jerry A. Fodor in the following way: "[...] one way a system can be autonomous is by being encapsulated, by not having access to facts that other systems know about." (Fodor, 1983, p. 73), hence by not being able to recourse to other systems' databases and processing sequences.



Vertical pipe architecture
Figure 2: Barrett 2005, p. 265

The mechanisms do not gain access to other mechanism's "knowledge" either, they are impenetrable to other systems: "The penetrability of a system is, by definition, its susceptibility to top-down effects at stages prior to its production of output." (Fodor, 1983, p. 74). In Figure 2, p. 33, penetrability is visualized when mechanism 2's intermediary input stage (hence, an input that it receives in the middle of its processing) is changed by mechanism 1's output. This mechanism, however, would have to compute at the same time as mechanism 2 as it would have to produce some output that changes mechanism 2's intermediary input. If we take Fodor strictly at his word, mechanism 1 would then be parallel to module 3 as shown in Figure 1, p. 28 (the illustration of parallel pipe-like structures), if there was an arrow back to module 2 whose intermediary input it would change: For the penetration to be a top-down effect, there needs to be a (here: sequential computation) hierarchy in the first place. And, finally, the inputs of many systems do not depend on other systems' outputs. Briefly summed up, the argument goes as follows: It would be evolutionarily advantageous for an organism (or, rather, for the promotion of its genes) if its entire cognitive system was able to serve its function [29] even when one of its parts is defective. This means that generally all mechanisms that function properly should be largely independent of all mechanisms that do not function at a point in time. If a mechanism 'breaks', the other ones can still function and help the organism to (further) procreate or help their kin procreate despite the outage of one mechanism. This would be fulfilled under the following conditions:

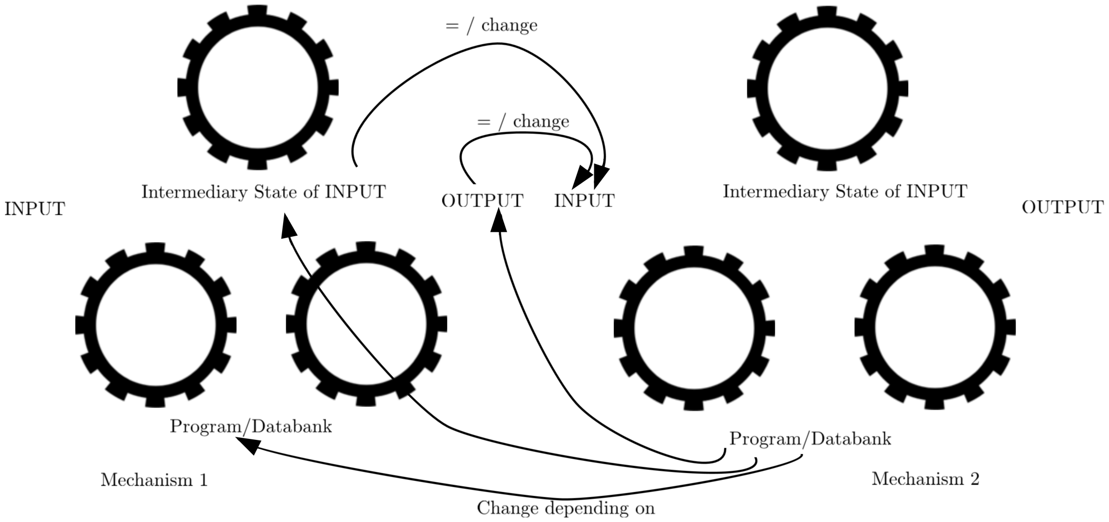


Figure 3: Possible Interactions between Cognitive Mechanisms

Firstly, mechanisms that serve particular functions (and that are individuated by the functions they serve) would optimally be highly connected within themselves, but not so tightly connected to other mechanisms. Their inner bandwidth should be broader than their outer bandwidth. This means that the amount of information which flows within one functionally individuated mechanism is bigger than the amount of information which flows between those respective mechanisms: The entire cognitive system should have a modular structure so that a cognitive defect only destroys one mechanism and only one function can no longer be fulfilled. The extreme version of this independence would be functional mechanisms that have inputs and outputs but do their computational processes completely without recourse to other mechanisms' databases, program sequences or outputs: They

[29] Which I have defined before as an evolutionary function: The promotion of the genes underlying its architecture.

depend on inputs and produce outputs but are otherwise secluded from other mechanisms and only use information internal to themselves while they are computing. Secondly, those mechanisms should be highly independent of other mechanisms' information as inputs, too, so that even if one mechanism fails to function, the other mechanisms should not break down due to a lack of required input from the other mechanism. This would hold in an architecture where all mechanisms' computations are generally largely independent of all other mechanisms' outputs.

One version where all mechanisms function mostly independently of each other would be a case where mechanisms go straight from (sensory) input to (motor) output and do not use each other's outputs at all.

A different version of system where all mechanisms function relatively independent of each other would be a relatively central architecture with one mechanism that computes all other mechanisms' outputs. Breakdowns of any but the central module would then not affect any other functions if the central module did not use other modules' outputs to further compute another module's output. Jerry Fodor proposes a structure similar to that one in "The Modularity of Mind" with specialized sensory input modules (different layers, transducers etc.) that feed into a central structure that conducts computations over those heterogeneous inputs (Fodor, 1983). A different architecture where mechanisms are highly independent of other mechanisms' information as inputs, too, is the type where mechanisms do use each other's outputs but are able to establish workarounds in case of defects such as bypassing a structure where a defective module's output is replaced by another module, as suggested in Anderson's "Neural reuse" theory (Anderson, 2010). His suggestion, however, is less probable in our picture of functionally highly specialized modules, as most modules that serve a function seem to need particular inputs. It is also important to keep in mind that the computations that are considered here are cognitive computations and not neural activation as in Anderson's article, hence, we are talking about a different (Marr-)level (Marr 1982): Anderson includes the implementation level, too, while we have restricted ourselves to the functional and, possibly, algorithm level.

The 'programs' of our highly specialized cognitive mechanisms, depending on which form a computation takes, are not able to deal with all sorts of inputs. Especially if we take seriously the condition that different modules serve different functions and solve different problems, one module is not able to handle all kinds of information as inputs. According to Robert Kurzban (Kurzban 2012, p. 40) there are no domain general modules; each module is fit for certain types of inputs (or tags, that is). These may not be types such as the categories made in common language (like 'mammals'), but rather something such as what Kurzban and Barrett call 'actual domains' (like 'hairy, warm animal') that rely on certain (in most cases more directly perceivable) features as opposed to 'proper domains', the domains the function 'has developed for' (Barrett and Kurzban, 2006).

This means that the outputs of each module consist of representations and whatever kind these may be, they are different symbols/representations for every module and not the same ’currency’. A symbol/representation would be, for example, something like ’cat’ (for a fictional module that has different kinds of mammals as output (’cat’, ’dog’ etc.)) or, if instead we wish to choose a fictional example that has more resemblance to an ‘actual domain’ output, ’small, hairy, warm animal with triangular ears’ or ’that which walks around houses and gardens and makes ’meow’ noises’.

Let us assume that the (fictional) functional module that normally would have the kind of mammal as output is defective. Another module has the kind of mammal as input (for example, because it computes whether the kind of mammal can jump very high or not). It would usually receive its input from the defective module with the mammal-kind output. This module will not be able to compute (use as input) another module’s output such as ’computer’ in a way that makes sense in terms of evolutionary functionality.

Hence, it is rather unlikely that a module can bypass one module’s defective output by instead computing a different module’s output if modules are as highly specialized as many Evolutionary Psychologists claim and, especially, if we follow Barrett and Kurzban’s account of modularity (Barrett and Kurzban, 2006) who claim that only certain forms of input fit certain modules specialized in those forms of input.

I conclude: The inputs of most mechanisms of our evolved system should not depend on the outputs of most other mechanisms: Either there are very few mechanisms involved in a chain of computations so most mechanisms go from (sensory) input to (motor) output directly or their output is fed to only a few different modules that receive it as an input (this would be the case with a central structure or only a few modules overall). Alternatively, and less probably, the mechanisms’ outputs are fed into many different modules, but those modules do not depend on those mechanisms’ particular outputs but are either able to compute without their outputs at all or their outputs are replaceable by different modules’ outputs, i.e. they can be bypassed. If the mechanisms are not completely independent of each other, according to the ’evolvability argument’, it should be more evolutionarily advantageous for more recently developed mechanisms to rely on phylogenetically more ancient mechanisms’ outputs/information items than the other way round.

Now that I have outlined the most likely computational way a mechanism is built if it has developed as an evolutionary module, I will examine whether Trolley Dilemma judgement mechanisms compute this way.

2.2 Are Trolley Dilemma Judgement Mechanisms Mechanistic Modules? Testing Mikhail’s Mechanism for modularity

As I have explained in the introduction, different scientists have proposed that the mechanism computing Trolley Dilemmas has developed as a result of evolutionary processes. They have argued that because of this, it is built up in a functionally modular manner.

I have argued at the beginning of Chapter 2.1, p. 17, that if a cognitive system is built from functional modules (that have developed as evolutionary adaptations), the whole system is also likely to consist of 'mechanistic modules', as I have called them: If we have congenital cognitive functions that enable (formerly) evolutionarily advantageous behavior, hence, if the structure of our mind is a product of evolution rather than upbringing, this structure is (according to the watchmaker argument) likely to have assembled as single informational processing units: During the genesis of the structure, each processing unit would have grown at a time in addition to the already existing ones with every successful mutation and formed a stable state (hence, promoted the fitness of the individual carrying it) by being able to solve an environmental problem. This refers to its computational architecture, not its material implementation.

The evolutionarily most advantageous form for those modules to develop is the following: To a large degree independent of each other, or at least of phylogenetically newer mechanisms. The aim in this chapter is to find out whether Trolley Dilemma judgement mechanisms have this computational structure. If they do, this would be evidence in favor of an Evolutionary Psychological explanation of the development of those mechanisms; if not, this would provide counter-evidence for them to have developed as evolutionary adaptations. The method for testing this assumption is to consider the computational structure that might be used to process Trolley Dilemmas.

As Michael R. Waldmann (Waldmann et al., 2012) rightly points out, John Mikhail is the first (and, to my knowledge, only) philosopher who proposed computational mechanisms that can account for at least three different, possibly morally salient distinctions that test subjects made in the assessment of Trolley Dilemmas. In his paper "Universal Moral Grammar: Theory, Evidence and the Future" (John Mikhail, 2007a), earlier in his doctoral thesis (Mikhail, 2000, p. 167) as well as in his book "Elements of Moral Cognition" (Mikhail, 2011, p. 169), he sketches computational steps that lead from Trolley Dilemma inputs to those judgement outputs that have been found in empirical studies.

Mostly, the inputs are written scenarios about runway trolleys that are heading towards five people and are going to kill them and the outputs are judgements of the subjects presented with those scenarios about the permissibility of killing one person in order to prevent the trolleys from killing the five. Whether most people find killing the one a permissible act varies according to how the person is killed.

Cushman et al. (Cushman et al., 2007) have found that people with very different cultural, educational and religious background judge very similarly in a certain set of Trolley Dilemmas: "[...] in the context of the trolley problems we studied, all of the demographically defined groups tested within our sample showed the same pattern of judgments [....]" (Cushman et al., 2007, p. 15).

Although there have been more studies with different variations of Trolley Dilemma scenarios, (e.g. Waldmann and Wiegmann, 2010, p.), as mentioned in the introduction, I will restrict the scope of discussion in this thesis to the Means/Side-Effect distinction (or what I will later define as a version of the Doctrine of Double-Effect, (see Chapter 6.2.1, p. 148), the Personal/Impersonal Harm distinction (that I will hereafter refer to more accurately as "Close Contact Harm Principle" and the Action/Omission distinction. I made this decision because those are the principles John Mikhail uses for his computational models and they are the empirically most tested ones, e.g. by (Schwitzgebel and Cushman, 2012), (Cushman et al., 2006) and (Powell et al., 2012).

The following dilemmas were presented to subjects by Cushman et al. (Cushman et al., 2007): "Denise is a passenger on a train whose driver has fainted. On the main track ahead are 5 people. The main track has a side track leading off to the left, and Denise can turn the train on to it. There is 1 person on the left hand track. Denise can turn the train, killing 1; or she can refrain from turning the train, letting the 5 die. Is it morally permissible for Denise to turn the train?" (In the picture, right and left track are exchanged, but the structure is the same.)

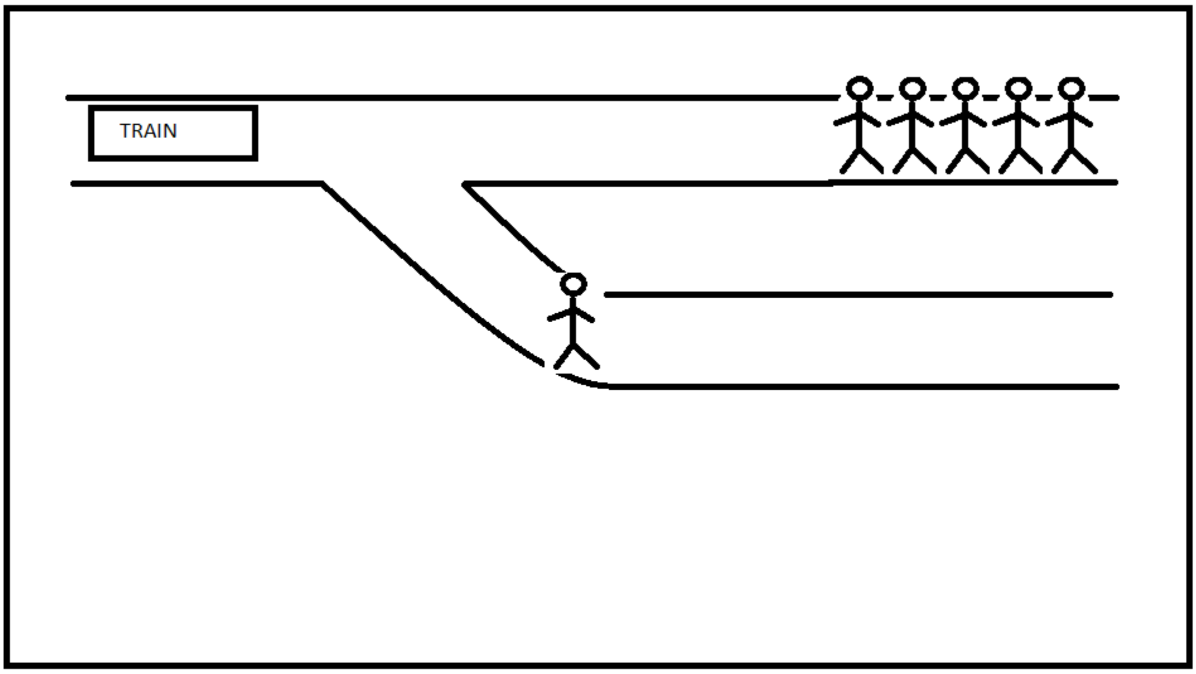


Figure 4: Denise/Switch case, Cushman et al. 2007, p. 6

"Frank is on a footbridge over the train tracks. He sees a train approaching the bridge out of control. There are 5 people on the track. Frank knows that the only way to stop the train is to drop a heavy weight into its path. But the only available, sufficiently heavy weight is 1 large man, also watching the train from the foot bridge. Frank can shove the 1 man onto the track in the path of the train, killing him; or he can refrain from doing this, letting the 5 die. Is it morally permissibly for Frank to shove the man?" "Ned is walking near the train tracks when he notices a train approaching out of control. Up ahead on the track are 5 people. Ned is standing next to a switch, which he can throw to turn the train to a side track. There is a heavy object on the side track. If the train hits the object,

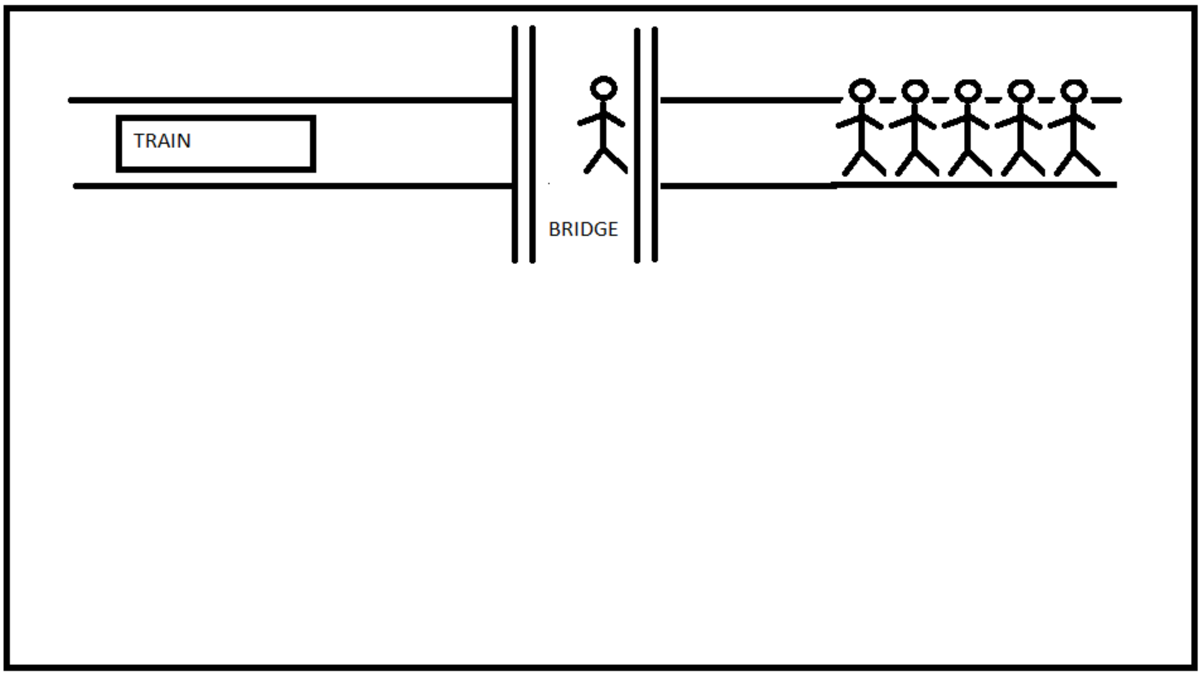


Figure 5: Frank/Push case, Cushman et al. 2007, p. 6

the object will slow the train down, giving the men time to escape. The heavy object is 1 man, standing on the side track. Ned can throw the switch, preventing the train from killing the 5 people, but killing the 1 man. Or he can refrain from doing this, letting the 5 die. Is it morally permissible for Ned to throw the switch?"

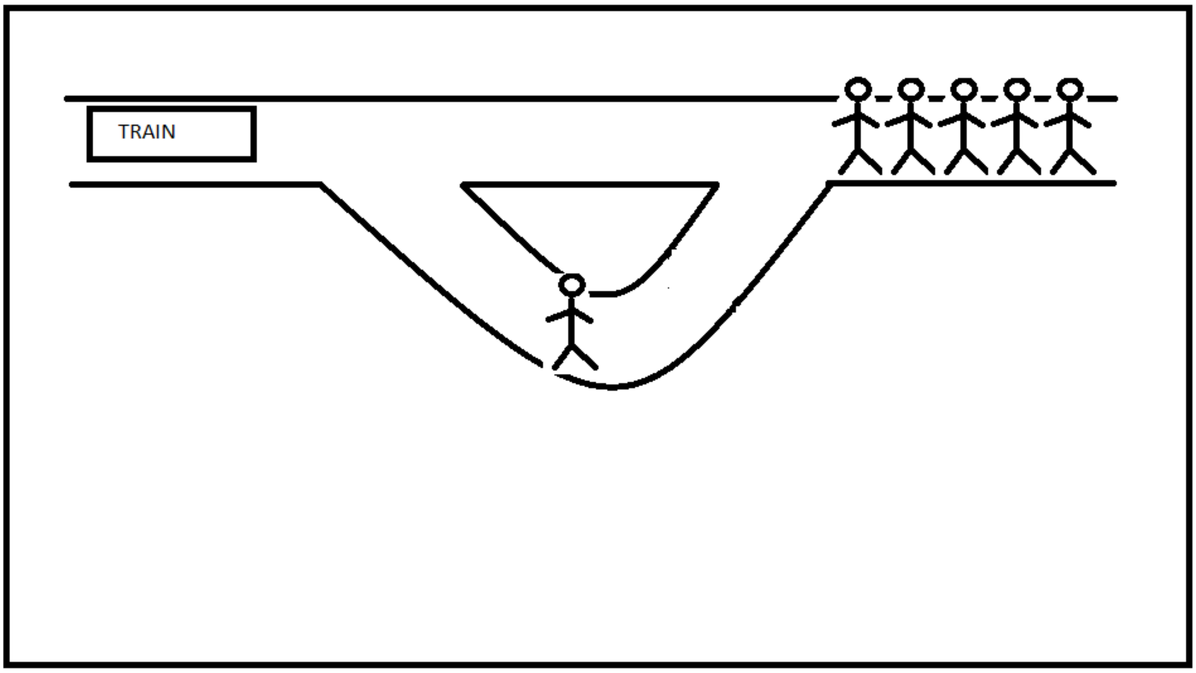


Figure 6: Ned/Loop case, Cushman et al. 2007, p. 6

"Oscar is walking near the train tracks when he notices a train approaching out of control. Up ahead on the track are 5 people. Oscar is standing next to a switch, which he can throw to turn the train on to a side track. There is a heavy object on the side track.

If the train hits the object, the object will slow the train down, giving the 5 people time to escape. There is 1 man standing on the side track in front of the heavy object. Oscar can throw the switch, preventing the train from killing the 5 people, but killing the 1 man. Or he can refrain from doing this, letting the five [sic] die. Is it morally permissible for Oscar to throw the switch?" [30]

All dilemma texts and pictures are copied from (Cushman et al., 2007, p. 6). The experiment was conducted in form of an online survey with about 5000 participants. [31] The participants had to choose whether it was morally permissible for the respective protagonist (Denise, Frank, Ned, Oscar) (Cushman et al., 2007, p. 6) to act ("turn the train", "shove the man", "throw the switch", "throw the switch") or not. The results were that

- 89% of the participants said it was morally permissible for Denise to turn the train, [32]
- 11% of the participants said it was morally permissible for Frank to shove the man,
- 56% of the participants said it was morally permissible for Ned to throw the switch and
- 72% of the participants said it was morally permissible for Oscar to throw the switch.(All results: Cushman et al., 2007, p. 7)

Interestingly, the ratio in which subjects judged the respective actions to be permissible did not differ much amongst different kinds of population groups:

> "'Our analyses generate two central conclusions: (1) in the context of the trolley problems we studied, all of the demographically defined groups tested within our sample showed the same pattern of judgments and (2) subjects

[30] They were copied along with the following caption: "*The description presented here represents an abbreviated version of the actual text that is described in the supplementary material, along with control scenarios. The schematic illustration is provided here, but was not given to subjects.*", italics in the original.

[31] According to the authors, the "Subjects were voluntary visitors of the Moral Sense Test website (http://www.moral.wjh.harvard.edu) from September 2003 to January 2004. Overall, there were some 5,000 subjects responding to the dilemmas targeted in this paper, covering 120 countries, but with a strong bias toward English-speaking nationalities. The website was promoted through print and online media coverage, online discussion forums, and word of mouth. All procedures were conducted in accordance with the Institutional Review Board of Harvard University, and followed the testing procedures of other web-based research projects." (Cushman et al., 2007, p. 4/5)

[32] Text and figures diverge here: While the figures on page 7 say that 85% found it permissible for Denise to turn the train and 12% for Frank to shove the man, the text on page 9 says that 89% voted that it was permissible for Denise to turn the train, whereas 11% judged it permissible for Frank to shove the man. The tables on page 9 indicate that the figures in the text section are correct because they repeat them; this is why I have chosen these values above. (All in (Cushman et al., 2007)

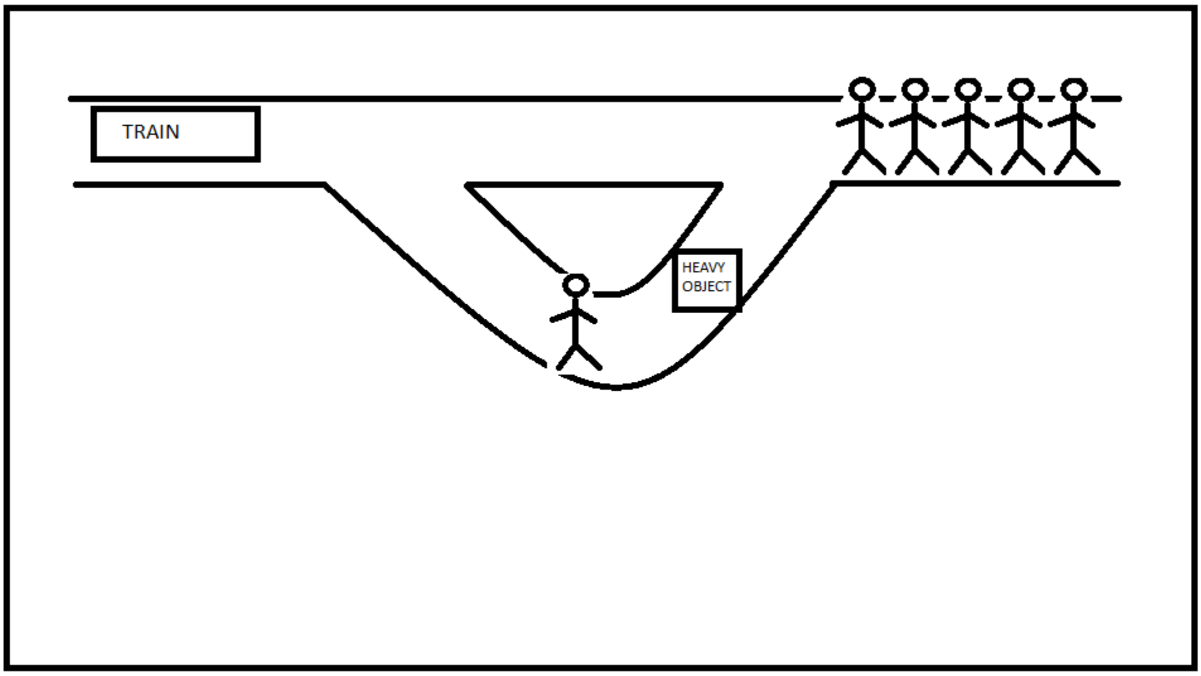


Figure 7: Oscar/Loop with Heavy Object case, Cushman et al. 2007, p. 6

generally failed to provide justifications that could account for the pattern of their judgments."(Cushman et al., 2007, p. 15). [33]

As the authors of the study report, "the observed pattern of judgments was consistent with at least three possible [...] distinctions:" The distinction between Personal and Impersonal modes of harm (Greene et al. 2004); the distinction between harm as intended and harm as foreseen side effect (Doctrine of Double-Effect); and the distinction between the redirection and the introduction of threats (Cushman et al., 2007, p. 15). Some of the authors, however, later dropped the latter distinction and henceforth tested for the distinction between harming someone by action or by omission instead (e.g. Dwyer et al., 2009, p. 497) and more research was dedicated to the action/omission distinction, which is why I will examine this claim instead of the redirection claim. This systematization contains, among others, the following principles:

1. Harms that have been caused by actions are worse than harms that have been caused by omissions.

2. Doctrine of Double-Effect: Harms that are caused as means for a greater good are worse than equivalent harms that are merely caused as a foreseen side-effect of an action.

3. Harms that involve physical contact are worse than equivalent harms that are generated by a non-human, causal intermediate effect like a thrown stone or a pushed button. (Formulations based on (Dwyer et al., 2009, p. 497))

[33] More recent studies relativize both findings; for the latter, see e.g. (Cushman et al., 2010); for the first, see Chapter 5.3, p. 113

In the following section, I will first show how John Mikhail thinks these principles are cognitively computed (from a situation as input to the judgement as output) and then show why I do not think they are computed by mechanistic modules.

John Mikhail hypothesizes that people compute Trolley type situations using a mechanism that goes through the following steps:

> "(i) identifying the various action descriptions in the stimulus [mostly: Trolley Dilemmas in written form], (ii) placing them in an appropriate temporal order, (iii) decomposing them into their underlying causative and semantic structures, (iv) applying certain moral and logical principles to these underlying structures to generate representations of good and bad effects, (v) computing the intentional structure of the relevant acts and omissions by inferring (in the absence of conflicting evidence) that agents intend good effects and avoid bad ones, and (vi) deriving representations of morally salient acts like battery and situating them in the correct location of one's act tree" (John Mikhail, 2007a, p. 146), citing (John Mikhail, 2007b).

Mikhail represents the computational mechanism in the form of rather complex representation trees but knowing the exact procedure is unnecessary to understand this argument. I will go through this procedure with an example to make it easier to understand the single computational steps cited above:

Let us assume that the Trolley Dilemma input is the following: "A runaway trolley is about to run over and kill five people, but the driver can push a button that will turn the trolley onto a side track, where it will kill only one person. Is it permissible to push the button?"

Step (i) and (ii) and (iii), identifying the action descriptions in the stimulus, placing them in an appropriate temporal order and (partially) decomposing them into their underlying causative and semantic structures, would split the action in question into the following parts:

> Pushing button at time t(0);
>
> turning train at time t(+n), hence at some later time;
>
> preventing train from killing men at time t(+n), hence at the same time as turning the train and, as a side effect of turning the train, causing the train to hit man at time t (+n+o), hence at some later point of time than turning the train;
>
> committing battery at time t (+n+o), hence at the same time as the train hits the man; committing homicide at time t (+n+o+p), hence after the train hit the man. (For a more thorough explanation of the timeline notation, see Mikhail, 2011, p. 124/125.)

John Mikhail's definition for battery is the following: "The prohibition of intentional battery forbids purposefully or knowingly causing harmful or offensive contact with another individual or otherwise invading another individual's physical integrity without his or her consent." (John Mikhail, 2007a, p. 145), citing (Epstein, 2004) and (Shapo, 2003). [34]

Thus we have computed a written input and broken it up to its temporal order. Mikhail writes: "Presumably this task is accomplished by relying on auxiliary verbs ("can," will," etc.) and other temporal clues in the stimulus, but I set aside for now a closer examination of these operations." (Mikhail, 2011, p. 171). Additionally, we have identified that causing the man to be hit by the train and thereby killing him is a side-effect and not the means to preventing the train from killing the five men. If we continue to step (iv) to compute the good and bad effects (applying principles such as that killing is bad and preventing something from killing someone is good), we will find that the main effect (preventing the train from killing the five men) is good and the side-effect (causing the train to kill the one man) is bad. Now we can either directly apply the version of the Doctrine of Double-Effect that only mentions means and side-effects or we can go on computing that, in the absence of counter-evidence, the agent's goal is to achieve the good effect, not the bad effect, and thereby find out about their intentions (v), namely to save the five and not to kill the one.

We now still have to compute acts of harm that could have happened to achieve the good effect (vi). These can include acts such as battery or personal/impersonal harm as a means or side-effect to saving the five: In the footbridge case, for instance, battery or personal harm is a means to saving the five which makes it less permissible, although in overall terms it would have been permissible as the agent intended the good and not the

[34] Note that Mikhail's definition of battery is not a version of the Personal/Impersonal distinction, as he explicitly states (John Mikhail, 2007a, p. 149). Mikhail believes that prohibition of battery and the Doctrine of Double-Effect are the main distinctions that make people judge actions permissible or impermissible and that it is not the Personal/Impersonal distinction. His evidence is that "Acts appear more likely to be judged permissible in these circumstances as counts of battery that are committed as a means decrease from three (Ian) to two (Victor) to one (Ned), and as these violations become side effects (Oscar, Walter, Hank) and additional structural features come into play." (John Mikhail, 2007a, p. 149); the Personal/Impersonal distinction alone could not account for the judgement pattern in the cited cases. The definition of the Personal/Impersonal distinction he applies is the following: "A violation is personal if it is (i) likely to cause serious bodily harm, (ii) to a particular person, (iii) in such a way that the harm does not result from the deflection of an existing threat onto a different party. A moral violation is impersonal if it fails to meet these criteria". (John Mikhail, 2007a, p. 149), citing (Greene and Haidt, 2002). However, both principles involve physical harm, and the judgement pattern of the dilemmas he mentions can be explained by a compound of the Doctrine of Double-Effect (or by a simplified version of it counting how often harm was done as means or as foreseen side-effect) and battery, but also by a compound of the Doctrine of Double- Effect and the Personal/Impersonal harm distinction. The main differences between battery and personal harm (that are important in the context of Trolley Dilemmas) are that personal harm involves serious bodily harm and battery involves knowingly causing contact that does not need to culminate in serious bodily harm: Offensive/harmful contact is sufficient for an action to be called battery (John Mikhail, 2007a, p. 149). Both principles cover very similar Trolley Dilemma type cases, namely those in which people are touched/pushed before they are sacrificed. The distinction does not seem too important to Mikhail either, because in a publication with other Trolley Dilemma researchers he seems to have agreed to proposing the Personal/Impersonal harm distinction as basic moral principle and does not mention his principle of battery (Cushman et al., 2007, e.g. p. 15).

bad effect as a goal. In the case we are now discussing, there is no battery involved, but in the footbridge case, battery (and also close-up, personal harm) as a means would be involved (shoving the man from the bridge as a means to stop the train) and that would make the action less permissible (example and computational mechanism: all in John Mikhail, 2007a.).

This description of the cognitive computation of a Trolley Dilemma stimulus is very simplified; for a broader and more thorough discussion including all presuppositions for this to work and for a key and reasons for the notation, see John Mikhail's Elements of Moral Cognition (Mikhail, 2011). The important part of this description is that steps of this process are or presuppose semantic processing, deriving causality from written action descriptions, ascribing intentions to agents, and weighing sums (five deaths against one death).

I will first expound why this is important and then elaborate on those steps one by one. In the next section, I will conduct some exegetic work and show that Mikhail thinks that his mechanism is mechanistically modular: John Mikhail speculates that the computational process he proposed might be innate, domain-specific and 'partially inaccessible', all notions connected to classical mechanistic modularity (Fodor, 1983):

> "Although each of these operations is relatively simple in its own right, the overall length, complexity and abstract nature of these computations, along with their rapid, intuitive and at least partially inaccessible character [8-10,33,36,37], lends support to the hypothesis that they depend on innate, domain-specific algorithms. However, this argument is not conclusive [43-45], and further research is needed to clarify the relevant conceptual and evidentiary issues." (John Mikhail, 2007a, p. 147/148), citing (Greene and Haidt, 2002), (Greene et al., 2001), (Greene, 2004), (Mikhail, 2008), (Cushman et al., 2007), (Cushman et al., 2006), (Nichols, 2005), (Sripada and Stich, 2005) and (Prinz, 2007a).

In the same context, he refers to a text by Jerry Fodor about modular mechanisms (Fodor, 1985). Mikhail compares his mechanism to language and vision mechanisms (John Mikhail, 2007a, p. 146) because all three mechanisms recover properties "when the stimulus contains no direct evidence" for them "from the stimulus by a sequence of operations that are largely mechanical." (John Mikhail, 2007a, p. 146). In the text Mikhail cites in this section, Fodor emphasizes that perceptual mechanisms are typically encapsulated. [35] Language ("recogniz[ing] [...] word boundaries in unmarked auditory patterns") and vision ("to recover a three-dimensional representation from a two-dimensional stimulus"

[35] "Although perception is smart like cognition in that it is typically inferential ['a lot of inference typically intervenes between a proximal stimulus and a perceptual identification', (Fodor, 1985, p. 2)], it is nevertheless dumb like reflexes in that it is typically encapsulated." (Fodor, 1985, p. 2)

[36]), the two mechanisms Mikhail compares to his mechanism, are well-known candidates for modularity and Mikhail refers, amongst others, to Fodor, Chomsky and Marr when he mentions language and vision ((Mikhail 2007b, p. 146), (Mikhail, 2011, p. 120)). Fodor is famous for his modular account of perception, Chomsky for his belief in innateness and his flirt with the idea (or metaphor?) that the computational apparatus for language is like an 'organ', and hence works more or less autonomously, and Marr for providing modular computational explanations for visual processes (see Craig, 1998, p. 635), including recovering three-dimensional representations from two-dimensional stimuli.

To sum up, Mikhail hypothesizes that his mechanism is innate, domain-specific and 'partially inaccessible' and draws very close analogies to, according to Fodor, typically informationally encapsulated mechanisms that have been treated as standard candidates for modularity in the past. Note that the properties 'rapid' and 'intuitive', which he attributes to his moral judgement mechanism (see quotation above), are also typically ascribed to modular mechanisms. Although Mikhail does not explicitly say so, I hope to have shown that he more than hints at his mechanism being mechanistically modular. While Mikhail suggests that his mechanism is mechanistically modular, other proponents of the Linguistic Analogy suggest that a 'moral faculty' (see, for instance, Hauser, 2006) could be responsible for computing Trolley Dilemma inputs. Those 'moral faculties' are, as the name says, domain-specific to morality.

I will show that the single elaborated computational process that has been proposed so far for judgement-making in Trolley type cases (for a less elaborated, and also non-domain-specific, non-modular computational model, see (Cushman & Young 2011)) is neither likely to be mechanistically modular in the ways described above (especially in respect to shallow connections to other modules) nor domain-specific. I agree that cognitive Trolley Dilemma judgement mechanisms (if there are such) probably compute in the way described by Mikhail or in a similar way: They need to progress from written Trolley Dilemma type situations to judgement outputs; to 'find out' whether the three principles apply, they have to compute the causal structure; to extrapolate the latter, they have to compute what the sentence means and possibly the temporal structure. Mikhail writes:

> "An interesting question is whether these computations must be performed in any particular order. Offhand, it might seem that the order is irrelevant; however, this impression appears to be mistaken. In fact, it seems that these computations must be performed in the order depicted in Figure 6.2(a), at least in our 12 [Trolley type] cases, because to organize the deontic structure of these actions, one must already grasp their intentional structure; to recognize their intentional structure, one must already grasp their moral structure; to recognize their moral structure, one must already grasp at least part of their

[36] David Marr developed a computational account for this aspect of vision.

> causal structure; and finally, to recognize their full causal structure, one must already grasp their temporal structure.” (Mikhail, 2011, p. 167/171).

If I can show that Mikhail’s mechanism is not domain-specific (hence, ‘built’ to compute a certain input as a module specialized on solving a particular ancestral problem) and informationally encapsulated, I have shown that it is not mechanistically modular in the sense it should plausibly be if it is innate.

And if all other possible Trolley Dilemma computational mechanisms compute in a way that is similar in the relevant aspects (namely the steps shown above), neither would they be mechanistically modular in that sense. This means that it becomes much less plausible that cognitive Trolley Dilemma computational mechanisms are innate. To show Mikhail’s mechanism is neither domain-specific nor informationally encapsulated, I will argue that at least some of the single computational steps are not domain-specific and therefore, the whole mechanism is not mechanistically modular.

Therefore, it is important that the Trolley Dilemma computing process contains or presupposes such steps as semantic processing, deriving causality from written action descriptions, ascribing intentions to agents, and weighing sums (five deaths against one death): These steps are not domain-specific, and I will show that for every single one of them. Not being domain-specific here means that the tasks in the single steps rely on mechanisms that are used for tasks other than Trolley (or moral) computations, too.

And, importantly: If those mechanisms are not domain-specific, they are not informationally encapsulated because they have access to inputs from non-moral sources as well. And the entire judgement process is very unlikely to be isolated in terms of information either, while the same applies to its component parts.

One could argue now that the judgement mechanism might contain components that can (and do) compute non-moral inputs as well at times but that the actual moral judgements are computed in an informationally encapsulated manner, that they compute straight from the moral stimulus input to the moral judgement output in the steps Mikhail described. The mechanism might not be modular, but its computations could be. This would be no problem for my argumentation line for two reasons: If we assume that the same physical mechanisms perform these computations in both cases (non-Trolley-Dilemma type and others), these mechanisms would likely have access to the same information during any kind of computation. This means that even if the process does not retrieve this (non-domain-specific) information, the mechanisms would theoretically be able to retrieve it (if the process does not prevent them from accessing the information, but why would it? The burden of proof is on the side of the objector to show that it does).

But, of course, we are talking about the computational, not the implementation level (compare Marr, 1982, p. 27). Mikhail hypothesizes, in his citation above, that the computations he proposes “depend on innate, domain-specific algorithms.” (John Mikhail, 2007a, p. 147/148). If we assume that the Trolley Dilemma judgement process is carried out by the same algorithms as other (non-moral) computations, those algorithms are, obviously, not restricted to the moral domain. If we imagine them as a chain of subroutines, those subroutines have access to just the same kinds of information in the Trolley Dilemma case as in any other case.

Even the proximal stimuli (hence, their 'actual domains') seem to be very similar: How does a social situation, presented either in written form or in real life, where someone intends something, differ from a moral situation whose computation depends on intentionality, formally? Do they have something that 'triggers' them and sends them along a 'moral pathway' that then restricts the information access? What could that be?

I hypothesize that the computations Mikhail's mechanism makes are, in point of fact, a chain of computations by mechanisms that compute inputs from other domains as well. If the entire mechanism is not informationally encapsulated and the partial mechanisms of the (hence non-)module interact with other (non-?)modules, the picture we get would be a network of mechanisms and certainly not one modular mechanism that does indeed only receive inputs and outputs and otherwise does not exchange much information with the other mechanisms. In addition, if domain-specificity (or informational encapsulation) does not apply, the practicality argument would cease to hold. If the moral module were to break down, many other functions would be impaired as well because its components would be responsible for or involved in so many different types of computations that are not connected to morality. Again: A mechanism like this is not modular in the sense that would satisfy the mechanistic modularity conditions I have elaborated before with the 'evolvability' and the 'practicality' argument.

So, let us start with the semantic processing: Identifying the action descriptions in the stimulus and ordering them first temporally, then causally. Without going into particulars as to how people compute situations, to understand the written sentences and identify relevant action descriptions in a stimulus, (not to mention the visual system translation one needs to identify the letters at all) one needs skills that are required for language processing as well (and, when the stimulus is presented in spoken form, those modules one needs for word distinction a previously mentioned by Mikhail).

Understanding sentences, the actions they recount and their temporal and causal structure, however we do this, is something we do not only do when we are presented with Trolley Dilemma cases. We do it every time we read and understand something. This part of the computation cannot be something that we developed specifically to compute moral situations (or if we did, we use it for many other instances as well). That means it cannot be domain-specific. Of course, this mechanism might have developed twice, once for the general function of reading and understanding situations and the other one for reading and understanding Trolley Dilemma (or moral) situations.

To anticipate this objection: Some theorists (e.g. (Marcus, 2006); see also (Calabretta et al., 2000), abstract) hypothesize that some cognitive mechanisms get copied in the course of evolution and the copies are then used for different, but similar purposes. Hence, it might be that those reading and understanding skills got copied and ‘pasted’ into the Trolley Dilemma (or moral) computation module to serve merely moral computation purposes. But, as I will set out in the following section, the understanding-and-attributing-causality part is only one of many mechanisms that are needed for Trolley Dilemma (or other moral) computations. But to successfully compute Trolley Dilemmas (in a way that gives us a systematic judgement pattern), we also need abilities from different domains.

This, however, would contradict our evolutionary theses as described in our ’evolvability’ argument: A copy of the reading and understanding skills would, without all those other steps needed to compute Trolley Dilemmas, not be any more useful than the original copy and hence have no evolutionary advantage before it has evolved to fully compute moral situations. And as the other steps, as I will show, are or rely on domain-general mechanisms, too, on that duplication-and-deviation account, all of those steps would have had to be copied at once to build up a functional module and (possibly; see next Chapter 2.3, p. 56 about Functional Modularity) lead to an evolutionary advantage. Hence, a module that is solely a copy of a different module generating action representations and attributing temporal and causal orders would have no evolutionary advantage and hence no reason to persist and be passed on. The same holds for the domain-general mechanisms that the next computational steps under discussion include or presuppose, for example: Attributing intentions to agents. Mikhail writes that in the absence of contrary evidence, people represent the good effects but not the bad effects of an action as intended and judge based on this distinction (e.g., John Mikhail, 2007b, p. 8). Judging moral dilemmas depending on the intentions we attribute to the agents, however, according to Baird and Astington 2004 (Baird and Astington, 2004) correlates with performance in false-belief tasks, at least in children: 5- -and 7-year-old children who scored better in false-belief tasks differentiated moral situations based on information about the agents’ motives more often than those who scored worse: They, for instance, evaluated a child’s turning on a hose to collapse her brother’s sand castle worse than a different child’s turning on a hose to help take care of the garden.

This could mean that whether someone intended a good or a bad effect in a situation with both outcomes only becomes salient to children as soon as they have the concept of intentions at all (hence, in every situation, not only Trolley/moral dilemmas) and/or are able to attribute them to other people, e.g. the agents in moral dilemmas. According to Baird et al., 4-year-olds already start judging according to intentions, and such judging according to intentions, as I mentioned above, is correlated to performance in false-belief tasks (Baird and Astington, 2004). 79% of the 5- and 6-year-olds tested by Powell et al. judged actions where someone is harmed intentionally and as a means to saving the agent worse than actions where someone is harmed in a not explicitly intentional way and as a foreseeable side-effect to saving the agent (Powell et al., 2012, p. 191). Hence, judging according to something similar to the Doctrine of Double-Effect [37] seems to develop to some extent after judging based on the agent's intentions which, in turn, is contingent upon performance in false-belief tests.

All those results suggest that there is no domain-specific 'moral module' mechanism, at least no mechanical one, that 'ripens' at a certain point of time. They suggest that children develop the ability to attribute false beliefs in any kind of situation first, then they learn distinguishing agents' intentions and judge accordingly, and at about the same time or a little later start judging more complicated cases that involve intentionality as well as means vs. side-effect according to the agents' motives and/or the causal necessity structure (means/side-effects). This would indicate that, if there are any modular mechanisms at all, those mechanisms are built upon false-belief skills and attribution-of-intentionality-skills that are applicable (and applied, as false-belief-experiments prove; see e.g. (Wimmer and Perner, 1983)) to any kind of input, inside and outside the moral domain.

To sum up: Attributing intentions and judging agents based on them seems to be contingent on performance in false-belief tasks; judging based on agents' motives and the means/side-effect structure of an action seems to be contingent on judging based on agents' intentions: Judging based on the agents' motives and the means-/side-effect structure is a conjunction of two properties, hence judging based on one of them (the agents' motives) should be a precondition for being able to judge the conjunction; and all of those abilities seem to develop one after the other. This gives us reason to believe that those mechanisms (for recognizing false beliefs and attributing intentions), that can compute on all kinds of actions (hence, have all kinds of actions as input domains), are component parts of the Trolley Dilemma judgement mechanism (hence, Mikhail's mechanism or a very similar mechanism). This makes processing Trolley Dilemmas a chain of computations and the Trolley Dilemma judgement mechanism an accumulation of other mechanisms with other than moral functions.

In Cushman and Young's 2010 article "Patterns of Moral Judgment Derive From Nonmoral Psychological Representations", they agree that at least part of the cognitive process that leads people to decide according to the Doctrine of Double-Effect (or, in this case, means versus side-effect) and the Action/Omission Principle is domain-general: "On this model some patterns observed in our moral judgments are derived from the architecture of cognition in non-moral domains such as folk-psychological and causal cognition, rather than from a domain-specific moral faculty." (Cushman and Young, 2011, p. 17). The subjects attributed, for instance, less intentionality to a boat driver who lost some seaweed that was on his boat as a side-effect to boating over to photograph some "playful seals" than to one that lost it as a means to being able to photograph them: They thought that "the driver intended for the seaweed to fall off" to a larger degree (The

[37] Powell et al. do not test for classical Doctrine-of-Double-Effect-situations; see Development Chapter 4, p. 91!

question they posed was: "To what extent did [agent] intend to [outcome]?" (Cushman and Young, 2011, p. 8)) (Cushman, 2014, p. 5). [38]

This shows, on the one hand, that Mikhail's suggestion that people ascribe intentionality based on the causal line of events might be right: Mikhail writes that people attribute the agent's intentions "by inferring (in the absence of conflicting evidence) that agents intend good effects and avoid bad ones." (John Mikhail, 2007b, p. 10). He constructs this in a way that, as Cushman writes, "[a] handy test of the means versus side-effect distinction asks: could the victim be removed from the situation without interfering with the agent's plan?" (Cushman, 2014, p. 2). Hence, if we compare the Loop and the Loop-with-Heavy-Object situation, we attribute intentional battery to the agent in the Loop case because had the victim not been there, the train would have just looped back and killed the five on the main track. Hence, if the agent intended to save the five by diverting the train to the loop track, he must have taken into consideration that the train hit the one person and was stopped. The harm was intended because it was a means to saving the five.

In the Loop-with-Heavy-Object situation, however, the harm is computed as unintended because the victim's presence was not necessary to stop the train: Had the victim not been there, [39] the stone would have stopped the train.

Hence, harming the victim is only a side-effect, but not a necessary means to saving the five. Subjects seem to attribute intentionality in a very similar way in Cushman and Young's experiment: The means/side-effect structure seems, at least partially, to determine how much intention people ascribe to the agent. This confirms Mikhail's model of how those dilemmas are computed.

On the other hand, it once again shows that the mechanisms in play when Trolley Dilemmas are computed are likely the same as in non-moral situations. This constitutes evidence against the thesis that the Trolley Dilemma judgement mechanism is domain-specific, informationally encapsulated and has developed evolutionarily as proponents of the Linguistic Analogy hypothesize. [40]

[38] The website that Cushman and Young 2011 mentioned as source for their scenario descriptions is not online at this point of time (http://moral.wjh.harvard.edu/methods.html; 06/08/15); hence I am not able to provide any more particular descriptions of the scenarios used to test the means/side-effect effect on ascription of intentionality.

[39] Of course, we could individuate the action differently and maybe redefine either the preconditions to assume that the agent intends something (you could say that someone who foresees something already intends it) or redefine the event of the train hitting the victim as not identical with harming or killing them, but that would be (and has been) material for several other articles (see, e.g., (Nelkin and Rickless 2015), (Fischer, Ravizza, and Copp 1993).

[40] They also show that the attribution of intention (maybe in connection with the distinction of means/side-effect) seems to be responsible for permissibility judgements: In cases where the agent explicitly did not intend to harm someone but accidentally did so, people did not judge the situation where he harmed someone as a means differently than when he harmed someone as a side-effect (e.g. a burglar who knocked a barrel over that accidentally hits a bystander and forces the policeman who was hunting the burglar to help the bystander vs. a burglar who knocked a barrel over to make the policeman dive out of the way and accidentally hits a bystander who gets hurt are both judged the same) (Cushman and Young, 2011, p. 12).

Cushman and Young show that the same holds for the action-omission distinction: Whether something is the consequence of an action or an omission seems to be the basis of attributing causality to an agent ("How much of a role did [agent] play in causing [the outcome]?" (Cushman and Young, 2011, p. 8)) in moral and in non-moral cases, and this causality (maybe in connection with the action/omission distinction) seems to be crucial for judging harmful actions worse than harmful omissions in moral cases. [41]

Just as with the distinctions between means and side-effects, the distinction between actions and omissions in moral situations seems to be computed by a mechanism that is not restricted to the moral domain. The mechanism attributes causation to agents based on whether something happens because of their action or whether it happens because of their omission (Cushman and Young, 2011).

After I have shown that analyzing temporal order and causality as well as ascribing intentions and ascribing causal responsibility to agents are likely computed by domain-general mechanisms, I will now show that weighing sums such as five deaths against one death is likely computed by domain-general mechanisms, too. I will follow two strings of evidence:

Firstly, children seem to be able to judge moral cases that do not involve weighing sums earlier than they are able to judge moral cases where they have to weigh sums. So, weighing sums does not seem to develop at the same time as, for instance, judging cases according to the agent's motives or at least it seems to interfere with children's judgements up to a certain age, leading them to judge differently than adults and differently than they would in cases without outcome information. This learning (or maturation) process seems to take place in different steps than the learning (or maturation) of the Trolley Dilemma judgement mechanism (if it develops in steps at all, see Development chapter) and the weighing sums mechanism does not seem to be an integral part of the moral judgement mechanism as it seems to lead to distorted assessments as compared to adults' assessments or children's assessments when there is no outcome information given.

Secondly, people show the same patterns when they judge economical outcomes as when they judge moral dilemmas. If the system used for both is the same, that means that this part of moral dilemma assessment is not solely moral either.

Let us start with the developmental part. I will cite evidence about ultimatum and dictator games for this part. In ultimatum and dictator games, one player gets to share

[41] An example for non-moral cases of actions/omissions are the following examples: "Ed is driving to the theater with a cord hanging out the side of his car. He approaches a rock resting by the side of the road. If he does not slow down, the rock will be knocked off the road by the cord and fall down a steep cliff. If he does slow down, he'll be late to the theater. Ed keeps driving quickly and knocks the rock off the side of the road." vs. "Jack is driving to the theater with a cord hanging out the side of his car. He approaches a rock that is about to fall off the side of the road and down a steep cliff. If he slows down, the cord will block the path of the rock and prevent it from falling, but Ed will be late to the theater. Ed keeps driving quickly and the rock falls off the side of the road." (Cushman and Young, 2011, p. 7). Those 'dilemmas' are, of course, only non-moral if the falling rock does not have any immoral consequences such as harming people (and the falling is not immoral itself).

some valuable resource (e.g. money) with another person and can decide how much they wants to give to the other person and how much they want to keep. In ultimatum games, that other person can reject the offer; in that case, neither of them is going to keep their share of the resource (and people tend to do that when the offer is very low, even though they do not get anything in that case which, economically, is even worse than getting a little) (see, for instance, Thaler, 1988). In dictator games, the other person cannot reject the offer; the 'dictator' can give them as little as they want to without fearing not getting anything themselves. Knowing that people might reject offers that are too small in the ultimatum game, and being selfish, it would be rational to not give anything in the dictator game (or, being a compound of just and selfish, to give less in the dictator game than in the ultimatum game).

Four-year old children made a smaller distinction between dictator and ultimatum games than adults did (they gave only 7% less in dictator games than in ultimatum games as opposed to adults, who gave 20% less) (Lucas et al., 2008, p. 84).

Lucas et al. interpreted this the following way:

> "[C]hildren seemed to understand that they did not need to offer as much in the dictator game as in the ultimatum game. But their ability to perform a cost/benefit analysis was limited. They did not seem to appreciate the degree to which they could 'shade' their offers without penalty." (Lucas et al., 2008, p. 84).

Powell et al. apply this as a possible reason for their results that 5/6-year olds judged actions as bad both when only harm was done and when the harm produced "prevented similar harm from being done to a larger number of individuals", hence the effect was mainly good (Powell et al., 2012, p. 192).

Although it is a big jump from differences in outcomes of economic games to weighing sums in moral judgements and economic games might be far more complex to understand than outcomes in dilemmas (and the reason for children to offer only a little more in ultimatum games than in dictator games could have reasons other than non-understanding, e.g. a stronger sense of fairness), Powell et al.'s and Lucas' results – together with the results by Yuill et al. 1984 who found that 3- and 5-year-olds "rated an actor with a good motive more favorably than an actor with a bad motive only when the value of the outcome matched the value of the motive" (Baird and Astington, 2004, p. 39), referring to (Yuill, 1984) – might be explained by a story showing that very young children are not very good at weighing outcomes and might get distracted from their moral reasoning when they have to do that. [42]

[42] I am aware that there are many possibilities to explain the data differently, too, and they are only part of the case I am making for a domain-general mechanism for weighing sums; I can say in my defense that the authors of the papers about moral judgements that I have cited seem to interpret the results similarly to the way I do.

In this case, either weighing sums is not part of the moral judgement mechanism and matures later than the abilities to judge morally or moral cases including weighing sums are just too complex for children to judge them as adults would because weighing sums is not easy for them and the domain-general sum-weighing mechanism is one step of many domain-general mechanisms that compose the chain of mechanisms needed to judge Trolley Dilemmas. I will continue with the second string of evidence which is even stronger confirmation that a domain-general mechanism does the sum-weighing part of judging moral dilemmas: In an experiment by Rai and Holyoak, people had to rate whether they agreed to an action in the following setting: "A high-speed train is about to hit a large railway car with 10 (40) people. An employee of the train company is sitting in an office at the station. He could press a button on a control panel which would move a small railway car with two people into the path of the train. This will slow the train down and give 8 of the 10 (40) people on the large car time to escape. However, the action will kill the two people on the small car. To what extent is sacrificing the two people on the small railway car the right thing to do?"(Rai and Holyoak, 2010, p. 316). People agreed with sacrificing the two significantly more when 8 out of 10 people escaped than when 8 out of 40 people escaped. In a follow-up experiment, subjects where asked how many people needed to be saved to make sacrificing the two people "the right thing to do" with 10 or 40 people on the train. Even if they answered for both dilemmas successively (similar to the above text and with the two versions only differing in terms of the number of people who cannot escape), "the threshold number of lives that needed to be saved to justify taking action roughly tripled when the number of potential victims was 40 as opposed to 10."(Rai and Holyoak, 2010, p. 318).

Interestingly, the phenomenon that people would sacrifice a smaller amount if they could save a smaller fraction of what was at stake was first observed in economic studies in which "willingness to incur a cost in order to save on a product depends on the relative savings one will experience rather than the absolute savings." (Rai and Holyoak, 2010, p. 315), referring to (Thaler, 1999); people would, for instance, rather drive 20 minutes to save $5 on a calculator that costs $15 than on a jacket that costs $125 (Rai and Holyoak, 2010, p. 315), referring to (Kahneman and Tversky, 1984, p. 347).

People seem to have the same kinds of mechanisms when making consumer choices as when they are judging moral dilemmas. As Rai and Holyoak rightly argue, deontological, basic utilitarian and other consequentialist frameworks would all predict that what matters is the absolute number of saved lives (in the deontological case, in most frameworks, none should be sacrificed) and would hence predict different outcomes (Rai and Holyoak, 2010, p. 315).

We have seen that the effect does not seem to be genuinely (or exclusively) moral, as it emerges in non-moral contexts such as economic thinking and is not part of most prominent moral theories. This, again, seems to be evidence that the sum-weighing part of judging Trolley Dilemmas is computed by a domain-general mechanism.

Additionally to those behavioral data, Shenhav and Greene have found some fMRT evidence that the brain mechanisms that compute Trolley Dilemma Type cases are similar to those that compute economic choices. They presented their subjects with dilemmas of the following type: “You are driving a rescue boat in the ocean, heading east towards one drowning man. You receive a distress signal informing you that a small boat has capsized in the opposite direction, and all the people aboard are now drowning.

You know that if you immediately change course and go full speed, bearing west, you will reach these people in time to save them. However, if you do this, the one man to the east will certainly die. If you do nothing and hold your course, the one man will be saved, but you will not reach the people to the west in time to save them.

You also know that the only other rescue boat in the area is much further to the west, so would be unable to reach the one drowning man. But there is a chance the rescue boat will reach the group drowning to the west.

Consider each of the following scenarios and, for each one, determine how morally acceptable you think it would be to change your course to head toward the group to the west. Individuals drowning to the west: [variable Magnitude value appears here] Probability of alternate rescue boat reaching them: [variable Probability value appears here]” (Shenhav and Greene, 2010, Supplementary Information).

Unlike the classical dilemmas, the number of people who would die if the one person does not get sacrificed varies here. Additionally, the authors introduced non-zero probabilities for the group of people to be saved even if the one person is not sacrificed. They found, amongst other things, that “BOLD signal in the vmPFC/mOFC correlated with the “expected moral value” of decision options”, this moral value being the interaction between the number of lives at stake and the probability that they will be saved. Others (Shenhav and Green (Shenhav and Greene, 2010, p. 671) cite, amongst others, (Chib et al., 2009) for the vmPFC) have found that those brain regions play a role in decisions where people have to weigh between different kinds of rewards for themselves (such as food and money). Chib et al. hypothesize “that a specific region of vmPFC holds a representation of value regardless of the categories of goods presented, and regardless of the specific type of comparison being performed.” (Chib et al., 2009, p. 12319) and Shenhav and Greene suggest that this could mean that computations about decisions that involve weighing other people’s lives against each other in moral contexts might be computed by the same mechanisms as those that involve weighing other decision values (food, money), hence by domain-general mechanisms. They also found results that “suggest that an individual’s sensitivity to lives saved/lost in the context of moral judgment is in part determined by the same mechanisms that determine that individual’s sensitivity to the probability of loss and to overall reward in the context of self-interested economic decision making.” (Shenhav and Greene, 2010, p. 673) and summed up: “in implicating domain-general mechanisms, our results speak against the hypothesis that moral judgments are produced by a dedicated, domain-specific “organ” for moral judgment.” (Shenhav and Greene, 2010, p. 674).

Although I will mostly confine myself to the computational level, I allowed for this excursion because the information was too significant to keep out of this analysis.

In the last paragraphs, I have cited evidence that the mechanisms weighing sums in Trolley Type Dilemmas are the same ones responsible for weighing sums in economic decisions and hence not domain-specific to the moral realm and, possibly, not even for the social realm.

In the previous chapter, I have shown why the modularity claim is important for proponents of innateness: Modules evolve more easily and, if they are not very interconnected with other modules, it is likely that a breakdown of one mechanism will not render the others non-functional. I have shown that modules do not necessarily have to fulfil all of Fodor’s conditions to have those evolutionary advantages; they have to be relatively informationally isolated from each other and especially older modules should not depend on more recent modules. I have shown why those properties are most likely not given in domain-general mechanisms.

I have further shown that not only do many proponents of the Linguistic Analogy claim that the Trolley Dilemma judgement mechanism has developed evolutionarily and is a relatively independently working ’organ’, but also that John Mikhail claims that the specific mechanism that he believes computes those judgements has many properties that are commonly ascribed to (classical, mechanistic) modules (even if he does not explicitly call it modular); a reason why he does this is probably the connection I have established between modularity and evolutionary advantages, as Mikhail is one of the proponents of the innateness hypothesis (Mikhail, 2008). I have cited empirical evidence suggesting that large parts of Mikhail’s mechanism that contains the steps necessary to divide Trolley Dilemma cases along the lines people seem to divide them (Doctrine of Double-Effect, Action/Omission and Close Contact Harm/Distant Impersonal Harm) are not domain-specific and hence neither modular in the evolutionarily relevant sense nor specialized in computing moral judgements. In this context, I have also shown that Mikhail’s steps or very similar steps are necessary to compute Trolley type Dilemmas and hence this kind of criticism applies to all kinds of mechanisms that judge Trolley Type Dilemmas in the given patterns.

One possible objection remains: That Trolley Dilemmas or even morality are not natural kinds, that the computation of those dilemmas is specialized and modular, but not for the realm of morality, but for the greater realm of social cognition.

I have three answers to this:

Firstly, the question of this dissertation is whether the Trolley Dilemma judgement mechanism is an evolutionary adaptation. If this mechanism is used for other tasks as well, it ceases to be a Trolley Dilemma judgement mechanism and starts to be a social cognition mechanism. Of course, by researching Trolley Dilemmas, what ultimately interests us is not only how people make judgements in Trolley Dilemmas, but also how they develop the principles that they judge in accordance with. Hence, I examine the hypothesis that there is a moral module, as claimed by proponents of the Linguistic Analogy. This is not only a matter of terminology. The questions posed to subjects in Trolley Dilemma cases are formulated in normative terms: Moral permissibility (Cushman et al. 2007), "should" (Ahlenius and Tännsjö 2012), moral acceptability (Shenhav and Greene 2010). The Doctrine of Double-Effect that has been widely used as an example has been treated as a classically moral principle in literature over the years (see Introduction 1.3 p. 9), Trolley Dilemmas stem from research on Ethics, and "trolleyology" is usually treated as branch of moral psychology/philosophy. Proponents of the Linguistic Analogy claim that a "moral organ" exists (e.g., Hauser 2006) and John Mikhail calls his paper about the Trolley Dilemma judgement process "Moral Cognition and Computational Theory" (Mikhail 2007b). I have shown that the people who claimed that there was an evolutionarily developed mechanistic moral module responsible for the Trolley Dilemma judgements were wrong. Whether a module for social cognition exists is simply a different question.

Secondly: If a module for social cognition existed, it would be much broader and more domain-general and hence less interesting in terms of innateness claims: It would have very diverse inputs and might not even be largely predetermined.

Thirdly: Related to the second claim, and more importantly, weighing sums seems at least to be present in economic considerations as well and those belong, if at all, only partially to the social domain. This is even more obvious for recognizing causal structures of sentences: If we are generous and regard those economic considerations as something we can use in trade activities (and trade as part of the social domain) and those causal structures as part of language (and not logical thinking which could be applied to anything), and language as part of the social domain, we end up with a bloated 'module' for social cognition that might as well be a domain-general inductive learning mechanism.

I have shown that Trolley Dilemma judgement mechanisms are unlikely to compute in a mechanistically modular way. But there is the second notion of modularity which makes it more probable for a mechanism to have developed (or, as I will argue, which means that it actually has developed) as an evolutionary adaptation: Functional Modularity. Hence, I will examine whether we have reason to believe that Trolley Dilemmas have developed as functional modules.

2.3 Functional Modularity and Evolutionary Adaptations

Scientists who call themselves Evolutionary Psychologists have a certain view of how behavior can be explained. They hold the view that the mind consists of a large set of psychological functions. If they are massive modularists (which most of the scientists cited here are, see section about massive modularity), they believe that the mind is composed solely of modules that have developed as a result of natural selection. This natural selection mainly took place in the Pleistocene. Functional modules can be tested and predicted in different ways; there is, however, a critique on how they are predicted and tested that applies to Trolley Dilemmas as well and might make it difficult to test whether cognitive mechanisms of Trolley Dilemmas have developed according to that paradigm or not. In this section, I will explain what Psychological Functional Modules are and how they can be tested, whether the empirical results would lead us to expect that the Trolley Dilemma judgement mechanism is a functional module and what might be alternative explanations for the intercultural similarity of answers to Trolley Dilemmas.

Psychological Functional Modules are domain-specific psychological mechanisms that promote behaviors that were advantageous for spreading the genes underlying them in a Pleistocene environment. To test whether something is a Functional Module, the two most promising methods are the following: Testing whether the trait is a solution to a challenge that might have been present in a Pleistocene environment (Reverse Engineering) or whether a trait can be predicted by challenges we know have been present in a Pleistocene environment (Reconstructive Engineering). I will further expound this in the following section and after that show why those testing methods are problematic if the postulated module's function lies in the social domain.

More precisely, I will first give an account of what Evolutionary Psychological functions are. After that, I will expound ways of devising with Evolutionary Psychologist hypotheses about which psychological mechanisms qualify as candidates for Functional Modules as well as how to test them. The passages about the testing methods will end with a critical assessment of the respective method. This is to show that Evolutionary Psychologist hypotheses are difficult to test, especially in the realm of social interactive behavior. I will cite David Buller's "Adapting Minds" frequently in this section as I think he has done excellent work in systematically listing the problems with many Evolutionary Psychologist methods (Buller, 2006a).

The following are the main assumptions about Evolutionary Psychological Functions:

- they developed in the Pleistocene (1.8 million to 10,000 years ago) as a result of environmental challenges (Buller, 2006b, p. 197);

- they are likely to be common to every human being (e.g. Buss 1995, p. 11);

- they are inheritable (see also (Leda Cosmides and Tooby, 1994) (Kurzban, 2012) (Pinker, 1997), (Buss, 1995));

- they are solutions to ancestral functional problems, not necessarily to problems in a modern society. (e.g.(Buller, 2006b, p. 199) (Buss 1995, p. 10)).

The narrative on which Evolutionary Psychology is grounded goes as follows: In the Pleistocene, the ancestors of all modern humans used to live in small groups as hunter-gatherers. This is supported by archaeological evidence. Compared to the 10,000 years after this, the Pleistocene time span (almost 1.8 million years) was very long. This is why most of human evolution took place during the Pleistocene. Living as hunter-gatherers, some kinds of environmental challenges stayed relatively stable for all humans along that time-span. [43] Brain traits developed as a reaction to those challenges. Particularly, those brain traits were selected for that were the biological basis of psychological mechanisms that promote adaptive behavior: Individuals who showed behavior that led to the spreading of their genes (be it in terms of procreating themselves or by 'helping' kin [44] with a similar set of genes to survive and procreate) had psychological mechanisms that enabled or promoted that behavior. If genes were responsible for the development of those psychological mechanisms, those genes would stay in the gene pool. Hence, the individuals with those genes would possibly endow their offspring with the set of genes responsible for that behavior which would for their part support the promotion of the offspring's genes.

So the 'chain' goes as follows: Genes → brain structure → psychological mechanisms → behavior: Genes encode and enable the production of a brain structure that is correlated to psychological mechanisms [45] that lead to a certain kind of behavior. The very genes that, along this presumably causal chain, lead to behavior that helps spreading them are systematically selected for. This is why humans have evolved in a way that they are endowed with many psychological mechanisms. Those mechanisms are fit to solve environmental problems in the Pleistocene in a way that either gave themselves or their kin a bigger chance to procreate or a chance to have a larger number of descendants.

The more special such a mechanism is, the better it is fit to solve problems: "[A] jack of all trades is necessarily a master of none", as goes the often cited passage by Cosmides

[43] Stephen Pinker formulated this (and some facts about the mind's architecture) and gave examples for some challenges in the following quote: "The mind is a system of organs of computation, designed by natural selection to solve the kinds of problems our ancestors faced in their foraging way of life, in particular, understanding and outmaneuvering objects, animals, plants, and other people." (Pinker 1997, p. 21).

[44] And by doing that promoting their relatives' genes. As this set of genes is similar to their own genes, an individual is not unlikely to thereby promote genes they have in common and that cause the same psychological mechanisms to develop in their kin's offspring as have been at work in themselves (see (Hamilton, 1964)).

[45] Most Evolutionary Psychologists are committed to computational theory of mind and do not talk much about the implementation of those psychological mechanisms (see e.g. (Pinker 1997, p. 26); (Kurzban, 2012, p. 27); (Dahlgrün, 2015, p. 21)). Tooby and Cosmides write: "The idea that low-level neuroscience will generate a self-sufficient cognitive theory is a physicalist expression of the ethologically naive associationist/empiricist doctrine that all animal brains are essentially the same. In fact, as David Marr put it, a program's structure 'depends more upon the computational problems that have to be solved than upon the particular hardware in which their solutions are implemented' (1982, p. 27). In other words, knowing *what* and *why* places strong constraints on theories of *how*."((L Cosmides and Tooby, 1994, p. 46, emphasis in the original), citing (Marr, 1982, p. 27)). *Some* arguments of Evolutionary Psychologist discussion, however, do depend on the brain structure ("implementation") level (see e.g. (Geary and Huffman, 2002), (Buller, 2006b)).

and Tooby (Leda Cosmides and Tooby, 1994, p. 89) or, further elaborated, by Kurzban: "*Specialization yields efficiency.*[...] As an object's shape conforms to the requirements of a particular task, the shape simultaneously - and necessarily - becomes worse at others" (Kurzban 2012, p. 32; emphasis in the original). [46] So if there were recurring problems for humans in the Pleistocene environment, and solving them had the effect that the genes of those who solved them stayed in the gene pool, it might well be that mechanisms have evolved that enable or cause behavior that is apt for solving those problems.

Based on this, we would expect two factors to predict how specialized those mechanisms are:

1. The extent to which solving those problems would enhance the evolutionary fitness in each instance that problem was solved, and

2. The rate in which those problems would occur over many generations of humans. [47]

The more important the solution of an environmental problem was for procreation, and the more regularly that problem occurred during many generations, the more specialized we would predict the mechanism to be that has evolved to solve that problem (see also Cosmides and Tooby, 2006, p. 63). A combination of both is the salience of a problem for the overall survival [48] for a population. The more salient such a problem is, the more well-functioning and therefore specialized Evolutionary Psychologists would expect the psychological mechanisms to be that 'help' to solve that problem.

A snakebite, for instance, would possibly kill the person that has been bitten, thereby severely impairing their chances of procreating or helping kin to survive and procreate. [49] Something less fatal like eating a plant which causes a person to become sick (this is an imagined example) would maybe lead to something like the person not being able

[46] Kurzban uses the legs of Olympic Games aspirant Oscar Pistorius (who is now more famous for killing his girlfriend) as an example: He has prosthetics from his knees downwards and a pair of legs designed for running, his discipline. They make him an excellent runner. With those running prosthetics, however, it is very hard for him to walk because they are especially designed for running and thereby not very suitable for walking. Human non-mechanical legs might be less perfectly suitable for running but better fit for walking than Pistorius' running prosthetics: They are less specialized and therefore a trade-off between walking and running ((Kurzban, 2012, p. 42/43), referring to: (Bramble and Lieberman, 2004)).

[47] See also (Geary and Huffman, 2002, p. 668), who make the point that brain mechanisms that respond to problems which are caused by stable features of the environment can be expected to be less flexible than those brain mechanism that respond to problems caused by changing features of the environment. They explain this lack of flexibility as the existence of structures that are already specialized to solve those steadily recurring problems. Changing features of the environment, in contrast, would lead to flexible structures that adapt to the respective mode of the environment. Those stable features would thus pose recurring problems (as they constantly have the same properties, including the problem-posing properties) and, as predicted above, lead to a less flexible brain mechanism that is therefore very good at solving those problems (specialized) but less flexible in solving other problems (again, specialized).

[48] For the sake of simplicity, I will in the following text imply that survival in the context of evolution means having an increased number of offspring that survives or promoting the procreation of kin with a similar set of genes.

[49] The snake example has been extensively cited in the Evolutionary Psychologist literature, see e.g. (Buss 1995, p. 6).

to get a sufficient amount of food for two days, thereby increasing their risk of death (e.g. by starving, additional illnesses due to an impaired immune system, falling prey to animals because they is not able to fight,...). It would, however, not pose as big and immediate a threat as a snakebite by a very venomous snake. So one would expect the mechanisms against that mildly poisonous plant to be less specialized than the mechanisms against snakes: In order to survive, it is much more important to behave in a way that protects from snakebites than in a way that protects from eating the sick-making plant because the risk of death for someone who gets bitten by a venomous snake is higher than their risk of death from eating the plant. This is why Evolutionary Psychologists would predict that the mechanisms that lead to behavior that reduce the risk of being bitten by a deadly venomous snake are more specialized than those that reduce the risk of eating the sick-making plant: "[...] the more important the adaptive problem, the more intensely natural selection specializes and improves the performance of the mechanism solving it." (Leda Cosmides and Tooby, 1994, p. 89).

If there were only very few venomous snakes, or if there were only venomous snakes for a period of 100 years, [50] but a large amount of that mildly venomous plant everywhere that humans lived during the entire Pleistocene, one might expect psychological mechanisms against those plants to be more specialized: The problem of avoiding those sick-making plants would then be far more urgent to solve as they would pose a recurring threat. Even if eating them would not pose an immediate death threat, if people would frequently and over many generations eat them and thereby increase their risk of dying, while individuals or lineages of individuals with a working mechanism against eating those plants would have a higher change of surviving. As specialized mechanisms function better in solving problems than more general mechanisms (see above), the expected outcome would be that a more specialized mechanism against eating the plant is selected for.

To sum up: Mechanisms that are specialized in solving certain problems are better at solving them. If the solution to a certain problem was very important to our ancestors' survival in the Pleistocene environment, today's humans can be expected to have a very specialized psychological mechanism that solves this problem.

This feature of specialization in the paradigm of Evolutionary Psychology is not only important to determine which degree of specialization is predicted. According to massive modularists (which I hope to have shown constitute at least a notable proportion of Evolutionary Psychologists), the fact that specialized mechanisms are more successful in solving tasks leads to the assumption that the brain is made up entirely of modules that serve the function of solving different survival and reproduction problems that arose during the Pleistocene. Kurzban writes:

[50] The following argumentation only holds if during those 100 years, venomous snakes would not have killed the entire population apart from a few individuals who are endowed with a psychological mechanism that lowers the chance of being killed; if that was the case, the remaining individuals would possess a very specialized anti-snake mechanism.

> ”Whenever something useful is made, whether through human artifice or the process of natural selection, it must assume some form that enables it to carry out its task. There are no general-function artifacts, organs, or circuits in the brain because the concept itself makes no sense. [...] Well, yes, it seems to me that everything designed is designed to do something in particular [...].”(Kurzban, 2012, p. 41)

This also applies to domains of social behavior which I will take to include cognitive systems for the computation of Trolley Dilemma scenarios:

”[T]here is no reason to • that specialization applies to everything we have ever discovered or created that has some function *except* the parts of the human brain designed for social behavior. [...] Yes, these might be very complicated functions, but the logic of solving these social problems is exactly the same: efficiency from specialization.” ((Kurzban, 2012, p. 39), emphasis by author; referring to (Symons, 1979), (Tooby & Cosmides 1992)).

This sounds like the very strong claim that the entire brain consists of functional modules that either are perfectly suited to serve urgent survival problems that humans have met in the Pleistocene or a little less specialized for less urgent problems, all of them ‘designed’ by nature to fulfil certain tasks. This does not leave space for behavior that would not have served any purpose whatsoever in a Pleistocene environment. To assume that this is what Evolutionary Psychologists say is to build up a straw man or at least to confine the field of Evolutionary Psychologists to a very few extremists.

Even within the paradigm of Evolutionary Psychology and Massive Modularity, we can find some ways to explain behavior that is not very useful: There are different reasons for those systems to be ’buggy’. [51]

The first reason could be constraints on the working of those modules that impact their ability to fulfil their ‘purpose’ perfectly. [52] The following list does certainly not include all of those constraints but can explain some of the ‘shortcomings’ that evolutionarily adapted functions might have, according to Evolutionary Psychology.

Firstly, as already mentioned, in less specialized functional modules there have to be trade-offs. To use Kurzban’s analogy again, legs are ‘designed’ for walking and running. Therefore, prostheses especially designed for running might be better at this task than original human legs but are less suited for walking with slowly: They are specialized for running and hence they are deficient when it comes to walking. Transferred to psychological mechanisms, this means that if a psychological mechanism serves a broader function, there have to be trade-offs when it comes to the single ‘tasks’ included in this broader function.

Let us assume (and this again is a fantasy example, inspired by (Barrett, 2005a, p. 279) and his ’predator lookup table device’ that can ’tag’ animals as predators), a mechanism

[51] In the sense of a ’bug’ in a computer program.

[52] Please excuse my teleological language here; it is, again, for the sake of simplicity and I presuppose that most readers agree with me that adapted traits have not been designed, but selected for.

is ‘designed’ to deal with dangerous animals but is not specialized for certain kinds of animals. It would lead us to react to encounters with crocodiles in just the same way as we react to encounters with venomous snakes. With crocodiles, it might be wise to run away whereas with venomous snakes, the snake is probably going to be quicker anyway and staying still would be the best strategy to survive. So a mechanism built to survive animal attacks would either send us running and screaming or make us stand still, and neither reaction would be ideal solution to both kinds of attacks. [53] Let us for illustrative purposes assume that the mechanism would respond to an input such as ”dangerous animal” with a flight reflex. A mechanism ‘designed’ to protect us against animal attacks would then lead to behavior that would increase survival chances against those kinds of animals that do not move too quickly but be counterproductive against animals that react to quick movements with an attack and move quicker than humans. This behavior would then be the result of a broader-‘purpose’ mechanism and be successful in many cases, but counterproductive in some other cases. This behavior shown towards snakes would then be a less-than-perfect reaction based on a mostly useful psychological mechanism (see also Machery, In press, online manuscript on http://www.pitt.edu/~machery/papers/Discovery_and_Confirmation_in_Evolutionary_Psychology_FINAL.pdf p. 9).

Secondly, a mechanism might not give rise to the optimal behavior in a situation simply because no mutation ever produced the mechanism that would have led to optimal behavior (or the individual carrying it did not procreate despite their evolutionary advantage in form of the mechanism). Kurzban cites insects of the species ”Phyllium celebicum” that are shaped like a leaf and therefore hard to identify when they sit surrounded by leafy shapes (Kurzban, 2012, p. 31). The leaf form is supposedly a result of the fact that it was hard for predators to spot animals of that shape amongst leaves. There might be a better way of hiding for an animal, such as adapting the form and colors of the ground it sits on so it can hide on any type of ground (as some octopuses do) (Kurzban, 2012, p. 34), but this might just not have been one of the random mutations one of the animals ancestors underwent. So the mutation that would have brought up the perfect solution was just never in the gene pool and could therefore not be selected for. [54] Thirdly, there could be constraints by the architecture that is already given. Kurzban compares this situation to

[53] Such a reaction is not particularly far-fetched. Actually, I could well imagine people jumping and screaming when seeing a snake and have observed the same reaction several times when people discovered spiders nearby. Anyway, this example is only for illustration purposes. Arguably, venomous snakes and dangerous crocodiles will provide entirely different optical inputs due to the sizes in which an average individual of each species would be dangerous to humans. Modules might react to optical input and not to categorical input, as has been suggested by Buss (”The fear is triggered only by a narrow range of inputs, such as long, slithering, [sic] organisms perceived to be within striking distance.” (Buss 1995, p. 6)) and therefore crocodiles and (mostly small) venomous snakes could be unlikely to have the same functional module to react to them. Reflex reactions could then be entirely different to both animal species. Another kind of trigger could however be the conceptual input of ”dangerous animal” or the more immediate input of ”unexpected motion”.

[54] In Kurzban’s example, it might have been too energy-consuming to be selected for.

the way new buildings in a city are constrained by ”the layout of the buildings and streets that are already there.” (Kurzban, 2012, p. 35).

To sum up, ”[...] while it’s perfectly reasonable to expect circuits in the brain to be well engineered to solve adaptive problems faced by our ancestors, it’s also reasonable to expect any number of imperfections, suboptimalities, and so on [...].” (Kurzban, 2012, p. 35). The fact that not every behavior produced by a psychological module perfectly serves a function is not only part of the Evolutionary Psychologist story, but will become more important later when it comes to the difficulties in individuating different psychological modules in terms of the functions they serve: How can we tell apart a ’bug’ and a ’feature’? How can we decide whether some behavior is due to constraints in the building of the psychological mechanism causing/enabling it or whether it is part of the solution to an environmental problem? And how can we deduct from several clusters of behavior which psychological mechanism is responsible for it and whether and if so which function it served originally?

One of the main presupposition of Evolutionary Psychology is that the brain consists of a conglomerate of mechanisms that have been evolutionarily selected for and therefore serve a function. This function is to cause or enable behavior that leads to enhanced chances of reproduction for the individual endowed with it or to enhanced chances of reproduction for the individual’s kin. Those mechanisms are functionally modular.

Note that Functional Modularity and Mechanistic Modularity are not the same, even if argumentations for them often co-occur: The modularity of Mechanistic Modules lies in their flow of information, while Functional Modules are modules because they evolved to serve a certain function and no other. The question Mechanistic Modularity answers is: How does it compute? The question Functional Modularity answers is: How did it evolve? What was its function? The answers to those questions if the mechanisms are modular are: Informationally encapsulated for mechanistic modules, and because it mutated and solved environmental problem X for functional modules.

Although, as shown above, mechanistically modular computation in the sense defined in the chapter about Mechanistic Modularity has several evolutionary advantages, functional modules do not necessarily compute in a mechanistically modular way.

Various methods can be used to test whether something is mechanistically modular or functionally modular: For Mechanistic Modularity, you have to test which information is accessible to which part of the mechanism (e.g. whether mechanisms from other domains influence computation in the presumably modular domain, as we did in the chapter about Mikhail’s Mechanism); for Functional Modularity, you have to test the form/function fit in general terms or see whether there is evidence that something cannot have been learned (Poverty of Stimulus arguments) and must, hence, have been genetically implemented.

The concepts are dissociable: We can find Functional Modularity without informational encapsulation, as in Barrett and Kurzban who proclaim “that a broader notion of modularity than the one Fodor advanced is possible: In particular, a modularity concept based on the notion of functional specialization, rather than Fodorian criteria such as automaticity and encapsulation." [55] (Barrett and Kurzban, 2006, p. 628/629).

And Dwyer and Hauser agree that the "moral faculty" might not be modular in the Fodorian sense, e.g. informationally encapsulated or inaccessible to explicit moral justification mechanisms: They and others (Dwyer et al., 2009, p. 499), based on (Cushman et al., 2007) had argued for informational encapsulation of the moral judgement mechanism on the grounds that many people do not give coherent answers to the question why they found one case permissible and the other impermissible, hence the judging operation seems to be inaccessible to conscious reasoning. When Jacob and Dupoux refute that, arguing that explicit reasoning is part of the moral judgement mechanism too and hence conscious reasoning must have access to at least some moral judgements, (Jacob and Dupoux, 2007, p. 374/375), they reply: "Hence, [the] L[inguistic]A[nalogy] is not undermined, even if [the] M[oral]F[aculty] is not a module in the Fodorian sense." The context, mentioning that it is not module "in the Fodorian sense" as well as the fact that innateness is an important feature of Chomsky's linguistic theory and Hauser's concept of a "moral faculty" (Hauser, 2006), gives reason for interpreting this citation as the claim that the "moral faculty" might not be modular in the Fodorian sense, and might not be informationally encapsulated, but is modular in some other sense, possibly our "functional modularity" sense.

As we can see, both kinds of modules are domain-specific in different senses: Mechanistic modules compute domain-specifically; they compute, from input to output, only using the mechanisms for a certain domain: A domain-specific moral mechanistic module, for example, would not use the same mechanisms (and algorithms, including the same biases) as a module that computes economical costs. A functional module might use the same mechanisms for parts of its computation as used by other mechanisms, as Barrett and Kurzban 2006 propose. Confer et al. agree:

> "Psychological adaptations are not separate 'modules' in the Fodorian (Fodor, 1983) sense of informational encapsulation; rather, they often share components and interact with each other to produce adaptive behavior." (Confer et al., 2010, p. 111) also referring to (Barrett and Kurzban, 2006).

Functional models have evolved to compute certain kinds of (proper) inputs so the (behavioral) output is evolutionarily advantageous. Barrett and Kurzban write: "As a direct and inseparable result of this evolutionary process of specialization, modules will become domain specific: Because they handle information in specialized ways, they will

[55] And, in a trivial sense, we can find informational encapsulation without functional modularity: Imagine, for instance, a non-animated mechanism that does not fulfil any function and does not exchange any information with the other components of this useless device. Both non-functionality and informational encapsulation would be given.

have specific input criteria." (Barrett and Kurzban, 2006, p. 230). This domain specificity does not lie in the mechanisms that compute the input but in the input constraints and the functionality. Barrett and Kurzban write: "[...] we define domains as individuated by the formal properties of representations because, we believe, this is the only possible means by which brain systems could select inputs. As a corollary, by virtue of the fact that formal properties determine which inputs are processed, a mechanism specialized for processing information of a particular sort can, as a by-product, come to process information for which it was not originally designed [...]." (Barrett and Kurzban, 2006, p. 630). [56] - The genes for a moral module, for instance, would have stayed in the gene pool because it optimally computes moral situations; it might, however, also compute social situations that have formally similar inputs and (since modular mechanisms can "share components and interact with each other", as I have cited Confer et al. 2010's view above) thereby use mechanisms that have developed to compute food trade-offs.

Tooby and Cosmides agree that they are not committed to Mechanistic Modularity, because "[t]he criteria for calling a device module are inconsistent and vague (some view information encapsulation as criterial; others emphasize specialization, etc.), especially when compared to the crisp criteria for calling a device an 'adaptation'." (Cosmides and Tooby, 2006, p. 63). They, however, commit to Functional Modularity. [57]

[56] If you find the definitions here inconsistent, you are not alone, but I hope they are not caused by my representation of the distinction between functional and mechanistic modules. Quite a few inconsistencies arise with the concept of a mechanism that has developed to compute inputs of a certain domain advantageously and that is not exclusively reserved for those operations; for further criticisms of this position, see (Dahlgrün, 2015, p. 47 ff.) and (Grossi, 2014, p. 15 ff.). To name one of the problems: "Systems can accept only specific types of input, but they can process information for which they were not originally designed. According to B[arrett] and K[urzban], the fact that a system is activated by other types of input (that is, lacks domain specificity) does not undermine its modularity." (Grossi, 2014, p. 17).

[57] criteria for calling a device module are inconsistent and vague (some view information encapsulation as criterial; others emphasize specialization, etc.), especially when compared to the crisp criteria for calling a device an 'adaptation'." (Cosmides and Tooby, 2006, p. 63). They, however, commit to Functional Modularity.

3 Are Trolley Dilemma Judgement Mechanisms Functional Modules? - Evolutionary Psychologists' main evidence re-evaluated

As I have shown, many of the methods that can empirically test whether something is a mechanistic module are useless for functional modules because the latter do not have to accord to any restrictions regarding the computational mechanisms they use and only to very broad input restrictions ("[...]systems specialized for speech perception process only transduced representations of sound waves [...]" (Barrett and Kurzban, 2006, p. 630), to cite one of the vaguest examples). [58]

Hence, from an Evolutionary Psychologist point of view, we can gain insights and come up with testable hypotheses about Functional Modules mainly via two methods: 'Reconstructive Engineering' and Reverse Engineering.

I will first elaborate on those ways of testing Evolutionary Psychologist hypotheses. After that I will explain why they are problematic, especially when applied to the social domain or domains of human interaction. This is relevant for Trolley Type cases because I take them (as they test for rules for human interactions) to be part of this domain and those problems apply to them as well. Next, I will expound what could be further criteria that indicate that a psychological mechanism has developed as an evolutionary function and after that show why those criteria are not fulfilled in Trolley Dilemma judgement mechanisms.

What I call 'Reconstructive Engineering' [59] has the following rough form:

> Ancestral Problems → What Kind of Behavior Would Solve them? → Psychological Mechanism → Prediction: Complex of Behavior, Parallel Behavior in Apes, Hunter/Gatherer Societies.

To illustrate this idea, I will adopt Robert Kurzban's example of a toaster (Kurzban, 2012, p. 33). He uses it to make the idea of backward engineering clearer, which I will also do in the next chapter, but it can also serve the purpose of an analogy for forward (or reconstructive) engineering. In the case of forward engineering, we knew someone wanted toasted bread and needed something to toast it. Maybe they have had some means to toast the bread before, such as an open fire, but these did not work optimally because

[58] Barrett and Kurzban 2006, referring to (Sperber, 1994), define 'proper' domains as the domains something actually evolved to process and 'actual' domains as the domains something can compute; "the proper domain of a face recognition system would be, putatively, faces of conspecifics [...]. The 'actual domain' might be [...] perhaps not only faces but the wider set of stimuli that have formal properties that cause them to be processed by the face recognition system." (Barrett and Kurzban, 2006, p. 631).

[59] I call it that because we could analogize the process of natural selection to engineering and therefore the reproduction of this process is a kind of 'Reconstructive Engineering': We try to reconstruct the starting conditions of the process, the features that have been selected for and predict the end state, the complex of behavior that has been selected for.

the means were not made especially for toasting bread. The bread would burn sometimes and you would have to turn it over in order to toast the other side as well. So they would try to engineer something that particularly fits the problem of toasting bread: Something that roasts slices of bread regularly from both sides. These are the functions we would predict the toaster to serve. The next task is to reconstruct what would have been built by someone who wanted to invent a device to toast bread. It should have something like one or two slots fitting a slice of bread and a heating device that roasts both sides of the bread when it is inserted in those slots. In our analogy, we have not found the toaster yet, but are looking for something that serves those functions. If there is something that seems to be like what we predicted and does not serve a different function better this will be evidence that it was developed to toast bread. If it serves a different function better it has probably been constructed to serve that function. If it has, for example, all the toaster features we predicted but you can also drive in it, it might be a car that accidentally has parts that heat up and are shaped in a way they can hold toast. In that case, and considering it is inconveniently big to put it in your kitchen, we would conclude that it has been built to drive in, not to toast bread.

Another reason why something might have features that are needed in a toaster but obviously does not serve the function of toasting bread for the purpose of eating it, would be something that has side effects that would make it unable to serve this function. If something had slots that are bread-shaped and heat up but was very radioactive and would contaminate every single slice of bread that is put in it, we would probably conclude that this item, tool, was not made for toasting bread for human beings to eat.

In the Evolutionary Psychologist kind of 'Reconstructive Engineering', the researchers start with knowledge or assumptions about evolutionarily salient problems or tasks in a Pleistocene environment (in the analogy, the problem was that someone had bread and wanted it to be toasted). Those kinds of problems or tasks might be difficulties connected to procreating, raising offspring, surviving, helping kin to survive etc. According to Evolutionary Psychology, the kind of behavior that leads to a large number of progeny or kins' progeny over many generations is selected for.

After having identified such a problem such as the problem of surviving encounters with venomous snakes, the researchers then infer what kinds of behavior would solve those problems and make predictions about that kind of behavior (in the analogy: The features the toaster should include one or two slots and a heating compartment). In the case of snake encounters, they might for example predict that people who come across a snake freeze even if they have never seen a snake before. [60]

[60] A weaker claim would be that humans easily learn fear of snakes (just as fear of darkness and fear of heights seem to be learnable more easily than fear of cars) [(Buss, 1995, p. 8), referring to (Marks, 1987)). In that case, the range of behavior that is predicted would be broader (any kind of fear reaction), and the fear would need an environmental trigger to be learned and would not develop without such a trigger. As it happens, monkeys seem to learn fear of snakes more easily than they learn fear of, say, artificial flowers

If they actually find that kind of behavior, that would be evidence for their claim that it has developed as a solution to that ancestral problem.

If the behavior does not solve the problem properly, this would be counter-evidence against its having evolved to solve that problem. If you, for instance, hypothesize that people needed some behavior to draw inferences and predict things about the world so they can react appropriately, which would then enhance their chance of survival, and that this behavior could be (conscious) reasoning, but this reasoning "leads to epistemic distortions and poor decisions" (Mercier and Sperber, 2010, abstract), this would mean that the behavior probably is not caused by a mechanism that evolved for this task. Mercier and Sperber instead infer from this and other properties of reasoning that it probably evolved for "the production and evaluation of arguments in communication" (Mercier and Sperber, 2010, p. 58). As in our example with the car that was better suited for driving than for toasting, they argue that they found a problem that reasoning solves better than making predictions about their general environment to be able to anticipate, for instance, dangers: According to them, it is better suited to persuade people in communication and evaluate their arguments. This discards the prediction hypothesis and replaces it with the communication hypothesis.

Hence, one way to develop Evolutionary Psychologist hypotheses is to make predictions about behaviors based on knowledge about the ancestral environment and salient, recurring problems in this environment and possible behavioral solutions to them and then identify complexes of behavior that match the predictions about those behaviors. If most humans do not behave as predicted, the hypothesis was wrong.

However, some issues arise with this kind of hypothesis testing: Many ancestral problems have many fitting behaviors that would solve them, leaving space for many different kinds of behavior to count as a solution to that problem. This opens up the space for possible evidence for Evolutionary Psychologist hypotheses in a way that might seem too wide to be appropriate for making a hypothesis more credible. This holds especially if we assume that the solution that has developed is not necessarily the ideal solution to the problem. Let's say that we found that people do not freeze when they come across snakes for the first time. What if they run away when they see a snake? This could then be evidence for a broader mechanism as proposed before, one that leads to the same reactions towards snakes and crocodiles. Or what if people act differently and less appropriately when they see a snake for the first time than when they see one for the second time, but all of those reactions are fear reactions? One could then say something similar, namely that there is a psychological mechanism that reacts to snakes just as it does to crocodiles, with fear, but that fear has been selected for. Or people react with fear to snakes after they have seen someone who was scared of a snake once in their lives, whereas they did

((Leda Cosmides and Tooby, 1994, p. 106), referring to (Mineka and Cook, 1988)); for counter arguments against the empirical evidence for innate fear of snakes, see (Grossi, 2014).

not react with fear after they saw someone who was scared of a flower. [61] We could then take this as evidence for a learning mechanism that is triggered if the individual endowed with it witnesses someone else being afraid of snakes.

Having to change the predictions so many times or even the need to come up with new ‘predictions’ that fit the evidence would weaken the case for the hypothesis. That there is some kind of behavior that fits the original problem is still evidence for a psychological adaptive mechanism. There should, however, be more evidence to enforce a ’Reconstructive Engineering’ hypothesis than just the fact that the scientists found some solution that fits the problem: Testing whether the form of the ancestral problem fits the function of the psychological mechanism or the behavior that it gives rise to is one mode of exploring adaptive functions, but the sheer quantity of behavioral solutions that might fit one evolutionary problem is one reason why there should be different ways to test whether some behavior is actually caused by a psychological mechanism ‘built’ to solve an ancestral adaptive problem.

This is especially the case when the predictions correspond to the kind of observations that we can make every day: What if researchers make predictions that fit their pattern of everyday observations? Many sciences use ’Reconstructive Engineering’ or Reverse Engineering or the fit between design and task to explore about functions or build up design features. To cite just a few examples: One could use ’Reconstructive Engineering’ to find out new ways of solving technical problems: Engineers could (and sometimes do) use biological traits, for example in animals, to solve technical problems. First, engineers come up with a problem (for example, how can we fly energy-efficiently). Then they come up with animals that might have an isomorphic problem to solve in their environments (flying without using much energy). They can then identify features in those animals that might fit the problem and use it to solve their own (for example, a shape with particularly low air resistance).

Pinker 1997 (Pinker, 1997, p. 21) has cited examples of Reverse Engineering: When Sony developers see a new product by Panasonic, they wonder what problem it solves and how which part of the design solves which problem and how those parts can be constructed. Another example is the reaction of many people when they face a new gadget: They wonder what it is made for and try which task it serves best. The third example comes again from biology (Pinker, 1997, p. 22): ”William Harvey discovered that veins had valves and deduced that the valves must be there to make the blood circulate.” All of those seem to be successful instances of ’Reconstructive’ or Reverse Engineering. But there are some features they do not have in common with psychological phenomena:

- They are not part of a world where social interactions play a very important role.

[61] Which is, as mentioned above, the case in monkeys but no experiments with humans have been made as far as I know.

- The problems they solve are directly observable in the environment and not part of an ancestral environment we cannot know everything about. [62]

And, most importantly for the following argument: Few phenomena are as easily and frequently observable (and observed!) as human behavior. We cannot help but watch human behavior every day (as I did when I stated in a footnote that I saw people screaming and jumping at the view of spiders). This is a feature that complicates the detection of new psychological functions: What if researchers use their own everyday empirical knowledge and predict exactly what they have experienced? This poses a problem because the prediction should at least to a large part be led by the Pleistocene problem that is to be solved by the behavior and not by previously observed behavior. In 'Reconstructive Engineering', ideally, the scientists assume that people frequently encountered a certain survival or procreation problem in the Pleistocene; the prediction about what kind of behavior might be apt to solve the Pleistocenic problem should be based solely on the solution to this problem.

Less ideally, it could be that the scientists (consciously or unconsciously) observed a certain kind of behavior and are led by this behavior when they hypothesize about the behavior that should solve the ancestral problem (or even when they reconstruct ancestral problems in a past world whose properties they can only infer today). The fact that the predicted behavior can actually be found is then explainable rather due to the fact that it has been observed before than due to the prediction being correct.

This is not a problem that generally disqualifies a method; it is common to many ways of psychological hypothesis-building that they might be based on everyday observations and can be solved by examining further features that come with psychological adaptive functions . In the worst case, however, most of the Pleistocenic problem scenario itself was influenced by those everyday observations. But the problem scenario is the basis for the predictions and the predictions are those everyday observations that the problem scenario was based on in the first place. This process exhibits a circularity that usual psychological hypotheses lack for the following reason: In (non-evolutionary) psychological research, the researchers usually develop a setting and test for behavior in this setting.

To demonstrate this with a concrete example: In the famous bridge experiment (Dutton and Aron, 1974), Dutton and Aron wanted to test whether "strong emotions are relabeled as sexual attraction whenever an acceptable object is present, and emotion-producing circumstances do not require the full attention of the individual." (Dutton and Aron, 1974, p. 511). They accordingly developed a setting where an "attractive" interviewer interviews people on two bridges: One precarious bridge "constructed of wooden boards attached to wire cables" with "a tendency to tilt, sway, and wobble" and "a 230-foot drop to rocks and shallow rapids below "(Dutton and Aron, 1974, p. 511) that, as they tested,

[62] I will return to those two differences later.

induced fear, and a more solid one that induced less fear. Part of the interview task was to write “a brief, dramatic story based upon a picture of a young woman covering her face with one hand and reaching with the other” (Dutton and Aron, 1974, p. 511), a measure for sexual arousal depending on the story’s content (Dutton and Aron, 1974), referring to (Murray, 1943). After the interview, the interviewer gave the respective subjects her number for further questions. Dutton and Aron found that the subjects on the shaky, anxiety-inducing bridge both wrote short stories that scored higher on sexual arousal and called the “attractive” interviewer back more often.

Back to our meta-level: The short stories and calling the interviewer are the behavioral component. The correlation between features of the setting (anxiety-inducing bridge and non-anxiety inducing bridge) and the behavior confirms the prediction that people will score higher on sexual arousal and call the interviewer more often when they are on anxiety-inducing bridges. The explanation is that, as hypothesized, the people misattributed their emotions (fear) induced by the shaky bridge as attraction towards the interviewer. Being scared caused them to feel attraction towards the interviewer.

Here is a short list of components of the experiment and their function.

- Setting: Bridges
- Parameters: Shakiness
- Correlation: Between shakiness of the bridge and erotic content of short stories/calling the interviewer
- Explanation: The shakiness causes anxiousness in the subjects that they misinterpret as attraction towards the interviewer
- Hypothesized Causation: Anxiousness → Attraction

Now let us draw an analogy to an Evolutionary Psychologist experiment. I have chosen an experiment by Buss (Buss, 1989) because it is not very complex and, to me, a good example of questionable Evolutionary Psychologist research. Although the article was published in 1989, Buss still frequently cites it in his own publications and treats the results as shown facts (Confer et al., 2010), (Conroy-Beam et al., 2015), (Buss and Shackelford, 2008); I hence assume that he still considers this research to be valid. The article is entitled “Sex differences in human mate preferences” and it is an example of ’Reconstructive Engineering’ in Evolutionary Psychology. Buss came up with the following ancestral environmental problem, based on parental investment and sexual selection theory: “Males may provide mates with food, find or defend territories, defend the female against aggressors, and feed and protect the young. Human males may also provide opportunities for learning, they may transfer status, power, or resources, and they may aid their offspring in forming reciprocal alliances.” (Buss, 1989, p. 2), referring to (Trivers, 1972, p. 142).

This, together with some assumptions about sexual selection and resource accessibility and translating resources into earning capacity, led him to the following predictions: “Females, more than males, should value attributes in potential mates such as ambition, industriousness, and earning capacity that signal the possession or likely acquisition of resources.” (Buss, 1989, p. 2). Please note the wide nature of the environmental problem he described: Although Buss lists the possible male contributions like a conjunction (Buss uses commas and ’ands’ to connect them), Buss treats it like a disjunction by only translating the part about feeding and transferring resources into his predictions, the parts connected to the “possession or likely acquisition of resources”. [63] Apart from the sheer number of things on his list, the content is questionable as well: In the 1972 paper by Trivers that Buss cites in this passage, Trivers only claims that the sex that is investing more in the offspring biologically (in humans’ case, the female by being pregnant for a long time) is going to prefer mates that invest in the offspring as well. The first part of Buss’s list is derived of the following list by Trivers: “A male may invest in his offspring in several ways. He may provide his mate with food as in baloon flies (Kessel 1955) and some other insects (Engelmann 1970), some spiders, and some birds (for example, Calder 1967, Royama 1966, Stokes & Williams, 1971). He may find and defend a good place for the female to feed, lay eggs or raise young, as in many birds. He may build a nest to receive the eggs, as in some fish (for example, Morris 1952). He may help the female lay the eggs, as in some parasitic birds (Lack 1968). The male may also defend the female. He may brood the eggs, as in some birds, fish, frogs, and salamanders. He may help feed the young, protect them, provide opportunities for learning, and so on, as in wolves and many monogamous birds.” (Trivers, 1972, p. 141/142)

As you can see, Buss did some stretching to ’translate’, when he, for instance, transforms “he may defend a good place for the female to feed, lay eggs or raise young” into “[m]ales may [...] find or defend territories”. But this shows only part of the randomness of his claims. He also notably does not incorporate an analogon to “brooding the eggs”, “build a nest” or “help the female lay the egg” which would be possible at least for “helping the female lay the egg”, namely: Playing the part of a midwife (or, that is, spouse). In his specifically human contributions, he could have added things like cooking, washing the baby and watching the child (the latter might be implied by “protecting” the offspring). All Trivers’ theory predicts is that the male invests in raising the offspring. Parental investment could mean anything. Nor does Buss refer to archaeological science when it comes to explaining why he chose this and no other list of male contributions.

We could, of course, say in his favor that he implicitly assumed a society where men hunted and women gathered and did the daily care for the children and therefore did not include any contributions of care for the family into his list. But even if he had had these

[63] If it was a conjunction, Buss would have had to show that the male does all these (and the female selects the males who provide all these).

hunter-presuppositions, it would have, according to Brumbach and Jarvenpa 2006, been a biased interpretation of archaeological findings: "The division of labour was highly variable and more flexible than commonly assumed, both within and across populations. There was no rigid or universally applicable "man the hunter/woman the gatherer" protocol, even with respect to the narrower scope of food procurement (i.e., ignoring food processing, storage, and distribution). Indeed, division of labour occasionally followed lines of age, ability, and experience, among other factors, rather than gender per se." (Brumbach and Jarvenpa, 2006, p. 524)

But not only is his list of contributions selective; Buss forms his predictions out of a subunit of the list, the one connected to goods, and translates this into material wealth nowadays, leaving aside things like "aid their offspring in forming reciprocal alliances" (Buss, 1989, p. 2). However, as unspecific Buss's formulation of the problem is, so are results as widely interpretable. Notably, the four most important things people in 37 countries sought in mates were the same in both men and women: Mutual Attraction/Love, Dependable Character, Emotional Stability and Maturity and Pleasing Disposition (Buss and Abbott, 1990, p. 19). Furthermore, Buss et al. stated that "the effects of sex are substantially lower than those for culture" (Buss and Abbott, 1990, p. 17), despite the title "Sex differences in human mate preferences". As the last sentences indicates, they, however, confirmed their predictions with, for instance, "good financial prospects" in 12th place in women and in 13th place in men (Buss and Abbott, 1990, p. 19), collection of findings: (Ruti, 2015, p. 57).

So why are the predictions of one study so much more problematic (Buss, 1989) than the predictions of the other study (Dutton and Aron, 1974) although they seem to make similar predictions, one about the correlation between being female and wanting a mate with capacities connected to economic success and the other between being on a shaky bridge and feeling sexually aroused?

It is because the two cases may not be as comparable as they seem. Let us make the same chart about the experimental components as with the Dutton case above.

- Setting: Pleistocene environment (with "problems": List of possible male contributions)

- Parameters: None

- Correlation: Being Female and reporting capacities that are connected to economic success as desirable in mates

- Explanation: Because it was evolutionarily advantageous for males to contribute to raising the offspring, too.

- Hypothesized Causation: Pleistocene Environment → [teleological causation] genetic endowment that shapes the brain of women in a way that they tend to desire capacities that are connected to economic success in mates

The most important difference between this Evolutionary Psychologist experiment (and any experiment using 'Reconstructive Engineering', that is) and non-evolutionary psychological experiments is the following: The correlations are used to test different things. In the Dutton experiment, the correlation between being on a shaky bridge (the setting) and feeling sexually aroused (and behaving accordingly, calling the interviewer/telling more sexually loaded stories, the observed behavior) is explained by a "simple" (and proximate, as opposed to: teleological and distal) causation between the correlates: Being on the bridge causes the subjects (possibly mediated by a row of other 'simple' causations) to feel attracted to the interviewer. Both, setting and behavior, are directly observed and the setting is manipulated/behavior in different settings is observed. In our Buss case, the correlation between being female and preferring potentially successful mates is the observed behavior and the setting is a Pleistocene environment. The explanation for the behavior is that Pleistocene environment teleologically (and distally) causes (favors selection in a way for) women to favor mates with capacities that promise economic success.

This means that the setting lies in the past. We cannot directly observe it or change its parameters (as in the bridge experiment, where they used one shaky and one solid bridge). We can only make informed guesses about the nature of the setting that caused humanity to develop the way it is now. Based on this information, the entire experiment is vulnerable to the circularity I have mentioned before. Whereas in 'regular' psychologist settings, we can directly observe the setting and change it because we have direct access to it, here have to construct it based on our information in 'Reconstructive Engineering'. As we know very little about the Pleistocene environment, we could come up with a very broad spectrum of Pleistocene challenges. So where do we get those intuitions from? In our Buss case, I strongly suspect that the setting, including the environmental challenge, was already strongly influenced by his assumptions about today's societies. And, even more strongly, the selection of predictions was biased in that he picked items out of a long list of possible solutions to the environmental challenge of paternal investment.

To sum up, I have shown the following problem with 'Reconstructive Engineering': If the same behavior that is observed in everyday life is the basis for the construction of the evolutionary challenge (the setting) and/or the predictions (the behavior), the experimental process becomes circular and predictions become rather self-fulfilling (at least in the case that the everyday observation was correct). Unlike 'regular' psychological experiments, this setting cannot be directly observed and changed and freed from biases because it lies in the past. If the everyday observations are the basis for inferring the Pleistocene challenge, we are basically facing a case of Reverse Engineering: We explain a complex of behavior by assuming that it was the solution to an ancestral problem that we tailor-fit to the complex of behavior. But if we claim that this ancestral problem was identified independently and predict the complex of behavior we previously modelled our problem on, the process is obviously circular and will not yield any valid results (Compare Machery, In press).

Of course, it is possible in principle to gather information about Pleistocene settings by, for instance, employing archaeological findings such as arthritis in joints of skeletons that show which joints have been strained and which kinds of pursuits strain the joints like that (e.g., Brumbach and Jarvenpa, 2006, p. 522). Maybe symptomatically, in his more than 400-page-long monograph, Buss devotes a chapter only one-third of a page in length to archaeological “Sources of Data for Testing Evolutionary Hypotheses” (Buss, 2008, p. 63).

The trivial (but possibly not always implemented) solution for a clean experimental set-up is: The information used to reconstruct the Pleistocene setting should be as independent as possible from both predictions and everyday observations and as precise as possible. This, I argue, becomes increasingly difficult with settings that include social interactions as they are hardly archaeologically manifested and it is very difficult not to be biased when it comes to something as important and omnipresent in our lives as social interactions. [64]

This, however, is not the only problem with ’Reconstructive Engineering’ and social interactions. As David Buller points out,(Buller, 2006a, p. 98) another problem is that environmental challenges that are connected to social interactions are contingent on ancestral humans’ psychologies. As Buller puts it:

> “[...]the finer-grained adaptive subproblems faced by a species are not independent of the morphology and psychology of that species. [...] [W]ithout knowledge of the morphology and psychology of a species, we can never specify the adaptive problems confronting it with anything but the most coarse-grained descriptions, and these will not inform us of the specific selection pressures acting on it. So, in order to identify the selection pressures that helped shape human psychology, we would need to know something about ancestral human psychology.” (Buller, 2006a, p. 98)

It is, for example, only evolutionarily advantageous to develop a tendency to not harm people close-up physically when other people have a similar tendency or will not attack you for different reasons. As we do not (and probably cannot) know much about this ancestral psychology, we are constrained in inferring Pleistocene challenges that have to do with the social order, human interactions, violence etc. in the Pleistocene.

A different way to identify ancestral challenges is deducing them from observations of societies with people who hunt and gather today, which is common practice among Evolutionary Psychologists. If we know how secluded societies who hunt and gather live today, we can infer how they used to live in the Pleistocene because they encounter similar

[64] Another problem might be biased archaeology, see (Brumbach and Jarvenpa, 2006)

environmental and social circumstances as Pleistocene people who hunted and gathered encountered. Right? (see, for instance, (Hauser, 2006), (Barrett, 2005b), (Pinker, 1997) for Evolutionary Psychological research amongst contemporary Hunter-Gatherers [65]). Wrong, according to David Buller, and he has collected some rather convincing evidence to back this up: Hunter-gatherer societies are diverse. Even those that live in similar regions (and might face similar environmental challenges) vary to a notable degree.

Buller, referring to anthropologist Laura Betzig (Betzig, 1998), writes:

> "Among these populations, the average daily caloric intake from foods gathered by women ranges from 2 percent to 67 percent, average paternal care ranges from ten minutes a day to 88 percent of the day, and mating systems vary. [...] There is considerable such variation among African hunter-gatherers in and around the region Evolutionary Psychologists believe was inhabited by our ancestors." (Buller, 2006a, p. 95)

Buller raises the problem that, depending on which of those populations we use as a model for our reconstruction of ancestral environments, we are going to get very different environmental challenges. However, Edouard Machery argues that these variations exist but do not make it principally impossible to reconstruct environmental challenges from the Pleistocene: (Machery, In press; online manuscript version, http://www.pitt.edu/~machery/papers/Discovery_and_Confirmation_in_Evolutionary_Psychology_FINAL.pdf, p. 6/7). He believes that we can gain at least some information about behavioral tendencies in societies of people who hunt and gather by collecting data from multiple societies and evaluating them statistically. Machery assumes that the majority of those people lived a life that is similar to the life ancestral hunter-gatherers lived; this is a point Buller doubts as well, citing anthropologist Robert Kelly: "[...] long before anthropologists arrived on the scene, hunter-gatherers had already been contacted, given diseases, shot at, traded with, employed and exploited by colonial powers, agriculturalists, and/or pastoralists. The result has been dramatic alterations in hunter-gatherers' livelihoods....There can be little doubt that all ethnographically known hunter-gatherers are tied into the world economic system in one way or another; in some cases they have been so connected for hundreds of years. They are in no way evolutionary relics." (Kelly, 1995, p. 25/26), cited after (Buller, 2006a, p. 94)

Hence, not only have the circumstances in which people in those societies lived changed, meaning that the continuity between the Pleistocene and the contemporary lifestyle has become weaker (if there ever was any), but they have also been exposed to circumstances that Pleistocene hunter-gatherers have never encountered. But assuming Machery is right

[65] Albeit in some cases the authors refer to research about contemporary societies who hunt and gather to show the cross- cultural universality of traits and not primarily, at least not explicitly so, because of similar environments that could allow analogies to ancestral environments; but see, for instance, (Pinker, 1997, p. 499) for deductions from contemporary societies who hunt and gather to ancestral hunter-gatherers.

and a statistical analysis of representative contemporary societies of people who hunt and gather would provide information about Pleistocene hunter-gatherers (assuming the latter lived in more or less homogeneous societies that were similar to the majority of contemporary societies with hunting and gathering people), this method could indeed be very useful. However, not many Evolutionary Psychologists seem to use it; not even the article cited by Machery, who attests to being "sympathetic to evolutionary psychology" (Machery, 2008, p. 263), as an example of Evolutionary Psychologists working with hunter-gatherer analogies (Barrett, 2005b) reviews results from societies who hunt and gather systematically, at least not explicitly; he only mentions that "anthropological literature" often states that hunters in traditional hunter societies "anthromorphize" animals (Barrett, 2005b, p. 214), citing his sources but not putting them in a statistical context. His sources might have applied such a method; at least one of his two sources for data about animal attacks on humans, however, does not compare large sets of intercultural data, but only interactions in the last century in different areas of rural Uganda (Treves and Naughton-Treves, 1999). Machery, who could have made a point supporting evolutionary psychology, [66] only cites one source that compares societies of people who hunt and gather as he proposed (Kaplan et al., 2000). I infer that Machery's proposal, even if it works, has not been implemented often and Buller's criticism of analogizing the ways of life of contemporary people who hunt and gather with the ways of life of Pleistocene hunter-gatherers remains sound and applicable to the vast majority of Evolutionary Psychologist research at this point in time; and possibly it will always remain applicable because even if we can see statistical trends in the variety of contemporary hunting/gathering lifestyles, those trends are not necessarily the same as tens of thousands of years ago.

A third source of evidence for 'Reconstructive Engineering' is drawing inferences from challenges that animals face and their solutions to ancestral human environments. Buller and Machery agree that one of Evolutionary Psychologists' strategies is to "identify what adaptive problems organisms with specific characteristics would face and what traits might have been selected. That is, evolutionary psychologists look for generalizations linking the possession of specific characteristics to specific adaptive problems and to the evolution or specific traits." (Machery, In press, online manuscript, p. 5)

The arguments I have found to be most prominent are paralleling humans' and other species' behavior and checking whether Pleistocene humans and those cited species likely share the environmental features that are hypothesized to be responsible for their behavior. One example is Profet's work on pregnancy sickness (Profet, 1992). [67] Her hypothesis is

[66] Although, in his article, he rather argues for the possibility of good Evolutionary Psychologist research in principle but concedes that "the research done by evolutionary psychologists is of uneven quality." (Machery, In press, p. online manuscript version p. 22). Considering his collaboration with Barrett, (Machery and Barrett, 2006), (Machery, In press, p. manuscript, online version, p. 22) he should, however, consider him as one of the higher quality sources.

[67] Rather than being outdated, this work is now canonical in the Evolutionary Psychology community and, for instance, cited by (Buss, 1995), (Schmitt and Pilcher, 2004) and (Ermer et al., 2007) as exemplary,

that experimental herbivores and omnivores, including Pleistocene humans, are likely to develop pregnancy-induced food aversions to keep the mother from eating teratogens that can harm the fetus (Profet, 1992, p. 354/355).

By comparing other mammals with humans, she parallels the experimental Pleistocene setting (humans being experimental omnivores) with today's mammals' setting: Because she cannot go back in time, she looks for analogous backgrounds. If she can establish a connection between being experimentally herbivorous or omnivorous and having pregnancy sickness or pregnancy-induced food aversions against teratogens, this strengthens her hypothesis that the mechanisms that selected for pregnancy-induced food aversions were the same in humans as in the other observed mammals.

Now this may function with traits like experimental omnivorousness, because we can be relatively sure that humans were experimental omnivores and we can establish the link between omnivorousness in humans and teratogen avoidance by researching the link between omnivorousness in other mammals and their teratogen avoidance mechanisms. But when it comes to mechanisms of social behavior, we cannot parallel them to other species because in many cases we do not know what kinds of social interactions were part of the Pleistocene environment and which factors were important, and this is why we do not know how to replicate the ancestral experimental settings, hence, which species to choose to establish a link between ancestral social environment and today's behavior. We, for instance, might not know whether to use chimps or bonobos as models for ancestral social settings and today's behaviors (both of them are closely related to humans, but bonobos' canine teeth resemble proto humans' more because they are smaller (Waal, 2012, p. 874)) and hence, without further evidence, cannot argue whether humans have rather evolved an 'innate' tendency to be xenophobic like chimpanzees or are rather friendly and empathic towards everyone like bonobos and hence, which if any of those human behaviors has evolved as an adaptation to ancient social and environmental structures and challenges that might be similar to today's ape group structures. [68]

As Buller writes, "nonhuman primate species differ considerably with respect to foraging, parental care, and mating system." And, as Buller argues, how close a species is related to humans might be a heuristic to see how similarly they behave, but, according to Buller, the "degree of relatedness isn't correlated with similarity in behavioral traits." (Buller,

successful Evolutionary Psychological research project; even Ermer et al. praise them while in their article criticize many Evolutionary Psychologists' methods: "According to the tentative guidelines described earlier for evaluating the quality of evidence, the nomological network of pregnancy sickness as a psychological adaptation has both exemplary breadth and exemplary depth." (Ermer et al., 2007, p. 648).

[68] DeWaal ends up arguing that both in-group biases and empathy have evolved in humans because they make evolutionary sense, but this argument is additional to his cross-species observations that show both empathy and in-group biases (Waal, 2012, p. 876). Additionally, according to him, both are present in many species. But he does not give any answers as to whether humans are more like bonobos or more like chimpanzees; his whole argumentation just stresses that we might have underestimated empathy as is strongly displayed by bonobos when we concentrated on research on chimpanzees to draw conclusions about human behavior.

2006a, p. 96), citing (Gittleman et al., 1996) and (Böhning-Gaese and Oberrath, 1999). The sources Buller cites measure the relation between genetic proximity and similarity in behavior in birds, whales and primates, among many others. They, however, do not support his claim in all its generality, partially because they have very limited measures of behavior such as group size in Gittleman et al., where they define group size as "average number of individuals which regularly associate together and share a home range" and home range as "average total area (km^2) used by an individual (or group in social species) during normal activities" (Gittleman et al., 1996, p. 185/186). They, however, warn that "comparative studies analyzing behavioral traits should not presume phylogenetic relations [to the behavioral traits]." (Gittleman et al., 1996, p. 188). We should not just assume that genetic proximity is a measure for similar behavior in species. And in Böhning-Gaese and Oberrath's study regarding birds, "for most behavioral and ecological traits, relatedness explained less than 1% of the variation among species." (Böhning-Gaese and Oberrath, 1999 abstract).

This means that although Buller's formulation might be overly general, the cited scientists agree that we cannot just assume that species that are most closely to humans will show the most similar behavior. The group sizes, for instance, might be very significant for psychological mechanisms; they varied between 2 and 45 in the family of monkeys (ceboid monkeys) that Gittleman et al. tested (Gittleman et al., 1996, p. 188) and were not significantly correlated to phylogenetic descent (Gittleman et al., 1996, p. 186). A group with two members, however, would need very different psychological adaptations than a group of 45, e.g. in terms of collaboration or cheater detection, also widely discussed topics in Evolutionary Psychology. (For a classic, see, for instance, Trivers, 1971). In a group of three, instead of cutting ties with people who do not reciprocate favors adequately it might make more sense to further collaborate with them if they were subtle cheaters, cheaters who always try to give back less than you gave them but do reciprocate, because you would not have much choice, whereas you could just 'switch' to more favorable friends in larger groups (Trivers, 1971, p. 47).

Buller raises the important question of how to choose the species for comparing Pleistocene humans to, especially when the environmental challenges we are looking for concern social interactions; which one of their groups parallels ancestral societies? (Buller, 2006a, p. 96). And Machery's criticism, that "the point is not to assimilate our ancestors to one of the remnant ape species. Rather, a few evolutionary psychologists have used this literature to reconstruct, admittedly speculatively, the evolution of some known psychological traits [...]." (Machery, In press, online manuscript version p. 6) does not invalidate this: To find out how those traits have evolved (as in Michael Tomasello's work (see, e.g., Tomasello, 2009)), we would have to identify common ancestral challenges as well; in this case, we would not parallel them with humans, but we would look for challenges of common ancestors (such as hunting together) that might be similar in today's monkeys.

But how would we identify common ancestors' challenges, and hence similar forms of living in monkeys today, if we do not know enough about the ancestors' environmental challenges? And, as Buller writes: "the most recent ancestor common to us and our closest relative, the chimpanzee, lived 5 to 7 million years ago, a good 3 to 5 million years before the Pleistocene. [...] By the time early humans emerged in the Pleistocene, the adaptive problems driving their evolution should have differed profoundly from the adaptive problems facing nonhuman Pleistocene primates. [...] As a result, even our closest relative's lifestyle holds few clues to the lifestyle of our Pleistocene ancestors." (Buller, 2006a, p. 95/96).

Hence, even if we have an idea about challenges Pleistocene humans might have faced, such as males somehow having to invest in parenting and hunting groups somehow collaborating, or that reciprocal support would have been a survival advantage for the individual as well as the group, the social environment of the species will have a major influence on which behavior is going to evolve; to know which kind of group (and hence which species) to pick as a model for Pleistocene societies, we would, again, need to know more about Pleistocene societies than most of the theories about ancestral challenges provide.

If the theory about the ancestral challenge, however, were to include very specific information about the social environment of Pleistocene humans (which, as shown above, is a challenging endeavor), then observing species that live in exactly this environment and seeing whether they display the behaviors that were predicted to evolve as a solution to that challenge could be productive. Another solution would be observing many species with those challenges and see whether there is a statistical correlation between species encountering Pleistocene-like challenges and showing the adapted behavior that would solve those challenges. Such a correlation, however, could also be due to simple learning or reactions to those challenges and not to inherited behavioral tendencies.

A fourth way to produce 'Reconstructive Engineering' predictions is very different: Modelling behavior based on presupposed ancestral challenges and making predictions. Those models usually use evolutionary game theory and calculate, based on assumptions about evolutionary mechanisms and environmental challenges, which kinds of behaviors would have been advantageous in an ancestral environment.

There have been several such modelling accounts; one canonical example is Trivers' article about altruism, in which he argues that because humans typically live (and lived) in groups, reciprocal altruism benefits both parties if the costs for each party to help the other party is lower than the cost they would have incurred themselves for getting something done; say, I do something for you that costs me less energy than it would have cost you and as reciprocal favor, you do something for me that costs you less than it would have cost me (e.g. you knit me a sweater because you can knit really well and I cannot, and in return, I revise your article because you cannot further revise it yourself. Both of us have profited because it took me much less time to revise your paper than it would have taken me to knit that sweater I needed and you got the revision you needed in exchange for knitting me a sweater, which did not take you long.) Trivers' example is a person who is drowning; their chance of surviving is 1/2 without help. Say someone tries to rescue them and the rescuer's odds of drowning during that action are much smaller than 1/2 and the drowning person's risk of dying is reduced considerably by this rescuing attempt. If the person who now probably was saved repays their rescuer by rescuing them when they are in the same situation and under the same circumstances, both have gained a considerable survival advantage (Trivers, 1971, p. 35/36).

Buss uses another model in the article I have discussed above when he tries to predict which traits women prefer in a partner: (Buss, 1989) Trivers' parental investment model (Trivers, 1972). As I have argued, this model is too unspecific to constrain the space of possible environmental challenges in the Pleistocene to a sufficient degree to lead to a valid prediction: The paternal investment Trivers predicts to be preferred in the Pleistocene females could consist in contributions as diverse as cleaning, cooking, hunting and protecting, to name only a few. I will not engage in a broad analysis of models as this would go beyond the scope of this dissertation, but I will bring up some general problems that this account may raise: The problem of identifying the right ancestral problem remains: A model can only work if we make a presupposition about the environments that set its parameters. Those assumptions are the following in Trivers' reciprocation account:

> "During the Pleistocene, and probably before, a hominid species would have met the preconditions for the evolution of reciprocal altruism: long lifespan; low dispersal rate; life in small, mutually dependent, stable, social groups (Lee and DeVore, 1968; Campbell, 1966); and a long period of parental care." (Trivers, 1971, p. 45)

Hence we do not evade the problem of finding ancestral challenges; we simply incorporated those ancestral settings into a model. The model is just as good as our empirical assumptions but it can, of course, help us make predictions (or in the case of 'Reversed Engineering': help us explain behaviors.) Another problem with simulations and models is that we can change the outcomes (hence predictions) a great deal if we change the parameters, sometimes even slightly. If we do not know much about our starting conditions (the ancestral environment), it might be tempting to change those parameters until the simulation predicts our actual behaviors. And, finally, we might find the same challenges in ancestral and modern societies. Behaviors that might have helped a group to survive in the Pleistocene and hence promote its members' genes could also help a group to survive nowadays; it might hence be rational to behave like this nowadays or societies that culturally transmitted rules to behave like this might have survived longer than societies that did not. To show that some predicted complex of behavior is an adaptation, it would hence be helpful to show that it is not advantageous in today's societies or that it comes with other behavior that is not explained by a cultural transmission theory. If we, for instance, seldom kill people that we spend a lot of time with, we could surely build an evolutionary model that would show that killing our kin and allies would have been bad for our own survival and for promoting our genes (and predict that non-killing behavior). However, a society where people routinely kill each other because there is no rule against killing each other would be doomed, too, and their moral system, even if it was passed on culturally, would eventually die out. Jesse Prinz writes: "I [...] think there are universal constraints on stable societies, which tend to promote the construction of rules against harm. [...] Harm prohibitions are not universal in form; they can be explained without innateness, through societal needs for stability; and the innate resources that contribute to harm prohibitions may not be moral in nature." (Prinz, 2007a, p. 375) We would hence have two alternative models that could explain human behavior (a genetically evolutionary one and a cultural evolutionary model) and would need additional evidence to settle the case. [69]

Hence, those models are not sufficient to argue that some mechanism has developed as an evolutionary adaptation and they suffer from the same problems as other methods of 'Reconstructive Engineering' because they need to make presuppositions about the ancestral environment before they start to predict behavior. They might, however, be a promising strategy in connection with other kinds of evidence.

Of course, modelling is not limited to Evolutionary Psychology. We can find models that challenge Evolutionary Psychological explanations, too, see e.g. (Levy, 2004).

I have shown four methods of accumulating evidence in 'Reverse Engineering': Hypothesizing conditions and challenges in the Pleistocene and predicting complexes of behavior that would solve those problems and might hence be adaptations to them; paralleling lifestyles of people who hunt and gather nowadays with ancestral lifestyles; examining whether correlations between hypothesized ancestral challenges and human behavior can be found in other species' challenges and behavior; and evolutionary game-theoretical modelling. I have also shown several problems and limitations for those methods.

Next, I will expound the method of Reverse Engineering and examine ways to collect evidence and their issues as I did with 'Reconstructive Engineering'.

Unlike 'Reconstructive Engineering', Reverse Engineering reasoning starts with behavior and ends with Pleistocenic challenges. It has the following structure:

> Complex of Behavior → Ancestral Problem that is Solved? → Psychological Mechanism → Predictions about not-yet-researched Components of the Complex of Behavior, predictions about behavior in apes, predictions about

[69] Or we could have both an adaptation that would give us a tendency not to kill those close to us and culturally transmitted rules that cannot wholly be explained by our adapted tendency and hence this tendency would have partially developed as an evolutionary adaptation and partially been learned.

behavior in societies of people who hunt and gather.

The scientist observes a complex of behavior that typically seems to be non-functional in today's environment. They then propose a possible ancestral challenge to which this behavior might be the solution and an adaptation for and, based on this explanation, ideally make predictions about not-yet-researched behavior that would help to solve that ancestral problem, hence behavior that would be part of the adapted mechanism that solves the ancestral problem. They might also make predictions that apes show this behavior or that people who face the same challenge, mostly people who hunt and gather, will show the complex of behavior, too.

As mentioned before, Kurzban brought up the analogy between toasters and evolutionary adaptations (Kurzban, 2012, p. 32/33): Toasters have several features such as two narrow slots and a heating system right next to them and a button that can make things sink into the slots and a mechanism that makes things jump out of the slots after a certain amount of time. If you have never seen people toasting bread before, you may wonder about the purpose of this device. Is it a mirror? Is it a weapon? But as soon as you have gained the idea of putting bread in there, everything suddenly makes sense. You can use a toaster as a weapon and you can use it as a mirror, if it is shiny. It is not very effective for either of these purposes though; there are certainly more functional designs for weapons and for mirrors. But it is perfect for toasting bread! The slots have just the right width for slices of bread to fit in but are narrow so the heating system is close enough to the bread to efficiently heat it up. Obviously, this device solved our challenge to comfortably get our bread toasted, and it does so very efficiently. It might be made especially for this task. We could now make predictions about other features of the device, what we would want it to have to make it ideal for toasting bread. It might, for instance, have a part that collects bread crumbs so they do not fall onto the surface the device is standing on. If we actually find that part, we feel reassured that the device was designed to toast bread. We might now recognize other toasters as well, things we find in the same context, in kitchens or on breakfast tables, that may look different but have similar features.

The analogy to brain functions is the following: Scientists might find some puzzling behavior, such as pregnancy sickness. This does not seem to make sense, does it? Evolutionarily, it seems to be a rather big disadvantage if people who are pregnant get sick in the mornings: It weakens them, sometimes alarmingly. In our analogy, the scientist has just tried to use the toaster as a mirror, but it is not a very good mirror: It distorts their face and has buttons in the middle of the surface.

But what if getting sick easily, especially when smelling certain foods, keeps pregnant people from eating and/or digesting toxic substances that might be of no harm to adults, but harmful to the fetus? That would make a lot of sense evolutionarily! It would have the side-effect of weakening the pregnant person, but at the same time keep her from harming her offspring. And that would be of great advantage to promote her genes! The scientist is at the stage where everything starts to make sense. They might also discover that people mostly get sick in response to certain smells, and those are fumes of substances that are considered harmful to embryos. In our toaster analogy, the scientist is just finding out that all the buttons and the ejection function are perfect for toasting bread.

If this is true, persons who suffer more from pregnancy sickness should also experience less miscarriages. The scientist is now at the stage of predicting that the toaster has a device for collecting bread crumbs.

And maybe other species have similar mechanisms, like pregnancy-induced food aversions? Is there a correlation between being an exploratory plant eater, hence being in danger of eating toxins, and pregnancy-induced food aversions that might keep those herbivores or omnivores from doing so? In our toaster analogy: Do other devices in similar settings have the same kinds of features? Is it probable that they were produced for toasting bread, too?

Margie Profet followed the above protocol after she hypothesized that morning sickness might have evolved because it keeps pregnant people from ingesting teratogens that might be harmful for the fetus and she wrote a widely-received article about her research (Profet, 1992). [70]

Recently, however, her findings have been questioned by Leslie Cameron who found in her review article that past experiments did not generally show a lowered olfactory threshold in pregnant people although, subjectively, smells seemed to be more unpleasant to pregnant people in comparison to non-pregnant people. More importantly, pregnant people did not find potentially harmful odors significantly more unpleasant than other odors and people with nausea/sickness in early pregnancy did not eat less vegetables that are potentially harmful to the embryo than non-pregnant people (Cameron, 2014), citing (Swallow et al., 2005) and (Brown et al., 1997). If morning sickness was an adaptation to protect the embryo by keeping the mother from eating harmful plants, it should serve exactly this function; the above findings challenge this assumption. The Reconstructive Engineering method has some general issues; the often-cited "just-so-stories" criticism (see (Gould, 1978)) points out one of them. According to this criticism, the theories about behaviors and their evolutionary roots are just matching stories that cannot be falsified. In principle, one just needs to make up a story that seems to explain the adaptation and potential stories can be quite adventurous, as very amusingly shown in a video of a festival where someone explains pretty convincingly how babies are designed perfectly to throw them across great distances (BAHFest, 2014).

Of course, we can constrain those stories by what we know or assume about Pleistocene environments, evolution and physics, but considering that we do not know much about

[70] If you wonder why I cite her in the section about Reverse Engineering although I cited her in the section about 'Reconstructive Engineering' first, my answer is: I cited her in the part where I discuss other species' behavior as evidence for theories, and the role that this evidence plays is similar in both kinds of theory construction.

social environments in the Pleistocene (as shown above), this still leaves room for plenty of different explanations that might even be mutually exclusive. If scientists could not decide between those hypotheses or they only had one hypothesis but it were unfalsifiable, that would indeed be a problem, at least to adherents of Karl Popper (see Popper, 1959).

Evolutionary Psychologists try to avoid this unfalsifiability by making predictions about other behaviors that might be part of the adapted mechanism, as Profet did in the case I have discussed above: She first observed pregnancy sickness and then predicted that people who suffer from pregnancy sickness would have less miscarriages. Those predictions should, however, not be obvious in everyday life (the problem I have discussed earlier in the section about issues with 'Reverse Engineering'): If you predict something like "if this is true, most people will eat every day", this might be true and this behavior might be selected for by the ancestral challenge you have identified before, but what you really are doing is just adding one more behavior to the complex you observed and suspect to be explainable by the ancestral challenge instead of making predictions. In terms of the toaster: If you predict the toaster to be made of matter if it is also designed to toast bread, the former might be true and it is also very functional if you want to hold or heat up bread, but it is neither very specific nor is it a real prediction: You probably knew that already before you made your hypothesis, so it probably was one of the features you explained with your hypothesis, if it is connected to your hypothesis at all. You cannot just take one of the features of your device away, make your hypothesis and then add it again by predicting it if you already observed it before.

Suppose that an Evolutionary Psychologist hypothesizes a distal cause for a complex of behavior, an ancestral challenge that selected for it. As Evolutionary Psychologists cannot change the parameters of the Pleistocene setting to find out whether the ancestral problem really was the cause (nor observe whether the ancestral problem really existed), they make additional predictions to test their hypothesis. In our morning-sickness terms: They hypothesize that the cause for pregnant people to be sick in the mornings is that the sickness developed as an adaptation for mothers to avoid plants that are toxic for the embryo. As they cannot remove toxic plants from the Pleistocene environment [71] to test whether that behavior still evolves [72] (and they do not know for certain whether Pleistocene humans regularly encountered toxic plants like that), they can just predict that not only do pregnant people develop morning sickness, but also that this sickness goes with avoiding or not digesting toxic plants. If this prediction was something we knew already to be true, this would make the whole theory more fragile as the Evolutionary Psychologist would not have the other means psychologists usually have to test their

[71] For obvious reasons, at least if the scientists have a mass.

[72] This is how you would change your settings if you were conducting a 'regular' psychological experiment to find out correlations and causalities: If the behavior disappears when the harmful plant is gone or the strength/frequency of the behavior is correlated to the frequency of digested harmful plants, that would confirm our theory.

theories, i.e. he could not change the settings (in the Evolutionary Psychologist case, Pleistocene environment).

Another method to check whether a Reverse Engineering hypothesis holds is, as in 'Reconstructive Engineering', to see whether other species which show the same complex of behavior are exposed to the same environmental challenges that the scientist hypothesized were the cause for the adaptation to evolve. Unlike in 'Reconstructive Engineering', however, we start with modern humans' behavior and not with our idea about the ancestral challenge. Hence it is easier to find the right kind of species that behaves like modern humans and to compare their environment to our hypothesized Pleistocene environment: In 'Reconstructive Engineering', we had to look for animals that live in an environment with the challenges that we think existed for Pleistocene humans. But if we did not know much about the rest of Pleistocene environment, the model environment that we were looking for in the other species could be inaccurate: Instead of 'reenacting' the Pleistocenic evolutionary story, the species' social environment might differ so much that the challenges would become incomparable.

By contrast, we know much more about human behavior today than we know about Pleistocene environments. If the scientist compares species that behave similarly to (most) humans in the respect they are examining, they can observe whether this behavior is correlated to whether those species regularly encounter the hypothesized Pleistocene challenge. However, the same behavior could of course have different causes in those species and today's humans; finding a correlation would hence strengthen the hypothesis, but not finding a correlation would not falsify it. The third method, collecting evidence in societies of people who hunt and gather, is to see whether they encounter the same challenges as the scientist hypothesized Pleistocene humans to encounter and whether they show the same behaviors as today's humans; this method plays a different role than in 'Reconstructive Engineering': We already know that many people show the behavior that we think is an adaptation; the fact people who hunt and gather show that behavior, too, could just be evidence that people behave similarly even if they have very different lifestyles and hence, the behavioral similarities might be genetically implemented. This way of collecting evidence rests on the universality argument; for an extensive discussion of this argument, please read the Universality chapter.

A different strategy, namely learning something about Pleistocene challenges from the environment inhabited by people who hunt and gather because their challenges might be similar, is, as I have shown in the section about 'Reconstructive Engineering', problematic: Those environments, particularly in terms of social organization, vary considerably and might not reflect Pleistocene ways of living at all. Pursuing either of these two aforementioned strategies in the first place rests on the assumption that the people have been secluded from different cultures. Hence, firstly, they may have a lifestyle closer to Pleistocene lifestyles because the continuity has not been disturbed and secondly, if they behave similarly to other cultures although they have been secluded and not influenced by them, this gives us more reason to believe that their behavior is genetic and not learned (see Universality chapter). As both Machery and Buller agree, this assumption is not true (Buller, 2006a, p. 95), citing (Kelly, 1995); (Machery, In press, online manuscript p. 6). Usually societies that hunt and gather have not been entirely secluded and there is not one hunting-gathering lifestyle but many, and I assume that there might not have been 'the typical' Pleistocene lifestyle either, but many different ones.

I have shown that many ways of testing Evolutionary Psychologist paradigms, both in 'Reconstructive' and Reverse Engineering mode, are problematic at best, especially when it comes to hypotheses about adaptations to social environment. Even if some of the methods may not in principle be unfit for this endeavor, they are used constructively in few experiments only. A more general problem with Evolutionary Psychologist theories is the following: Evolutionary Psychologists have a concept of 'innateness' that includes development as a reaction to 'triggers': Sometimes the genetic predisposition to behave in a certain way only translates into a phenotype when something happens, when the disposition is triggered. [73] Close to our topic, Universal Moral Grammar, Marc Hauser writes: "Adopting the analogy of language, one would expect a universally held principle of fairness that varies cross-culturally as a function of parametric variation; experience with the native environment triggers the culture's specific signature of fairness and fair exchange." (Hauser, 2006, p. 72).

Hence, an entire complex of short-term and long-term behavior could be triggered once the 'fairness' adaptation 'knows' which culture a person lives in: After getting some cues from their 'native environment', they will develop a fairness and fair exchange system that is typical for their environment.

Samuel writes about triggers:

> "One reason environmental insensitivity is unnecessary for innateness is that innate structures can be triggered – roughly, acquired via non-psychological, 'brute causal' processes (Fodor, 1981). To be sure, the notion of triggering is far from transparent. But what's clear is that triggering only results in acquisition where the relevant environmental factors obtain. Thus innate structures that depend on triggering are not invariant with respect to environmental factors." (Samuels, 2008, p. 29), citing (Fodor, 1981)

I cannot add much to that. But here comes the problem: As in Hauser, we may have several adapted fairness behaviors that get triggered depending on our environments, in

[73] Please do not confuse this concept of triggering the realization of genetic dispositions with the triggering of direct behavior, as in "automatic, reflexive responses to specific triggering stimuli" (Barrett and Kurzban, 2006, p. 636), or in evoking post-traumatic episodes, which is a more common use of the word.

his case, the culture we grow up in. Let me cite an example that makes the problem clearer. Buss writes: "Early father presence shunts children into a mating strategy marked by long-term monogamous relationships with high parental investment; early father absence shunts children into a mating strategy marked by early sexual maturation, early sexual intercourse, more numerous short-term sexual encounters, and a generally more promiscuous mating strategy." (Buss, 1995, p. 23). But those parametric accounts make this allegedly predisposed, specialized behavior very flexible. The more factors come into play, the less specialized the mechanism. How do we decide whether something is an adaptation or has developed as reaction to the environment?

I think the clearest distinction between genetically predisposed behaviors and learned behaviors is that a trigger would not provide sufficient information to learn the behavior. Otherwise, we could just call every environmental influence a 'trigger' and we could not decide between genetically predisposed, Pleistocene-challenge-specialized, adapted behavior and other kinds of behavior. In Buss' case, the environmental influences are sufficient to learn the behavior and in Hauser's case, he describes them too unspecifically to decide whether they would be sufficient to learn a culture-specific justice system. Those are only two examples, there might be others that introduce more specific kinds of triggers.

My only worry is: If we do not explicitly introduce the notion that the triggers should be insufficient to learn behaviors (e.g., they should not include all the information we would need to show some regular behavior as in moral or linguistic grammars, see Poverty of Stimulus chapter) if they trigger evolutionary adaptations, Evolutionary Psychologists might just introduce more and more parameters to explain behaviors until specialized behaviors are indistinguishable from non-adapted behaviors: Behaviors that have not been selected for with regard to a specific situation/ancestral challenge.

And, last but not least, we should be cautious about adopting Evolutionary Psychologist explanations when there are alternative explanations available (e.g., (Levy, 2004)), as the brain is very plastic and, as Buller argues in an essay, our psyche is very likely an adaptation to enable flexible responses to quickly changing environments (Buller, 2006b). In this framework, Evolutionary Psychologist explanations are not necessarily sparser or form a better fit for our epistemic networks: According to Buller's argumentation, it would make more sense both evolutionarily and according to neuroscientific research if our brains tended more to the flexible instead of being specialized to ancestral challenges.

Epistemologically, Evolutionary Psychology will be an even more critical endeavor if we believe that the adaptations have developed as solutions to problems in the social realm, as I have shown above. If we, however, pose Evolutionary Psychologist research questions, we should make sure to collaborate with people trained in Social Sciences (for alternative explanations) and Archaeology (for qualified information about Pleistocene societies).

Many proponents of the Linguistic Analogy have spoken out in favor of the innateness of the Trolley Dilemma judgement mechanism (and other moral mechanisms) (Dwyer, 2006), (Hauser, 2006) (Mikhail, 2008), (Harman, 2011).

Most of them have argued that 'descriptive adequacy', a description of how morality (or language) is computed, is necessary to argue whether morality or language are innate. They have, however, extensively cited Poverty of Stimulus, Universality and Developmental evidence for the innateness of the Trolley Dilemma judgement mechanism; they have also argued for it to be an adaptation and modular in the “Mechanistic” or “Functional Modularity” sense (see (Dwyer, 2008) in this chapter and (John Mikhail, 2007a) in the Mechanistic Modularity chapter). I have shown that it is almost impossible to gather evidence that something like moral principles are adaptations using the types of evidence cited above.

It is, however, rather obvious that generally having an aversion to, or finding it immoral to, harming people close-up and personally is functional and rational both in today's and ancestral societies as long as people lived in groups. The same holds for the Action/Omission Principle: If you harm someone by omission, you do not intervene when a person is in danger. For this to be a harmful action, a person has to be in danger already. The ways of harming someone actively, however, are indefinitely many. So, for any kind of society, if people were allowed to harm people actively, that would be much more dangerous than if they were allowed to harm people by not helping them. Hence, again, it would be evolutionarily advantageous not to slay all your peers (hence have an aversion to harming people actively) as well as advantageous for a society to more firmly enforce the rule that it is forbidden to slay everyone than the rule that it is forbidden not to help people. The situation is harder with the Doctrine of Double-Effect. If we place a focus on the “intention” part of the Doctrine of Double-Effect (doing something with a good and a bad effect is allowed if 1. A good consequence is intended; 2. The bad consequence is merely a foreseen consequence of the intended action), it becomes obvious again why behavior according to this rule would be useful in ancestral as well as in modern contexts: We can only control whether we intend harms; a rule against accidentally harming people would be hard to follow and an adaptation keeping us from accidentally harming people would have to prevent us from doing almost anything because every situation can result in accidentally harming people. The only thing we could do, of course, is to be careful not to harm people; but the difference in gravity of intentional killings and accidental killings (involuntary manslaughter vs. murder) in most penal systems might reflect, just as the name, that it is hard for us to control involuntarily harming people. Both an adaptation and a rule that kept us from intentionally harming people in our environment, however, would be advantageous both in ancestral societies as in today's societies. As our problem has the structure of Reverse Engineering, hence the proponents of the Linguistic Analogy already know which behavior they hypothesize to be adapted, our next step would be to find an ancestral challenge to which this behavior might be a solution; however, our manner of individuating those kinds of behaviors normally is to find behaviors that would make sense in ancestral societies, but not in todays'. This is not the case with all three principles I have mentioned above. More research might be needed, but this will be difficult.

The form-function (or behavior/ancestral challenge) fit, however, is not the only way to establish whether a complex of behavior has developed as an adaptation to an ancestral challenge: In the following chapters, I will present other kinds of evidence that the proponents of the Linguistic Analogy have used to argue for the innateness of the Trolley Dilemma judgement mechanism.

4 Development

Proponents of the Linguistic Analogy (and, less often, Evolutionary Psychologists) often mention two factors related to development as indicative that some trait or behavior is genetically inherited:

1. The trait develops early in life, e.g.: Children start talking very early in life. Many linguists have taken this as evidence that a great part of language does not need to be learned but is genetically 'programmed'.

2. The trait takes a typical developmental trajectory that is similar in every human, e.g.: Children typically start babbling (vocally or, if they cannot hear, with their hands) at an age of around 8-10 months (Petitto and Marentette, 1991) and acquire the first 50 words/signs of their vocabulary at an age of around 18 to 19 months ((Lillo-Martin, 1999), citing (Bonvillian and Folven, 1993) and (Nelson, 1973)), and children master languages better and learn them quicker before they reach puberty (Johnson and Newport, 1989, abstract), citing (Lenneberg, 1967). This is also taken to be evidence for an innate linguistic mechanism.

To cite a party from the Evolutionary Psychologist camp, Carruthers et al. write:

> "In other words, we now know that each individual's cognitive development follows a domain-specific trajectory for each cognitive domain [...]. However, we also know that within each domain there exists a well-ordered pattern of development, and that this pattern is uniform for all normal members of our species [...]. Moreover, there has been a striking trend in the developmental psychology of the past 25 years or so, finding that very young children are much more like adults, cognitively, than was supposed by Piaget. [...] This kind of evidence points strongly toward the existence of uniform, species-wide, innate cognitive endowment that consists (at least in part) of various domain-specific faculties." (Carruthers et al., 2005, p. 9)

Again, Linguists, proponents of the Linguistic Analogy and Evolutionary Psychologists seem to conceive a connection between early onset of behavior, a specific developmental trajectory (both in terms of the order of development as [74] in terms of the onset of changes) and innateness, but mostly do not explain the actual connection explicitly.

It is, however, relatively evident why behaviors that children show very soon after birth are more likely to be inherited, and the reason is connected to the Poverty of Stimulus argument: The sooner something comes up in development, the less input the person

[74] And these two terms are an addition to what Carruthers writes, they are not contained in the original text citation.

can have had before they learned it. To make an extreme case, if someone would speak a complete sign language right after birth, hence without having seen anyone speaking even a word of that language, it is very likely that this ability was somehow genetically 'built in', because they could not have learned it without having any linguistic input. The idea here being that the less time you need to develop some behavior, the less input you probably got during that time (see also: Cruz and Smedt, 2009, p. 5). As Marc Hauser puts it: "When a behavior emerges early in development, it is tempting to conclude that its foundation lies in nature. The reason is simple: based on timing, there has been insufficient experience for nurture to construct the details." (Hauser, 2006, p. 164). This, of course, is not necessarily so. [75] Regarding linguistics, for instance, there is already a great deal of input in pregnancy, so that children who are born are already able to differentiate between the language they have been exposed to in the womb and other languages (Moon et al., 2013).

Hauser further writes: "In addition, traits that appear late in development are not necessarily the result of nurture's handiwork, as evidenced by the emergence of facial hair in men and breasts in women, secondary sexual characteristics that arrive as a result of maturation and the achievement of puberty." (Hauser, 2006, p. 165). If we try to read this in Hauser's favor, he is not saying "either way, regardless of the empirical evidence, you cannot falsify our innateness hypothesis (at least not in terms of the onset time of the trait in question, even if it would strengthen our hypothesis more if the onset were early in development)", and we can read a more specific claim: This maturation, according to the second developmental factor proponents of the Linguistic Analogy relate to innateness that I have mentioned above, would likely happen in a fixed trajectory (with the phenomenal development of the trait occurring at similar ages overall, and a fixed order of stages of maturation across most people). This is a more beneficial reading of Hauser's claims because it might not falsify Hauser's innateness hypothesis if the trait in question does neither develop early nor in a certain trajectory, but it would strongly weaken it; this makes his hypothesis at least slightly more testable and therefore scientifically interesting. The connection between this manner of maturation and innateness is more tricky than the connection between early onset of a trait and innateness.

One possible reason to claim that something innate would mature along a trajectory is the organ-likeness that Hauser (e.g., Hauser, 2006) and Chomsky (e.g., Chomsky and Hornstein, 1980) attribute to the seat of morality and language respectively. Just like the heart, the liver or any other organ (or, as in the following citation, the cognitive mechanisms that look most promising for modularity theories), the 'moral organ' would follow its pre-

[75] In principle, it is conceivable that you only get a specific type of input during your time as an embryo and never again after birth; being older than a new-born would, in that case, not change anything about your learning from experience (if anything, it could change your skills because you learn by practicing). This, however, is a merely conceptual argument that clearly does not apply to the realm of morality: It is very improbable that children acquire most of their moral experiences in the womb.

determined path of growth: ”The findings of Experiment 6 [...] rais[e] [...] the possibility that, like language, vision, and other cognitive systems, moral development involves pre-determined critical stages, after which moral competence more or less stabilizes.” (Mikhail, 2002, p. 66).

A different, but related connection between a certain developmental trajectory and innateness could be: The less something is influenced by empirical learning, the less diverse interpersonal environments can influence its development to interpersonally diverse trajectories. That's why the traits show a typical, universal, unified trajectory based on the universally equal genetic endowment that leads their growth. Again, if a trait is influenced by external factors (and to a great part contingent upon them), its development might vary with varying environments. The language learning curve, for instance, would then be contingent upon the input and children with very little input would show a dramatically slower learning curve (I do not wish to discuss the innateness of language here, but this might just be the case (see Oxford and Spieker, 2006)). John Mikhail might have had in mind this unaffected development that is not influenced by any external factors when he wrote that linguists had argued that: ”knowledge of language is [...]the product of a distinct mental 'faculty', with its own, largely pre-determined developmental path. Again, traditional natural law theory is predicated on a similar belief: that every human being possesses a faculty of moral judgment - a conscience - whose normal development is unaffected by racial, cultural, or even educational differences.” (Mikhail, 2000, p. 357) [76]

> ”Acquiring data about the point at which children achieve the cognitive milestones that are required for making moral judgments will allow for the investigation of which processes are developmentally necessary and which developmentally sufficient for moral judgment. LA predicts that there will be a typical developmental pattern for the emergence of the various capacities that are at work in making moral judgments. [...] However, if empirical inquiries into the development of moral cognition show that this is not the case, LA will seem far less plausible as an account of moral cognition.” (Dwyer et al., 2009, p. 503)

The link between innateness and development in stages is, as I have shown, so weak that it is obviously not necessary: A trait that is innate does not have to develop in certain, typical stages. It could instead be there at birth already, or be triggered by the environment independent of a set time frame or it could develop in one step instead of a set order of steps. Neither is a trait that develops in stages necessarily innate. According to Jean Piaget's paradigm, for instance, morality develops in typical stages according to age,

[76] Mikhail writes that Rawls' Linguistic Analogy raises those questions but does not take any stance on this in his dissertation. In his later book chapter on Poverty of Stimulus, however (Mikhail, 2008) and his reply to an editor (Kirkby et al., 2013) he takes a more suggestive, although not decided stance towards inheritability of his Trolley Dilemma judgement system.

but he believed that those stages depended on experiences, not a genetically determined maturation process ((Haidt et al., 2007, p. 374), referring to (Piaget, 1948)).

As there are only very few studies regarding the development of Trolley Dilemma judgements in children and those studies have small respective sample sizes and, to my knowledge, no longitudinal studies exist, I can already predict that we will not settle the issue of whether Trolley Dilemma judgements change in typical stages, following a typical order with onset of changes at typical ages. This is why more empirical (and especially, longitudinal) studies are needed. I will, however, review all studies present that examine children's reaction to Trolley Dilemmas with regards to the onset and developmental trajectory of the judgement pattern.

I will present one caveat at the beginning: Just as late development is, according to Marc Hauser, not compelling evidence that something is no inherited trait, so early development of a trait is not compelling evidence that something is inherited. It is just one version of the Poverty of Stimulus argument. So no matter how old someone is, it is not the time at which a child becomes able to judge according to our three principles that shows that it cannot have learned these from its experiences, but it is the child's experiences themselves. If someone is very old but had poor moral input over time (did not witness many moral judgements/assessments per year), they might have had the same input as someone very young who had relatively rich moral input already (witnessed many moral judgements/assessments per year). Therefore, it is important to review single cases and (apart from examining the results and methods) see how much input children have likely had at the age at which they participated in the story.

Firstly, we will examine whether children aged 8-12 respond to the same principles as most adults did in the Cushman et al. 2007 study (Cushman et al., 2007). 6 of 15 children judged the Push Type Case permissible, while 14 of 15 children judged the Switch Type Case permissible. This means that, on average, children between 8 and 12 seem to make a significant difference between close contact harm as a means and no-contact, impersonal harm as a foreseen side-effect.

Importantly, the two dilemmas used here were a hospital setting (a doctor can cut up one healthy person, harvest their organs and distribute them among 5 otherwise terminally ill patients) and a train setting (a classical sidetrack case where someone can direct a trolley heading towards five people onto a side-track where it will only kill one, albeit with the agent being the driver). The first scenario was a rather juicy, drastically described situation in which "Dr. Brown" could "cut up" a patient [77] including a description of the sacrificed person as "perfectly healthy", their gender and their room number. The second one was a sidetrack case without all those factors ("switch his train" instead of "cut up the person", no gender and, obviously, no room number) (Mikhail, 2002, p. 125). The contrast between those scenarios is considerable: The first scenario gives far more

[77] The term "cut up" is repeated 3 times in the description.

material to imagine the situation, the battery involved in sacrificing the patient is more drastic than in the classical Push Type dilemmas and the revealing of the killed person's gender might make them seem more familiar. Apart from the relatively small sample (30 children), this makes it difficult to say which of these factors lead to the higher number of impermissibility judgements that the children assigned to the first dilemma. Interestingly, most of the cited justifications for impermissibility judgements indicate that the worst part about cutting the person up to distribute their organs is that they has not consented before and not the harm done to them (Mikhail, 2002, p. 125). [78]

1. However, Mikhail concludes: "Although the results of Experiment 6 are limited, they constitute at least some initial evidence that the moral competence of 8-12 year old children includes the prohibition of intentional battery and the principle of double-effect." (Mikhail, 2002, p. 64/65);

2. "[T]he results of Experiment 6 imply that at least some of the operative principles of moral competence, such as the distinction between intended means and foreseen side effect, are invariant throughout the course of moral development, at least between ages 8-65." and

3. "The findings of Experiment 6 call at least some aspects of this investigative procedure [79] into question, raising the possibility that, like language, vision, and other cognitive systems, moral development involves pre-determined critical stages, after which moral competence more or less stabilizes." (Last two citations: Mikhail, 2002, p. 65)

Setting aside the deficient methods I have mentioned above, Claim 1 may be justified. Claim 2 is warranted insofar as the judgements of older people seem not to vary much from the judgements of the children in his study. Claim 3 is formulated carefully enough as to be justified as it only raises a possibility. However, the evidence Mikhail presents might not even be sufficient to support this possibility: The possibility that moral development proceeds in pre-determined stages, vague as it is, is not warranted by what he shows: He only demonstrates that children of 8-12 years have gone through some kind of moral development (if the judgement distribution was due to moral features and mirrors a mechanism that produces judgements according to principles) and shows, at best, only the end product of a critical stage: The state after children started judging according to moral principles. Furthermore, the other cognitive systems that he mentions and compares to moral systems, language and especially vision, have their most productive phases far before the age of 8 years: Although languages can be learned better before puberty, ((Johnson and Newport, 1989, abstract), citing (Lenneberg, 1967)) the classical view is that acquisition

[78] This reminds me of something I have heard parents tell their children rather often, namely: "Don't do that, she doesn't want that!" or "Don't do that, she said no!"

[79] The view that moral development is something that happens gradually over the course of one's lifetime, and thus should be investigated by means of longitudinal studies.

happens in many different typical stages which lie to a great part in early infancy (Kuhl et al., 2005, abstract). [80]

The critical phase for the visual cortex has long been regarded as even more strict than the critical phase for language acquisition, with children deprived of binocular input up to the age of one year (but not four years) developing severe and afterwards hard-to-treat forms of amblyopia (decreased vision in one eye due to brain connections rather than eye malfunctions) (Morishita and Hensch, 2008, appendix 1). There are three reasons for this: Firstly, typically, as I just mentioned, children acquire their abilities in vision and language most quickly at a very young age. Secondly, language and vision acquisition are believed to have 'critical periods': Periods after which, if children have not acquired certain skills, they will have difficulties or be unable to acquire those skills. Hence, to show that there is a critical period in moral development acquisition, Mikhail should have shown that children under a certain age are very quick learners of moral principles or children who did not acquire them up to a certain age will not acquire them at all. Thirdly, if he mentions language and vision as typical examples of Poverty of Stimulus applicability, because their acquisition is at a very high level at a very young age (or if he argues for the Poverty of Stimulus variant at all, namely that children at a very young age have not had enough opportunity to learn moral principles by experience), he should have tested younger children: At the age of 8 years, children have had plenty of opportunities to acquire the principles (see Poverty of Stimulus chapter); the older on average that children are when they show a trait (hence, the more experience they have had), the less probable it is that the Poverty of Stimulus argument is applicable to them.

One of the most promising accounts for testing the development of all three principles discussed in this thesis (Close Contact Harm Principle; Doctrine of Double-Effect; Action-Omission Principle) comes from Powell et al. (Powell et al., 2012). They tested whether children aged 5/6, 7/8 and adults judge according to the three principles.

I will review the results according to the respective principle that Powell et al. tested, in the following order: Close Contact Harm Principle; Doctrine of Double-Effect; Action/Omission Principle.

Powell et al. presented two situation pairs (Powell et al., 2012, appendix) to 15 5- and 6-year-olds in order to test for the physical harm principle. First, I will describe the two situations of the first pair:

In one of them, a boy pushes another boy he does not like off a bike which causes the latter to break his leg. In the other one, he yells at him in order to make him fall and the other boy falls off the bike and breaks his leg. Note that the second case description explicitly mentions the boy's intentions to hurt the other boy and both cases are not dilemmas (they yield no benefit to anyone apart from, possibly, some sadistic joy or anger relief to the agent). This makes them structurally different from all the Trolley cases I have

[80] Although challenges to this view reveal a more complex picture (Abello-Contesse, 2009).

reviewed before. After presenting those cases, they asked which boy's action were worse. Hence the overall answers did not refer to whether the actions were permissible (obviously, most people would judge them as morally impermissible as no-one is saved and one person is harmed) but were designed to see which action types were judged better in relation to other action types. Despite the difference to the other experimental set-ups, the cases are fit to test the children's intuitions about harm by close contact harm and without physical contact because if children judge the action in the first situation (pushing the boy) to be worse than the second one (yelling at him), they do that although previous studies (e.g. (Nelson 1980), [81] (Baird and Astington, 2004)) indicate that if children are influenced by explicitly mentioned intentions ("[...] Nick jumped out and yelled to surprise the other boy and to make him fall off his bike" (Powell et al., 2012, Appendix)) this would rather bias them towards evaluating the (explicitly) intentionally harming agent worse than the not explicitly intentionally harming one: If explicitly mentioning the intention had an effect, it would probably be in the opposite direction. The case for the physical contact principle becomes even stronger in this experiment because people preferred non-close up, personal actions with explicitly intentional harm to harmful, not explicitly intentional actions involving physical contact.

The second situation pair can be read as a dilemma, albeit not a third-person dilemma: In the first situation, someone throws snowballs at a boy. In one situation, he grabs another boy without hurting him and pulls him in front of himself so that the other boy gets hit by three snowballs. In the second situation, someone throws snowballs at a boy and the boy calls a different boy who does not see the snowballs so the second boy gets hit and he is shielded. Again, the intention in the second case is explicitly formulated. The dilemma here is between the first boy getting harmed and the second boy getting equally harmed (or more, if we include the physical contact of pushing him and weigh it negatively); here, the structure differs from classical Trolley Dilemma Type cases although it is a dilemma because the person who makes the choice can save himself and not others and the question is not whether it is permissible for him to do so but whether doing it using close-up, personal contact is worse than doing so without close-up, personal contact. The question itself is formulated in a way that presupposes both actions are bad and therefore not morally permissible (the question is which action is worse or whether they are equally bad (Powell et al., 2012, p. 191)).

The results for both situation pairs are the same: 83% of the 5/6-year-olds judged the action of the boy who used close-up, physical contact worse than the action of the boy who did not (Powell et al., 2012, p. 191). This replicates Mikhail's results, albeit for

[81] In Nelson 1980 this was only the case when children were presented with drawings of the scenario, but as those children were only three years old and, in Powell's case, the scenario was acted out for the children with dolls, I think we can draw similar conclusions assuming that the children in Nelson's case judged differently when they only heard the stories because of understanding problems or because they had no visual stimulus.

younger, namely 5/6-year-old children.

The next situation pair was designed to see whether the same children judged actions resulting in harms that were foreseeable and unavoidable to be better than actions resulting in harms that were a "direct means to achieving" a goal (Powell et al., 2012, p. 191).

The first situation is identical to the last situation described: A boy calls an unsuspecting different boy to make him stand in front of him and shield him from snowballs. The second boy then gets hit by the three snowballs that were directed towards the first boy.

In the second situation, behind the boy [82] who is originally targeted by snowballs is a second boy. When the snowballs fly, the first boy ducks down which results in the second boy being hit by three snowballs (Powell et al., 2012, appendix).

In the first situation, the first boy intended for the second boy to shield him; in the second situation, the harm to the second boy was a foreseeable but (according to the authors) unavoidable side-effect. This, again, differs from classical dilemmas where one person is killed in order to save five as in our case one of the agents saves himself at the expense of equal harm done to others whereas someone who is not in danger has the choice of killing one person in order to save five in classical Trolley Dilemma Type cases. The question posed to the subjects is, again, which action is worse.

As mentioned above, this situation pair tests for a distinction between foreseeable, unavoidable harms and harms that were a "direct means to achieving" a goal (Powell et al., 2012, p. 191). This is only partially in the spirit of the Doctrine of Double-Effect as I have formulated it because the overall effect that the agent intends, saving himself, is not necessarily a morally good effect. For the Doctrine of Double-Effect, however, the intended overall effect needs to be a good one, otherwise it is not permissible to harm someone as a foreseeable and "unavoidable" side-effect. [83] However, even if both actions are perceived as bad actions it is interesting whether one action is regarded as worse than the other, because even if both actions are impermissible, the system at work could be similar to the Doctrine of Double-Effect but now not separating the judgements on the scale between permissible and impermissible but between impermissible and even less permissible. Powell et al. found that 79% of the 5- and 6-year-olds judged the situations where the boy harmed someone as a direct means of saving himself worse than the situations where the harm was a foreseeable and 'unavoidable' side-effect. It is not quite clear whether this difference is because the children ascribed intentionality to the first action only (where it is explicitly formulated) or whether it is due to the means/foreseen side-effect structure of

[82] Curiously, the agents in those scenarios are typically men and boys with only a few exceptions such as Mikhail's female doctor (Mikhail, 2002, appendix).

[83] This notion strikes me as rather vague as, had the other boy not stood behind the first boy, the first boy could have saved himself and the other boy would not have gotten hit either; the harm to the second boy was not as unavoidable as the harm in Loop Type cases where hitting one person is necessary to stop the train and the five persons would have gotten killed if the person had not been there; this Loop case of unavoidable harm, however, is classically the case where a person is harmed as a means and not as a foreseeable side-effect.

the situations or a compound of both.

If we grant that Powell et al.'s findings test for the Close Contact Harm Principle and the Doctrine of Double-Effect, we now have the same kind of results that Mikhail already obtained. We have a little more information now, such as that children of aged 5/6 already endow those principles and that people seem to continuously judge according to those principles, at least from the age of 5/6 on, without a break in between. The basic criticism, however, is that even children as young as five have already had much time to learn moral principles. As Cruz and Smedt write: "[...] as Haith (1998) observed, infants of a few months old have had several hundreds of hours of waking time." (Cruz and Smedt, 2009, p. 58), referring to (Haith, 1998). If we assume that children tend to sleep rather less than more as they grow older, 5-year-olds have at least 10 times as much waking time. The empirical question to solve is how much moral input they get during this time and whether it suffices to learn the two principles tested here (see Poverty of Stimulus chapter). We have not learnt anything about the developmental steps of the two principles either, as we do not know whether there are any transitional steps between not applying the three principles and applying the three principles; neither can we compare the studies with children of different ages and see whether the percentage to which they judge according to the principles changes because the situations used as stimuli, as I mentioned earlier, differ fundamentally.

If the salient principle that leads to this judgement distribution, however, is not the Doctrine of Double-Effect, but attribution of intention to the agents, we would have to introduce a different principle. The impact of intention on judging moral situations is well researched and John Mikhail, e.g., assumes that people judge the Doctrine of Double-Effect the way they do because they, "absent conflicting evidence", compute that in actions that have a good and a bad effect, the good effect was intended. If the harms, however, are a necessary means to the good effect (such as in the situation where the second boy shields the first boy from snowballs), that is counter-evidence because the good effect would not have been achieved without the bad effect and hence the bad effect was probably intended, too ((John Mikhail, 2007a, p. 148), see also (Mikhail 2011)).

Baird and Astington (Baird and Astington, 2004) show that there seems to be a typical age for children to learn judging moral dilemmas based on intentionality. Although they do not use cases similar to Trolley Dilemmas for their experiments, I will briefly review the findings: 4-year-olds differentiated morally between cases where the agent intended to do harm and cases where the agent did not intend to do harm, where actions were the same and outcomes were not mentioned. For instance, they judged turning on a garden hose worse when the agent intended to destroy their brother's sand castle by doing it than when they intended to help their mother by watering plants. However, the children made a significantly smaller distinction between good and bad intentions than did 5- and 7-year-olds (Baird and Astington, 2004, p. 41). This indicates that children typically start making moral judgements that rely on the agent's motivations between the age of four and five. This in turn suggests that there is one transitional period where children start making those judgements (the 4-year-olds already judged based on the agent's intentions, although they made a smaller differentiation than older children); the data do not suffice for an entire differentiated trajectory with several acquisition stages. However, the significant difference between the average judgements of children with only one year age difference is an interesting finding that might have indicated a pre-determined maturation process of moral judgement systems in children. [84] However, the ability to make judgements depending on the agent's intentions correlated positively with false-belief understanding (Baird and Astington, 2004, p. 42). This progress in judgement making, even if it is steered genetically in a way similar to puberty, is probably not due to the maturation of a moral judgement system but of a more basic ability, namely the Theory of Mind. Hence, moral judgements based on intentions rely on systems from other domains (see the Mechanistic Modularity chapter). This predictability of age based on judgement patterns does not inform us about a possibly innate mechanism dedicated to making moral judgements. It only tells us that the predisposition to judge based on agents' motivations was already there before the age of four or came at the same time as children acquire the false-belief skills necessary to understand the agent's intention at all (or that the mechanism that leads to judgements based on agents' motivations consists of, is part of or has as constituent part the false-belief mechanism). The story might go like this: As soon as children have the skills to attribute intentions, they can finally apply their tendency to judge according to the agent's intentions. (Or their false-belief skills trigger the maturation of their morality skills). Or it might go like this: Children acquire false-belief skills at the same time as they acquire their tendency to judge according to the agent's intentions. Or: There is an evolutionarily advantageous function of morally judging agents according to their intentions, and it matures around the age of four, and part of this is the ability to ascribe intentions and false beliefs to agents at all.

Basically, Baird's data say nothing more about the acquisition process than that children gradually start making moral judgements dependent on the moral agent's intention at the age of four and this correlates positively with false-belief skills. This is reconcilable both with theories of gradual learning and theories of gradual maturation, indicates that children younger than four mostly do not differentiate between the agent's intentions in moral judgements (which gives them, again, plenty of time to acquire this differentiation from their environment) and does therefore not warrant a Poverty of Stimulus argument

[84] Assuming that the learning environment is not so uniform as to warrant that all children have the same learning input by the same age and therefore the maturation process must have been steered by predetermined genetic mechanisms; it would then be an empirical question whether this is a correct assumption, testable by, for instance, examining whether children with very different moral environments show a similar maturation curve (make similar judgements at similar ages). Additionally, those data are averaged upon many children. Even though the age might predict moral maturation, there could be great variance in learning curves that may correlate with different amounts of moral inputs.

(because Poverty of Stimulus presupposes that an acquisition process cannot explained otherwise than by innate knowledge/tendencies).

In summary, Powell et al.'s data determine the age at which children already judge according to the Close Contact Harm criterion as 5/6: At that age, they already discern between harm as a means and foreseeable side-effects or else, as I argued, between explicitly intentional harms and harms as foreseeable side-effects. Baird and Astington's data suggest that children acquire the skill of attributing intentional harm and judging based on motives when they are around four years old. Depending on which is the salient feature that makes children judge cases as morally worse, to make a Poverty of Stimulus argument, we would then need to show that children of either four or five cannot have learned the principles by environmental inputs.

The results are similar for the action/omission distinction: Powell et al. tested twenty 5/6-year-olds, fourteen 7/8-year-olds and 34 adults with two situation pairs. The first one was a harm-only pair, the second one was a Switch Type dilemma case. In the first harm-only situation, a train is heading towards a puddle with one girl playing next to it. If it passes the puddle, the girl will get wet. The switch operator can, however, turn the train to a different track where no-one is playing but he decides to do nothing. In the second harm-only situation, the train is heading towards an empty track and the switch operator decides to turn the train to a track next to which a girl is playing who will get wet because the train passes a puddle again. For this situation pair, when they were asked whether the action in the first or in the second situation was worse or whether they were equally bad, adults considered the commissions/omissions equally bad more often; not counting those 'same' judgements, "all age groups judged commission to be worse than omission" at about the same rate (Powell et al., 2012, p. 190).

In the Trolley Dilemma Type pair, a train was heading towards a puddle close to five children in the first situation; the switch operator decided to turn it to a different track where only one girl instead of the five would get splashed with cold, dirty water. In the second Trolley Type situation, the train was heading towards the one girl and the switch operator decided not to turn it to the second track with the five children.

The 7/8-year-olds and adults tended to rate both actions to be good actions whereas the 5/6-year-olds rated them badly. The 7/8-year-olds did not significantly differ from the adults in their ratings as to which action they preferred; most of them did not make any distinctions between commission and omission in those cases, but the rest was "significantly more likely to judge that switching the train's path was better than permitting the train to continue its path" towards the one girl (Powell et al., 2012, p. 190). The 5/6–year-old children, however, did not make a significant distinction between commission and omission and, as mentioned before, judged the actions (or omissions) overall to be bad. Powell et al., in the light of other research (Lucas et al., 2008) indicating that preschool "children have limited abilities to engage in cost/benefit analyses more generally", hypothesize that "[i]t is possible that the youngest children could not compare and weigh the different harm and benefit elements of the harm/benefit stories." (Powell et al., 2012, p. 192)

If that is true, we can draw an analogy to the research about moral judgements based on intentions: Weighing harm against benefit is a capacity necessary to judge Trolley Type Dilemmas according to the Action/Omission Principle, just as ascribing intentions is a precondition for being able to judge moral situations according to the principle that intended harm is worse than foreseeable harm. But, unlike the Action/Omission Principle, we do not know whether children would have the tendency to judge moral actions depending on the agent's intentions before they are able to ascribe intentions (that the only thing they lack for judging that way is the ability to attribute intentions but were they be able to do so, they would judge intentional harm worse than foreseen harm). This is also a question that is hard to answer because attributing intentions is necessary to judge depending on the agent's intention structure. By contrast, the moral mechanism that leads to judging harmful omissions better than harmful actions seems to be in place at the age of 5/6 as the non-dilemma results before have shown. Weighing harm and benefits is crucial for judging dilemmas, but not crucial for judging different moral situations according to the Action/Omission Principle. The evidence we need to establish a case for innateness would be a Poverty of Stimulus case for children under five years (or evidence that they develop their action/omission distinction even earlier).

The results of a study by Baron et al. testing children from five to twelve are inconclusive: One experiment showed no significantly smaller omission bias in five-year-olds than in the older children (and no significant difference between all tested ages) and in the other experiment, second-graders had a larger omission bias than tenth-graders (Baron et al., 1993): Baron et al.'s study showed an age effect in one experiment and no age effect in the other one. [85]

In a different experiment by Knafo and Hauser that seems to have not been published but whose results Hauser mentions in his 2010 paper with Abarbanell (Abarbanell & Hauser 2010, p. 213), 29 out of 46 English-speaking, Israeli-born children found that an agent who opens a window to break a vase in order to wake up her dog so she can play with it acts worse than an agent who leaves the window open to break a vase in order to wake up her dog so she can play with it (for the exact formulations of the stories, see Abarbanell and Hauser, 2010, p. 222, Appendix A). However, they do not mention the children's ages.

And, lastly, in Cushman et al.'s 2007 study, age could account for 1.4% of the variance in percentages of permissibility judgements in Loop and Loop with Heavy Object Dilemmas. The tendency, however, was not quite consistent: the trend was that subjects gave the Doctrine of Double-Effect less weight the older they were (they judged both dilemmas more similarly), but the Close Contact Harm Principle received more weight, at least up

[85] Powell et al. write that the effect was very small in the 6-year-olds (Powell et al., 2012, p. 188).

to the age of 40. I do not think these data are strong enough to explain anything about how and whether adults learn to employ those principles when they get older, but they show that the judgements seem to vary somewhat depending on the subjects' ages.

In summary, I listed three reasons why a certain kind of trait development should be related to innateness of this trait:

1. If children who are very young show a behavior and cannot have had the environmental input to learn this behavior with empirical methods, the Poverty of Stimulus argument sets in. This is not necessarily connected with age, but normally the older children are, the more input they have had and, hence, the more opportunities to learn something. This means that in order to show that children cannot have learnt something by that age (and hence must have had innate tendencies that lead to their behavior), we neednot only to show that children show certain behavior at a very young age, but also present additional evidence that they (normally) have not received the decisive input by that age.

2. If the development of a trait follows a certain trajectory in order and age, this can mean that a trait that is genetically 'built in' matures like sexual organs in puberty.

3. This cross-individually uniform maturation could also mean that not only is the information 'built in', but it is also largely unaffected by environmental influences such as, in our case, moral input.

The studies all used very diverse stimuli and mostly did not insulate single principles optimally (e.g., the Push Type Case and the Switch Type Case in Mikhail 2002 differ both in terms of their close-up, personal/distant, impersonal harm structure and their harm as means/foreseen side-effect structure). Therefore they are not easy to compare but I will try to condense the results in the following paragraph.

The experiments I have reviewed show that 5/6-year-olds already seem to make a difference between close contact, personal harm and no-contact, impersonal harm (Powell et al., 2012). This replicated Mikhail's findings that children between 8 and 12 make the same distinction (Mikhail, 2002). According to Powell et al., 5/6-year-olds also seem to make a distinction between harming someone as direct means or as foreseeable and "unavoidable" side-effect (a feature that, besides the close contact, personal harm component, was also identified in Mikhail's dilemmas with 8-12-year-olds). [86]

Powell et al.'s results, however, might have been due to differences in intention ascription to the agents in the tested situations. According to Baird et al. (Baird and Astington, 2004), 4-year-olds already differentiated morally between cases where the agent explicitly intended harmful actions and cases where they did not, but not as much as did 5-7-year-olds. This differentiation correlated positively with false-belief skills. This could mean

[86] They tested the difference between close contact, personal harm as a means and no-contact, impersonal harm as a foreseen side-effect made for moral judgements.

that the Intention Principle might be acquired by the age of 4 or even earlier but is not put to use because children are not able to ascribe intentions before that age.

According to Powell et al., 5/6-year-olds differentiated similarly between harmful actions and omissions as did older subjects, apart from those giving more equivalency ratings, but only in non-dilemma situations where they did not have to weigh harms (Powell et al., 2012). Knafo and Hauser seem to have found the same action/omission differentiation in children (see Abarbanell and Hauser, 2010, p. 213). Cushman et al. (Cushman et al., 2007) showed that there were age effects accounting for judgements in 10- to 60-year-olds, but those were only small and inconsistent.

As I have argued earlier, although children seem to apply the principles from the age of four or five onwards, to make a Poverty of Stimulus argument out of those data, we require evidence about moral input before that age as well as evidence that with the hypothesized learning mechanism that needs the least input, this moral input would not suffice to learn judging according to the pattern demonstrated by 4-year-old children.

Additionally, the data do not show a typical maturation line of moral judgements but only an onset; the gradual acquisition of judgements that differentiate between harm by action/omission in dilemmas and between intentional harm and actions with no harm intended seems to be contingent on other mechanisms such as cost/benefit calculation and intention ascription. Additionally, these contingencies show that there is no dedicated ‘moral judgement module’ that receives moral situations as inputs and gives moral judgements as outputs; this is because there seem to be other mechanisms at work that are used in economic calculations as well as social situations with no moral background (false-belief tasks). This applies for Functional Modularity as well as for Mechanistic Modularity, as the ‘moral module’ would be less evolutionarily advantageous in terms of quickness and automaticity if it feeds on other modules that deal with broader economic or social cognition.

5 Universality

5.1 What does Universality have to do with Innateness?

Proponents of the Linguistic Analogy have several reasons to believe that the principles underlying the Trolley Dilemma judgements are congenital: As discussed in the previous chapter, they argue that they are acquired early in development and they follow a fixed developmental trajectory. Beyond that, they argue that the principles are cross-culturally similar, just like Universal Grammar's basic principles, and that this is further evidence that they are innate. They are, however, often careful not to fully commit to any universality claims and the conclusions that might be drawn from this universality:

> "Hence it is perhaps not surprising to discover that thought experiments like trolley problems, which implicate these concepts, elicit widely shared moral intuitions from individuals of different cultural backgrounds. Nevertheless, while Experiment 3 provides some initial support for the existence of moral universals, this support is obviously quite limited. More empirical investigation on a much wider scale is necessary before specific claims about universality could be defensible." (Mikhail, 2002, p. 43)

But although they are careful not to make overly extreme claims, universality is one of the core properties for their paradigm of Universal Moral Grammar. The name of the theory is telling: It shows that according to the proponents of the Linguistic Analogy, the basic principles that underlie our sense of morality are universal. One of the roles universality plays in Universal Moral Grammar is, as I mentioned, to support innateness claims, but the connection is often only made implicitly. In the next few paragraphs, I will analyze several text portions that establish a connection between universality and innateness, and show how they do so.

Often, proponents of the Linguistic Analogy mention the principles' universality together with other features that make something prone to be congenital, such as their tacitness (in this case, the inability of many persons to account for their moral judgment pattern explicitly), and together with other cognitive functions that have been most discussed as candidates for evolutionarily developed mechanisms, as here in John Mikhail 2007a, p. 144:

> "[T]he moral judgments that these [Trolley] problems elicit are rapid, intuitive and made with a high degree of certitude – all properties that one associates with probes that are used elsewhere in cognitive science, such as language, vision, musical cognition and face recognition[20]. Moreover, the judgments appear to be widely shared among demographically diverse populations, including young children; even in large cross-cultural samples, participants' responses

> to these problems cannot be predicted by variables such as age, sex, race, religion or education[36]. Furthermore, individuals typically have difficulty producing compelling justifications for these judgments: thus, trolley-problem intuitions exhibit a dissociation between judgments and justifications[36,37], thereby illustrating the distinction between operative and express principles [5]." (Citing (Jackendoff 1994), (Cushman et al. 2007), (Cushman et al. 2006) and (Mikhail, Sorrentino, and Spelke 1998))

Evolutionarily developed cognitive mechanisms may have served a very specialized function that was evolutionarily advantageous and can be impaired selectively by brain injuries (with those injuries leaving other cognitive functions intact). Possible candidates include language (Pinker and Bloom, 1990), (Scott-Phillips, 2010) and face recognition (Barrett, 2008, p. 176). Here, Mikhail lists a number of features that makes Trolley Dilemma judgment mechanisms candidates for being adaptations, among them, being "widely shared among demographically diverse populations".

Here is a different example where Kirkby et al. cite those cross-culturally prevalent principles as reason to believe that morality has evolved evolutionarily:

> "For example, most natural human moral systems appear to be deontic in their basic structure and to depend on distinctions such as act versus omission, mistake of norm versus mistake of fact, and intended versus foreseen effects [...]. None of this is to deny the existence of substantial cross-cultural variation, only to suggest that this variation may be sharply limited. It is this kind of principled limitation to the domain - something hostage to further empirical inquiry - which can justify postulating an evolved sense of morality." (Kirkby et al., 2013, p. 95)

Kirby et al. suggest that these deontic principles [87] have developed evolutionarily because they seem to be part of "most natural human moral systems", hence: Because they are (mostly) universal. Hence Kirkby et al. establish evolutionary development from universality. Harman et al. take the opposite route: "If there is no other obvious way for the principle to be learned, the hypothesis suggests itself that the principle is somehow innate and should be universal." (Harman, 2008, p. 346)

Harman uses a Poverty of Stimulus argument here to argue that a principle (in this case, the Doctrine of Double-Effect) is innate: If it cannot be learned (here: By explicit instruction),[footnote: for more POS, look at introduction to POS chapter] the best explanation of how it was acquired is that it is innate. And if the principle is innate, Harman would make the prediction that it is universal. So he argues that the Doctrine of

[87] Note that two of them, which seem to be the ones Kirkby et al. consider as most important or best established, are Act vs. Omission and the Doctrine of Double-Effect, two of three principles I consider most important and best established and to which I therefore restrict my discussion in this thesis.

Double-Effect is innate and concludes that it is universal; here, as opposed to Kirkby et al., the innateness is first and the universality follows. In both accounts, they are related or even correlated.

I have shown that at least some proponents of the Linguistic Analogy have established a strong connection between ”innateness” and universality. I will present two reasons for that:

1. The first one is a historical reason: The Linguistic Analogy is an analogy between Chomsky’s Universal Grammar and Universal Moral Grammar. Chomsky, at least initially, claimed that his Universal Grammar consists of innate constraints that make it possible to learn a language. He also claimed that every human was endowed with those constraints. He thought that those constraints mirrored in all natural languages and some structural properties were common to all of them (Chomsky and Hornstein, 1980, pp. 69, 210/11, 232).

   So people who want to establish the Linguistic Analogy have a reason to search for these universal properties in morality as well. And, starting from the empirical side, finding moral principles that are universal in a similar way that linguistic principles are, according to Chomsky, would give them reason to believe in the Linguistic Analogy. And, since Chomsky held the linguistic universal principles to be innate, the universal moral principles might, in analogy, equally well be.

2. The second reason is conceptual: If you believe that something has developed evolutionarily, there is reason to think that it is universal (every human was born with it), too: If something has developed evolutionarily, which means it proved to be advantageous in its environment, it is very likely to become prevalent in that environment until everyone has it. And this holds for every adaptation. Therefore, most Evolutionary Psychologists believe in a homogeneous ’human nature’. If you believe that all human beings are endowed with the same innate ’faculties’ (dedicated computational structures ’built for’ solving certain ancestral problems), and you believe that morality is one of those faculties, you would obviously think that everyone was born with the same moral faculty.

Tooby and Cosmides write:

> ”One payoff of integrating adaptationist analysis with cognitive science was the realization that, in long-lived, sexually recombining species (like humans), complex functional structures will be overwhelmingly species-typical. That is, the complex adaptations that compose the human cognitive architecture must be human universals (at least at the genetic level), whereas variation caused by genetic differences is predominantly noise: minor random perturbations around the species-typical design. This principle allows cross-cultural triangulation

> of the species-typical design, which is why many evolutionary psychologists include cross-cultural components in their research." (Cosmides and Tooby, 2006, p. 60)

As I have shown, universality is one of the main inspirations for proponents of the Linguistic Analogy to treat Trolley Dilemma principles as something that has developed genetically (see, for instance, Kirkby et al., 2013, p. 95, quoted on 106 .) [88] I will first argue that this is based on a mistaken premise: Although different cultures show similar patterns in their judgements, they do not judge according to universally equal principles. To make this point, I will review the study to which the proponents of the Linguistic Analogy usually refer to show that its results do not strongly support an evolutionary thesis, even if we assume a "principles and parameters" account: An account implying that we are genetically determined to develop certain moral principles, but, depending on how we grow up, certain parameters develop differently, leading to moralities that have universally common traits but parameters that depend on people's upbringing. I will then argue that furthermore, the subjects of this study are not sufficiently diverse to make universality claims based on their judgements alone. I will therefore review additional studies that may be more adequate to test for cross-cultural universality.

5.2 Are the judgments in Cushman et al.'s 2007 paper really universal? Inner-cultural and cross-cultural distribution

Before I start this sub-chapter, I want to issue a warning about the concept of 'culture': In social sciences, during the last twenty years, 'culture' has been a difficult concept. Cultural essentialist paradigms that implicitly shaped most everyday uses of the word 'culture' have been shown to be problematic both scientifically and politically.

These paradigms claim that 'cultures' are substantially homogeneous collectives, stable over time and that the differences between 'cultural' groups *simply exist (or have been caused by upbringing)* and have 'caused' the 'cultural' group (Sökefeld 2007, pp. 31, 32). In many publications, belonging to a 'culture' is treated similarly to belonging to a 'nation': "By 'cultural essentialism' I mean a system of belief grounded in a conception of human beings as 'cultural' (and under certain conditions territorial and national) subjects, i.e. bearers of a culture, located within a boundaried world, which defines them and differentiates them from others." (Grillo 2003, p. 158). Opposed to this paradigm are constructionist accounts whose proponents hold that people *actively differ*, that they *make*

[88] For another example, see Hauser et al. 2008: "For example, why are some moral judgments relatively immune to cross-cultural variation? Are certain principles and parameters universally expressed because they represent statistical regularities of the environment, social problems that have recurred over the millennia and thus been selected for due to their consistent and positive effects on survival and reproduction? Is something like the principle of double effect at the right level of psychological abstraction, or does the moral faculty operate over more abstract and currently unimaginable computations?"(Hauser et al., 2008, p. 135).

themselves part of a cultural group: That cultures originate from discourse and social actions. Seeing themselves as belonging to a culture leads people to rest their identity on it and adapt their practices to what they see as what 'their culture' does (Sökefeld 2007, p. 33).

Wikan 1999 [89] illustrates nicely how scientifically problematic (or even unsound) some authors find cultural essentialism: "[This] notion of culture as static, fixed, objective, consensual and uniformly shared by all members of a group is a *figment of the mind* that anthropologists have done their share to spread." (Wikan 1999, p. 58, emphasis added) Grillo comments on this: "[...] however, and pace Wikan, far from promoting an 'old' vision of culture most anthropologists, and other social and cultural theorists, champion a 'new' version, at complete odds with the 'old'. So far from being essentialist most contemporary anthropological accounts of culture are quite the opposite, to the extent that they are in sharp conflict with the predominant common-sense view [...]". (Grillo 2003, p. 158, emphasis in the original).

The critique of cultural essentialism states, in short, that it is plainly empirically false, that 'cultures' are not a 'given' constant over time and 'cultural groups' do not have clear-cut boundaries, but that instead those boundaries are "human constructs, underdetermined by existing variations in worldviews and ways of life. [. . .] Essentialist representations of culture eclipse the reality that the labels or designations that are currently used to demarcate or individuate particular "cultures" themselves have a historical provenance, and that what they individuate or pick out as "one culture" often changes over time" (Narayan 1998, p. 5), in brief: That 'cultures' are not a natural kind. Sökefeld 2007 provides the following example: According to him, Sabine Riedel shows in her 2005 monograph (Riedel 2005) "The invention of the Peoples of Balkans [sic]" how "the process of the differentiation of 'nations' which themselves appeal to a history that goes back centuries, started only towards the end of the 19th century, initiated, among other things, by the importing of the Middle-European concept of the 'cultural nation' into the Balkans." (Sökefeld 2007, p. 33, translated from German by the author of this thesis).

The political issue with the essentialist concepts of culture(s) is that "the human species is broken down into self-contained, closed totalities" (Taguieff 1990, p. 117), cited after (Grillo 2003, p. 163)) which brings the danger of cultural differences becoming "'naturalized' and rendered 'totally unbridgeable'" (Policar 1990, p. 105, cited after Grillo 2003, p. 163), which can come with the need to preserve the community as is, or to 'purify' it. This development can, at the moment, be seen in extremist right-wing movements such as "Die Identitäre Bewegung" in Germany, who commit to an "ethno-pluralist", cultural-racist view advocating that people should stick to members of 'their own culture'. Hence many social anthropologist share the view that Wikan 1999 expresses with her title

[89] I find this essay problematic for reasons that are not part of the topic of this thesis. They do not affect the following quote. However, I want to distance myself from part of the essay's content.

“Culture: A new concept of race”.

Hence, it is difficult for me to adopt the concept of culture in this text, bearing in mind that research of the kind involves interviewing a group of people individualized by their nationality with the hypothesis that they collectively differ significantly in the moral judgements they make and because of their collectively similar upbringing. [90] Making nationality/culture which, as we have seen above but also in the studies I will cite, are often used synonymously, into the independent variable and, in our cases, making moral judgements into the depending variable is problematic to me, but might be solved (or eased) by doing the following: Adopting the concept of culture in this text, but only under the premises that the cultures I am speaking of are fluid, that collective differences (in ethnical/cultural/national groups) do not mean that the differences have been there and have ‘caused’ the “ethnical/cultural/national” group (see Sökefeld 2007, p. 33), that those groups are not continual over time and permanent, and that those cultures originate from discourse and social actions. I will not treat nation or culture as permanent entities (because of the issues I have just raised). I will, however, tentatively accept the hypothesis of the authors that the collective similarities among and differences between ‘cultures’ in growing up are significant enough in respects that might influence their moral judgement when I review them. I want to add that, if I can show in this part that people from different nations/cultures differ in their moral judgements, this is not necessarily evidence for collectively different upbringing that caused different moralities in cultural groups, but mainly an argument against the “cross-cultural” (or “inner-cultural”) homogeneity of moral judgements that proponents of Universal Moral Grammar claim. [91]

I will now commence the reviews. Cushman et al.’s major online study (Cushman et al., 2007) is the most influential set of Trolley Dilemma experiments and I have cited them frequently in this thesis. Importantly for universality claims, among their subjects, factors like nationality did not influence the pattern in which people judged. The judgements’ relative stability with respect to nationality and other external factors such as age (that I have discussed in the “Development” section of this thesis) was the main reason for many

[90] Wikan 1999, p. 475, suggests instead doing research by conducting case studies: “For example, we can and often do say things like ”The Bongo-Bongo are polygynous.” Yet one could refuse to generalize in this way, instead asking how a particular set of individuals – for instance, a man and his three wives in a Bedouin community in Egypt whom I have known for a decade – live the ”institution” that we call polygyny. Stressing the particularity of this marriage and building a picture of it through the participants’ discussions, recollections, disagreements, and actions would make several theoretical points.”

[91] Additionally, even if we assumed that collectively different behaviors between people from different nations originated through their belonging to those nations, Valery Chirkov, the author of “Fundamentals of Research on Culture and Psychology: Theory and Methods” (who calls this assumption, if falsely made (which, according to his sources, happens often), “cultural attribution fallacy”) recommends using “classifications [. . .] based on different foundations – religion, language, philosophy, and geography” as a “sampling frame for comparative psychological research of cultures”, suggesting that nationality is only one of the factors that might be the origin of collective differences in behavior. In accordance with this view, I will analyze whether Protestants and Catholics judge Trolley Dilemmas differently in the Poverty of Stimulus chapter (Chirkov 2015, p. 188).

proponents of Universal Moral Grammar to assume that the principles were universal and, hence, innate (as in John Mikhail 2007a, p. 144, cited above).

As a short reminder: Cushman et al.'s 2007 study tested for the Close Contact Harm Principle [92] and the Doctrine of Double-Effect Principle (but the sample was too small to test for the relation between nationality and judgments in accordance with the Doctrine of Double-Effect (see 5.2, p. 111)).

The dilemmas testing for the Personal Harm Principle had the following structure:

In the first dilemma, from now on called Switch Type case, the participants were asked whether it is permissible to turn a train on a side track and thereby save five persons but kill one. In the second one, from now on Push Type case, the participants were asked whether it is permissible to shove a large man in front of a train to stop it and thereby save five persons but kill one. The dilemmas for the Doctrine of Double-Effect Principle had the following structure: A train is rushing towards five people on a main track. Between the train and the people, a side track diverts and loops back. In both settings, a person is standing on the side track; but in Scenario 1, a heavy object is behind him. Both men and object are heavy enough to keep the train from going back to the main track and killing the five. The participants are asked: Is it permissible to throw a switch that turns the train onto the side track, and thereby kill one person and save five? See Figure 5, p. 39 and Figure 6, p. 39.

In the dilemmas that tested for Personal Harm (Push Type vs. Switch Type cases), 89% of Cushman et al.'s subjects voted for the Switch Type case (that involved no direct, close-up harm) to be permissible, and 11% for the Push Type case (that involved direct, close-up harm). Here, the subjects judged fairly unanimously, but not so much in the cases that tested for the Doctrine of Double-Effect (henceforth Loop with Heavy Object Type vs. Loop Type): 72% of the subjects voted for the Loop with Heavy Object Case to be permissible and 56% for the Loop Case (all data in this paragraph: Cushman et al., 2007, pp. 6–8). The study is not perfect for examining whether moral judgements are made universally for (at least) two reasons:

1. Whether nationality influences the judgements could only be tested for the Close Contact Harm Principle (Push Type vs. Switch Type cases). In the scenarios testing for the Doctrine of Double-Effect Principle (Loop with Heavy Object Type vs. Loop Type), "chi square analysis could not be run for subpopulations defined by ethnicity or national affiliation because in each case only one subpopulation had an expected value exceeding 5." (Cushman et al. 2007, 11).

2. It is questionable whether the interviewed subjects and their nationalities were

[92] The Close Contact Harm Principle says that harms caused by a combination of personal force (direct muscle impact, as in pushing someone with a pole) and closeness are worse than harms caused by a combination of impersonal force (as in dropping someone through a switch-operated trap door) and spatial distance (Greene 2014, p. 23).

diverse enough to provide any strong evidence for cross-cultural universality; I will elaborate on this point later.

But even if we ignore these two points, we will find that the judgements were not that universal at all:

It is true that the judgement *patterns* were the same across nationalities: The ratio of subjects who found an action permissible to subjects who found an action impermissible in one dilemma scenario was the same across nationalities. Say, 89% of all subjects from Canada found it permissible to throw the switch in the Switch Type Case, but only 11% found it permissible in the Push Type Case. According to Cushman et al.'s study, the same (or nothing significantly different) holds true for all subjects from the United States and all other tested nationalities. I will call the distribution of judgement patterns *between* cultures the **cross-cultural** distribution. In the following, I will provide an example of cross-culturally different distributions: If in a study with thousands of subjects 10% vs. 90% of Canadian subjects judged an action permissible vs. impermissible, while 40% vs. 60% of US-subjects [93] judged it permissible vs. impermissible, they would display cross-culturally different judgement distributions.

We are only discussing the distributions of judgements (what I have called the "judgement pattern") here, so even if 50% of the subjects of both nationalities decided that an action is permissible, while 50% decided it was impermissible, i.e. the subjects showed the largest possible disagreement within their nationalities, the cross-cultural judgement distribution or pattern of Canadian and US subjects would still be the same, as subjects from both countries would equally disagree about the permissibility. This brings us to the next point:

It is also true that the *majority* of subjects judged according to the Close Contact Harm Principle: 89% of all subjects found an action permissible that required no direct physical contact and whose bad effect was a side-effect of the good effect, while only 11% found an action permissible that required direct physical contact and whose bad effect was a means to the good effect. Hence, a large majority of all subjects found an action permissible that is allowed according to the Close Contact Harm Principle and the Doctrine of Double-Effect Principle, while only a small percentage found an action permissible that was forbidden according to both principles.

This majority vanishes in cases that only test for the Doctrine of Double-Effect: If we compare the ratio of people who find it permissible to act in the Loop with Heavy Object Type vs. Loop Type case, we will find that it is 72% vs. 56%: Although more people find it permissible to act according to the Doctrine of Double-Effect, 56% (more than half of the subjects) would just as well find it permissible to act against it.

And this brings us to one of the core flaws of using Cushman et al. 2007's study as evidence for universal moral judgements: The subjects did not universally and equivocally

[93] Or whatever distribution makes a significant difference; this varies with the number of subjects.

judge in accordance with certain principles. As noted above, the percentage of people who thought an action was permissible that violated the Doctrine of Double-Effect was almost 56, hence, there was a huge (almost the largest possible) disagreement *within* the subject group: Within the group, subjects were far from making universally similar judgements. When doing within-subjects tests to see how many subjects actually stated it was impermissible to kill the person on the loop track when there was no stone AND it was permissible to kill the person on the loop track when there was a stone behind them, Cushman et al. found that only 33% differentiated between those cases in a way that could be attributed to the Doctrine of Double-Effect (hence, found an action that was reconcilable with the Doctrine permissible and a very similar action that violated the Doctrine impermissible): Only 33% (Cushman et al. 2007, p. 15) seemingly differentiated between actions solely because one of them was in agreement with the Doctrine of Double-Effect and the other was not. [94] Again, far from making universal judgements (which, as Universal Moral Grammar's story goes, is evidence for their genetic determination (or disposition) to judge according to the Doctrine of Double-Effect), only one-third of the subjects judged according to the Doctrine.

Hence, 'inner-culturally' (e.g., among all US participants), subjects did not agree on the permissibility of sacrificing one to save many in the iconic Trolley Dilemmas. I will call the distribution of "permissible" vs. "impermissible" votes in a population that is considered by the authors to belong to one culture, e.g. all people from the USA, the rate of **inner-cultural** agreement. If, for example, 56% of them considered an action permissible, while 44% considered it impermissible, this would mean that the population would not judge homogenously, even if the majority considered the action permissible, and inner-cultural agreement would be low.

Fitzpatrick 2014 raises and in the same sentence dismisses a similar issue: "Another concern is that the results usually reveal a significant minority opinion as well as a majority one [...]. This is, however, a pervasive phenomenon in psychological research, and thus not especially problematic for making general claims about moral psychology [...]." (Fitzpatrick 2014, p. 507). This might be true for the Switch/Push Type case distinction (which tests for several principles at the same time), but does not hold for the Loop/Loop with Heavy Object case distinction: 56% vs. 72% permissibility are both majority opinions, although one of them is not in accordance with the Doctrine of Double-Effect.

5.3 Are the judgement patterns the same cross-culturally?

As we have seen in the last chapter, although the inner-cultural disagreement was high, people of the six tested nationalities did not significantly differ in terms of their judgement

[94] Although even this decision could be attributed to other causes, see, for instance, Waldmann and Wiegmann 2010.

patterns in Trolley Dilemmas according to Cushman et al.'s study: Nationality did not significantly influence the ratio of yes-no-judgements. [95] While it is true that in this study, the authors did not find any significant effect of nationality on the judgement pattern, they only had enough subjects from the following countries to test for nationalities' impact on judgements: Australia, Brazil, Canada, India, United States and United Kingdom (Cushman et al., 2007, p. 9). This list is neither comprehensive enough to justify universality claims, nor are these countries' citizens representative for members of maximally diverse cultures. In a later paper, Linda Abarbanell and Marc Hauser, who also happens to be one of the authors of the Cushman et al. 2007 paper, refer to the latter and write: "[...] it is clear that this internet sample is insufficient for testing questions of universality and cross-cultural variation. In particular, although these studies originally sampled thousands of individuals, most were from English-speaking countries, most were formally educated, and many had read books discussing moral issues. Further, all were technologically savvy, at least in terms of their ability to find, log in, and navigate computer software designed to access the world wide web." (Abarbanell and Hauser, 2010, p. 208).

If the interviewed populations were too similar to provide (sufficient) evidence for cross-culturally universal moral judgements, how about populations that differ more in their upbringing? Do they exhibit similar moral judgement patterns to the ones Cushman et al. found? In this chapter, I will review literature apt to examine this: Papers that

1. test for the same moral principles as the iconic papers the Universal Moral Grammar proponents have cited and relied on, to check whether the following principles are universal as they hypothesize: The Close Contact Harm Principle, the Action/Omission Principle and the Doctrine of Double-Effect Principle and

2. whose authors claim that the subjects have a sufficiently different cultural background in contrast to, for instance, Cushman et al. 2007's study to substantiate universality claims in case they make similar moral judgements.

I will be reviewing most papers available at this time that satisfy these two conditions, as the tested populations vary in different aspects that the respective authors think may be responsible for moral judgements diverging from the ones Cushman et al. 2007 found.

I will start by reviewing a Trolley Dilemma Type study that was conducted with Hadza from Tanzania as it is the least elaborate, followed by a study by John Mikhail that is more persuasive. I have included the study's results, although they are preliminary, in this review because, firstly, "cross-cultural" studies researching Trolley Dilemma Type cases are rare, and secondly, some authors explicitly use them to back up universality theses:

[95] At least in the dilemmas that could be tested for influence of nationality; this was not possible for the dilemmas that tested only for Doctrine of Double-Effect. See 5.2, p. 111

"Though preliminary, these results provide further support for the universality of some of our moral intuitions." (Hauser et al. 2008, p. 136).

The study fulfils our two conditions that make studies interesting for universality claims about moral judgements in Trolley Dilemma cases: Firstly, the dilemmas were designed to be comparable to the original trolley cases and thereby test the same principles as they do. And secondly, the authors call the subjects "a small and remote group of hunter-gatherers living in Tanzania" when they contrast them to "our English-speaking, Internet-sophisticated, largely Westernized and industrialized subjects" (Hauser et al. 2008, p. 135), suggesting that their lifestyles differ sufficiently from their usual subjects to make universality claims.

According to Marc Hauser, preliminary results of the study show a trend for the subjects to make the same distinctions in terms of close contact harm as Cushman et al. 2007's subjects did (Hauser et al. 2008, pp. 135, 136). Since the authors expected the subjects to be unfamiliar with trolleys, they designed a study with stampeding elephants that were about to run over and kill five people; in the Switch Type case, a man in a jeep can either do nothing or drive toward the elephants, "turning them away from the five and around the grove where they will run over and kill one person." (Hauser et al., 2008, p. 136). In the Push Type case, "a person can throw a heavy person out of a tree to stop the herd and thereby save the five people." (Hauser et al. 2008, p. 136). In an older online version of the paper (https://pdfs.semanticscholar.org/6a26/55ecdc14cb54abaccac234eeba544b279e22.pdf, acc. 1/8/17), Hauser et al. mention that 12 out of 15 Hadza judged "these cases as do web-savvy westerners", (ibid., p. 298), which I suppose means that they judged the action to be permissible in the Switch Type case and impermissible in the Push Type case. In a later version, the authors do not mention these numbers anymore.

Those data, however, have not been published yet and, apart from the issue of Hauser's trustworthiness (only animal studies seem to be affected by his fraud (Carpenter 2012), so his Trolley Dilemma studies should be sound), the study has several problems: The sample is very small and herds of elephants, unlike trolleys, can take different paths and behave very differently than trolleys that drive in a track. Although throwing a man out of a tree is similarly unlikely (and unconventional) to stop a stampeding elephant herd as throwing a man in front of a trolley is unlikely to stop the trolley, we do not know whether the absurdness of this action would affect the judgement patterns. Additionally, the study has, to my knowledge and for reasons unknown to me, not been published. Hence, these results may speak for the universality of the Close Contact Harm Principle, but should be taken with a grain of salt. However, two studies conducted by John Mikhail ((Mikhail, 2011), originally: (Mikhail, 2002)) seem to confirm them by yielding similar results: When he tested 39 subjects "from the broader Cambridge, Massachusetts, community, all of whom had emigrated from China within the previous five years and most of whom had done so within the previous two years", 79% found it permissible to throw the switch in the Switch case and 14% to push the man from the footbridge in the Push case ((Mikhail, 2011, p. 331/332) and (Mikhail 2002, p. 107)). In comparison, 65 subjects whose majority were "United States citizens or members of other Western nations" (Mikhail 2002, p. 39) from Harvard University and from Cable News Network were asked the same questions (Mikhail, 2011, p. 325) and judged very similarly: 76% found it permissible to actively sacrifice one in the Switch case and 8% in the Push case (ibid., p. 326).

These experiments, too, fulfil the two conditions for papers that are useful for our universality case: Subjects from China are contrasted to "United States citizens or members of other Western nations" and the scenarios are "modelled on the Transplant and Footbridge Problems" [96] and "on the Trolley and Bystander Problems" (Mikhail 2002, p. 19). Mikhail writes that the results "suggest at least some operative principles of moral competence [. . .] are transnational and may be universal." (Mikhail 2002, p. 45).

Hence, not only did the results show that people's judgements were distributed in a similarly cross-cultural way (the ratio of subjects who found the Switch Case action and the Push Case action permissible vs. non-permissible was similar amongst subjects from China and subjects with unknown background), but they also showed some inner-cultural homogeneity: In both groups, a large majority found the Switch Case action permissible and the Push Case action non-permissible.

Unfortunately, however, John Mikhail only contrasted Push Type and Switch Type actions. The results are therefore conflated results for the Close Contact Harm Principle and the Doctrine of Double-Effect: The harm inflicted by a Push Type action is characterized by close, direct contact with the victim AND using the victim's body as a means to save the five, whereas the Switch Type action has none of those features. It would have been very useful to have a Loop Type case as contrast, too, i.e. to have a contrast where the victim is used as a means, but not killed by direct, close-up contact. However, Switch Case/Push Case contrasts usually elicit very polarized responses and the fact that those seem to be stable across subjects of mainly US/"Western" and Chinese provenience is an interesting result.

There are, however, three studies that contradict the evidence above which shows similar judgement patterns regarding the harm distinction in very diverse populations.

Firstly, Henrik Ahlenius' and Torbjörn Tännsjö's team presented Trolley Type Dilemmas to 3000 subjects, interviewing Russian and US inhabitants on the phone and inhabitants of China in a pen-and-paper survey (Ahlenius and Tännsjö, 2012). The study is not only very interesting in light of Mikhail 2002/2011's study contrasting Chinese and "Western" subjects, but also meets our two criteria that make the study interesting for

[96] Where the so-called "Transplant Problem" has the features I used to categorize a scenario as Push Type cases: The harm is inflicted from a close range, by direct contact with the victim, and the victim's body (or, in the Transplant Problem, the victim's body parts) is/are used as a means to save five others.

universality issues: Albeit only in a footnote, the authors tie their results (that I will come to shortly) to universality, suggesting that although the judgements were not universally homogeneous, “what remains universal [. . .] is the *ranking* of the options” (Ahlenius and Tännsjö 2012, p. 5, italics like the original), hence implicitly saying that the subject groups came from sufficiently different backgrounds to warrant claims about universality. And the scenarios are versions of the Switch, the Push and the Loop Cases, making them very comparable to the studies I have investigated previously in this chapter.

The subjects were asked whether they should perform an action that sacrifices one person in the respective dilemmas and their answer options were ”Yes”, ”No” or ”I don’t know”.

The Switch Dilemma yielded the following results:

> 81% of the US inhabitants answered ”Yes” when they were asked whether they should flip the switch in a Switch Dilemma, 13% answered ”No”.
>
> 63% of the Russians answered ”Yes” and 20% ”No” and
>
> 52% of the Chinese answered ”Yes” and 36% ”No”.

The Push Dilemma yielded the following result:

> 39% of the US inhabitants answered ”Yes” when they were asked whether they should push the big man, 56% answered ”No”.
>
> 36% of the Russians answered ”Yes” and 45% ”No” and
>
> 22% of the Chinese answered ”Yes” and 68% ”No”.

The Loop Dilemma yielded the following result:

> 60% of the US inhabitants answered ”Yes” when they were asked whether they should flip the switch and have the big man killed in order to save the five, 32% answered ”No”.
>
> 54% of the Russians answered ”Yes” and 23% ”No” and
>
> 34% of the Chinese answered ”Yes” and 52% ”No”

(all of the results in the paragraph above: Ahlenius & Tännsjö 2012).

This means that the preference order of judgements was the same as in the previous study: Participants of all nations seemed to find it less indicated to kill someone as a means to saving five than to kill the person as a side-effect and they found killing the person even worse if close contact harm was involved. The percentage of US inhabitants who agreed to flipping the switch in the Switch Type Dilemma and in the Loop Type

Dilemma was fairly close to the percentage of people who agreed to those actions in the Cushman & al. 2007 study (Cushman et al., 2007) with 89% (Cushman et al., 2007, p. 7) vs. 85% (Ahlenius & Tännsjö 2012) in the Switch Type case and 56% vs. 60 % in the Loop Type case; only the Push Type case differed with 11% (Cushman et al. 2007) vs. 39% agreement (Ahlenius & Tännsjö 2012) among US inhabitants. The agreement percentages from Mikhail 2002/2011 were closer to Cushman et al. 2007, with 76% of "Western" participants judging the Switch action permissible, but only 8% agreeing to the Push Type action.

The Russian and Chinese inhabitants showed the same order of agreement with highest agreement rates to the Switch Type Case, lowest to the Push Type Case and the Loop Type Case in between. However, inhabitants of Russia and even more so inhabitants of China agreed less to the respective actions overall: A smaller percentage of subjects found that they should sacrifice one person to save five in all cases, compared to the US citizens. Hence, while the Close Contact Harm Principle and the Doctrine of Double-Effect seem to play a role in moral judgements of inhabitants of all three nations, the study indicated that Russians and even more so Chinese are more reluctant on average to sacrifice one in order to save five.

However: The methodology of this study, too, is problematic: The Chinese subjects were asked to answer in a pen-and-paper survey, sitting at the same table as the person interviewing them. [97] The US inhabitants and the inhabitants of Russia were asked whether they should perform the respective action to sacrifice one, whereas the Chinese were asked: "Would it be morally permissible to ...?" (A procedure that has been criticized by Gold et al., 2014, p. 66). Answers to "should you" questions did, according to Gold et al., "not map neatly onto" moral [permissibility] judgements ((Gold et al., 2014, p. 66), citing (Gold et al. 2015)).

This is a valuable critique, but Gold et al. 2014 replicated the tendencies of Ahlenius & Tannsjö's experiment for Chinese citizens in comparison to British citizens (who, in Cushman et al. 2007, judged similarly to US citizens), albeit with real-life-scenarios: [98] They found that the Chinese participants were less likely to sacrifice one to save five.

Interestingly, in a follow-up experiment in the same paper with a hypothetical Switch Type case (that they were told to imagine) and the question "Would you pull the lever? (Yes/No)", followed by "Now please indicate how wrong or how right you think it would be to switch the lever: (*1 Definitely wrong to 7 Definitely right*)" and "Is it morally wrong for you to switch the lever? (Yes/No)", 76.36% of the British and 64.44% of the Chinese subjects "said that they would pull the lever to save the lives of the five people."

[97] Ahlenius and Tännsjö remark that earlier studies did not show any differences between data that were obtained per phone call and data obtained personally.

[98] This is why I will not broadly review them here: It is hard to compare real-life scenarios to thought experiments, and incorporating all of them, including how they differ from thought experiments, would go beyond the limits of this thesis.

(Gold et al., 2014, p. 70). The "wrong-right" ratings showed no significant differences between British and Chinese participants and 45.45% of the British as opposed to 35.56% of the Chinese participants said it was morally wrong to pull the lever (Gold et al., 2014, p. 71/72). These ratings differ significantly from those obtained in previous studies for the British participants: If 45.45% of the British said it was morally wrong to pull the lever, 54.55% at most found it morally right (and the rest found it neutral), whereas the Cushman study had found the British to judge just do as the other participants, with 89% judging it "morally permissible" to pull the lever in the Switch Type case (Cushman et al., 2007, p. 7).

These effects could be due to the sample size (55 British and 45 Chinese participants only) or to the limited scope of the recruitment strategy: The participants "were recruited through the University of Leicester's online e-bulletin, which is sent out to students and staff." (Gold et al., 2014, p. 69). Another cause for the divergent results may be the sequence of very similar questions that could have shifted participants' attention to the word "morally" because the different wordings forced them to reflect the difference in meaning ("What's the difference between 'How wrong or how right would it be to switch the lever' and 'Is it morally wrong for you to switch the lever'?") and the nature of the Trolley Type case: Two out of three were scenarios produced by humans rather than (un)fortunate circumstances: In one of those (hypothetical) scenarios, the agent could choose on a computer screen whether one orphan child should be deprived of one meal while five other orphans get one or the one orphan child keeps their meal while the five others end up empty-handed. These results cast doubt on the previous studies' results and call, again, for further follow-up experiments. If its results are reproducible with more controlled-for scenarios, this means that "British" subjects do not judge as other "Western" subjects but tend to endorse killing one to save five less than the latter. This would make their judgements more similar to "Chinese" people's judgements as in the previously studies cited, but distinguish them from many other groups' judgements.

As the methodology of the Ahlenius & Tännsjö 2012 study was far more similar to the other experiments reviewed here and had a larger number of participants than Gold et al. 2014, I will treat the 2012 study as more relevant to our purposes.

In another relevant study, Linda Abarbarnell and Marc Hauser tested for the Close Contact Harm Principle in a rural Mayan society. [99] They reported that their goal in doing so was "to extend existing empirical work on the nature of moral intuitions that, up to the present, has largely focused on Western, industrialized populations [...]." (Abarbanell and Hauser, 2010, p. 217). 30 adult subjects participated (Abarbanell and Hauser, 2010, p. 211).

The study fulfils the conditions cited above: Abarbanell and Hauser used dilemmas

[99] From parajes (areas) of the Cañada Chica, Cañada Grande and Pajalton (Abarbanell and Hauser, 2010, p. 210).

that tested for the Close Contact Harm Principle, the Action/Omission Principle and the Doctrine of Double-Effect Principle (p. 211, p. 212, table 1), just as, for instance, Cushman et al. 2007 did. And, as the quote above shows, the authors regarded the population they tested as sufficiently different from the populations other authors had tested. However, unfortunately, the published data are not directly comparable to studies testing aforementioned “Western, industrialized populations”: With their 5-point scale design that assigned a value of 1 to the Ts’eltal [100] translation of ”very impermissible”, 2 to ”a little impermissible”, 3 to ”a little good”, 4 to ”regular good” and 5 to ”very good”, Abarbanell and Hauser had a scale similar to Schwitzgebel and Cushman 2012’s study (Schwitzgebel and Cushman, 2012). The latter tested for “Order Effects on Moral Judgment in Professional Philosophers and Non-Philosophers” with very similar dilemma scenarios as Cushman et al. 2007, and their scale ranged from “(1) ‘extremely morally good’ to (7) ‘extremely morally bad’ with the midpoint (4) labeled ‘neither good nor bad’” (Schwitzgebel and Cushman, 2012, p. 138).

Cushman et al. 2006 (Cushman et al., 2006) could have provided another comparable experiment with their 7-point-design that ranged from 1 (“forbidden”) over 4 (“permissible”) to 7 (“obligatory”) (Cushman et al., 2006, p. 1083). But they neither published how many of the subjects had assigned equivalent permissibility values to actions in the respective dilemmas (e.g. the value ’3’ to taking action in both the Push and the Switch Type Case), as Schwitzgebel and Cushman did, nor did they publish the percentage of subjects who judged on the “permissible” or “impermissible” side of the scale. The latter makes the study hard to compare to studies that provided binary choices like Cushman et al.’s 2007 study, in which the subjects could only choose whether an action is permissible or impermissible. [101] It would have been possible to compare the study to Cushman et al. 2006b’s results by using Andrew Colman et al. 1997’s equations [102] (Colman, Norris, and Preston 1997).

However, their experiments differed from Abarbanell and Hauser’s not only in the extension of their scale but also in other aspects: Cushman et al. 2006b had very different labels on at least two points of their scale (“permissible” for the midpoint and “obligatory” for one pole of the scale in contrast to Abarbanell and Hauser’s “a little good” for the midpoint and “very good” for the end point), the wording of their dilemmas differed, and their subjects were tested in the company of other subjects (Abarbanell and Hauser 2010, p. 215). Furthermore, according to Andrew Colman et al., data drawn from 5-point Likert

[100] Ts’eltal is the main language of the Mayan communities from which the subjects came.

[101] This, along with most proponents of Universal Moral Grammar citing Cushman et al. 2007 and their greater number of participants, is also the reason why I do not discuss the study at length in this chapter. Their findings are consistent with Cushman et al.’s 2007 findings: People significantly preferred actions that were in accordance with the Close Contact Harm Principle, the Action/Omission Principle and the Doctrine of Double-Effect Principle than those that violated it. For further information, see Cushman et al. 2007b, p. 1084, table 1.

[102] $X7 = (1.47 \pm .23)x5 - (.40 \pm .15)$ $X5 = (.68 \pm .10)x7 + (.27 \pm .11)$

scales could only account for around 76% of the variance when transformed to 7-point Likert scales and vice versa even with identical questions. Considering this, it might be more useful to take the following approach:

We can review which actions Abarbanell and Hauser 2010's subjects preferred over others ordinally and whether those results compare to other studies' results. On the 5-point scale that assigned a value of 1 to the Ts'eltal translation of ”very impermissible”, 2 to ”a little impermissible”, 3 to ”a little good”, 4 to ”regular good” and 5 to ”very good”, the mean values assigned by subjects to the respective dilemmas were the following:

They assigned the Switch Type case a mean value of 4.07, the Push Type case 1.63, a case that is similar to a Loop Case [103] 1.90 and the same value of 1.90 to an omission that otherwise had the same features as the Loop Type action (Abarbanell and Hauser 2010, p. 211/212). Hence, (many of) the subjects had a strong inclination to judge the Switch Type case as morally permissible (4.00 would have meant ”regular good”) and the Push Type case as impermissible (a value between ”very impermissible” and ”a little impermissible”) with the Loop Type Case in between (but still on the ”impermissible” side); the preference order is the same as in Cushman et al. 2007's (Cushman et al., 2007) results. The Doctrine of Double-Effect seems to be morally relevant to the subjects of this study: On average, the subjects judged the Loop Case as much less permissible than the Switch Case. Abarbanell and Hauser did, however, not “find evidence for a distinction between harm caused by contact and harm caused by non-contact” (Abarbanell and Hauser 2010, p. 217), hence, the Close Contact Harm Principle. In this and a further condition, the “rural” subject group did not judge actions and omissions significantly differently (other than most previously tested population samples). Abarbanell and Hauser further replicated this pattern with two more conditions which tested for the Doctrine of Double-Effect and the Action/Omission Principle, but did not follow up on the Close Contact Harm Principle. Hence, both the preference for harm as a side-effect over harm as a means (Doctrine of Double-Effect) and no significant preference for “harmful” omissions over harmful actions (Action/Omission Principle) proved to be rather stable in “rural” subject groups. I will examine this later in chapter 5.7, p 134 and first focus on an intercultural comparison of judgements with respect to the Close Contact Harm Principle. Unfortunately, the “urban” subject group was not tested on the Close Contact Harm Principle, but does draw a distinction between harmful actions and “harmful” omissions. This shows that features of subjects such as their access to the internet, being more educated and living in larger cities are more likely to have made the difference between judging in line with or against the Action/Omission Principle than their identity as Maya. I will return to the Action/Omission Principle later in this chapter.

The most spectacular finding of this study is that the “rural” group did not significantly endorse the Close Contact Harm Principle: Not only proponents of Universal Moral Gram-

[103] Causing someone's death as a means to saving five without touching the victim.

mar, but also many proponents of other kinds of innate 'moral sense' have hypothesized this to be one of the very core principles of human morality, [104] which makes it even more unfortunate that (and, possibly, makes us wonder why) Abarbanell and Hauser did not follow up on their results. To sum up, while many of the previous studies (most of them designed best to test the Close Contact Harm Principle and the Doctrine of Double-Effect) indicate that the principles that matter to people in moral judgements might be the same cross-culturally, because their preference order for harmful actions in Trolley (Type) Dilemmas is the same in most cases, the 2010 study by Abarbanell & Hauser yielded different results: The Mayan group they interviewed that was less involved with the nearby city judged neither in line with the Close Contact Harm Principle nor in line with the Action/Omission Principle: They did not prefer actions where harm was inflicted from afar and without direct contact to actions harming people by directly touching them. Nor did they prefer harmful omissions to harmful actions. While the authors have replicated the latter tendency, they did not follow up on the former one and the sample size was rather small. The results, however, contest the other studies whose results were more uniform 'cross-culturally'. This calls for more experiments testing different demographic factors' impact on moral judgements, such as size of the city/village people live in, internet access, extent to which people feel responsible for their peers and others. The only other study testing people who live in a smaller-scale community, mentioned by Hauser et al. 2008 on p. 298/299, has not been published and had an even smaller sample size than Abarbanell and Hauser's 2010 study. Furthermore, the last study was not the only one to contest a consistent, cross-cultural preference order in accordance with the Close Contact Harm Principle and the Action/Omission Principle.

5.4 Inner-cultural agreement rates vary cross-culturally

Another challenge tor Universal Moral Grammar's universality claims goes as follows: The inner-cultural homogeneity of judgements differs cross-culturally and even for the same population group (British) in different experiments: Inhabitants of China and Russia showed higher inner-cultural disagreement than US inhabitants in Ahlenius & Tännsjö's 2012 study. While "British" subjects had similar disagreement values as the former in Gold et al.'s 2014 study, their disagreement values were similar to those of US inhabitants in Cushman et al. 2007. Although the value of the ratio of inhabitants of China who found actions permissible to those who found them impermissible was closer to 1 (or 50:50) than that of subjects from other nations (hence the population was close to being

[104] See, for instance, older papers by Joshua Greene: "[...]I have at times [J. Greene 2008; Greene and Haidt, 2002] suggested that the negative reaction to causing "personal" harm, as in the footbridge case (Greene et al., 2001, 2009), reflects a domain-specific, innately supported affective response." But now, as he recently wrote, his view has shifted and he is "increasingly convinced that learning plays a dominant role in generating these patterns of judgment." (both quotes: Greene 2017). This, too, supports the findings of this thesis.

split in half judgement-wise), they agreed more about dismissing the Push Type action than US inhabitants. These values are a symptom of the shift towards finding killing one person impermissible in all scenario variants in subjects from China, Russia and, in Gold et al. 2014's results, British subjects, all in comparison to US subjects. Hence, not only were more subjects from those countries inclined to judge against harming one person than their US equivalents, this also led to great inner-cultural disagreements about, for instance, the Switch Type actions: In Ahlenius & Tännsjö's 2012 study, while most US participants uniformly agreed to them (81%), a shift towards disagreement in subjects from China means only 52% of them found them permissible while 36% found them impermissible, splitting the population nearly in half. But even the studies with relatively high inner-cultural agreement rates do not necessarily mirror inner-cultural agreement rates in grammaticality judgments. Hence, in this aspect, the Universal Moral Grammar analogy seems to be flawed:

In the Push Type case in the Cushman et al. 2007 study (Cushman et al., 2007), a relatively large proportion of subjects seemed to agree on the Close Contact Harm Principle (or at least judge it identically): 89% stated it was acceptable to turn the train on a side track, thereby running over one and saving five. Only 11% thought it was acceptable to shove a large man in front of a train who then gets hit and killed, and thereby save five. In both cases, a large majority of participants seems to agree on those judgements. If we compare this to grammaticality judgements in a language community, 89% is not such a large proportion: Usually, almost 100% of the native language speakers tend to agree to the grammaticality of sentences. [105]

Notwithstanding the Linguistic Analogy, the results might still hold as evidence for universal moral principles: We could take those results as evidence that in many folk moralities, close contact harm is not acceptable in the opinion of a large majority of people, [106] while sacrificing a person to save five is acceptable to most if the person who gets killed is standing on a side track. If we judge by majority, according to some studies, there seems to be a 'right' answer to those Push and Switch Type Cases.

We have applied two measures that indicate whether moral principles are universal: Cross-cultural preference orders and inner-cultural agreement rates.

The latter is the percentage to which a population finds the action in one single dilemma case permissible. As mentioned before, in Cushman et al. 2007's Switch Case, 89% found it permissible to throw the switch, whereas the others had a different opinion. Hence,

[105] Although, of course, people tend to disagree about the grammaticality of some sentences: Linguistic grammaticality seems to be a property that can come in degrees (e.g. Chomsky's famous sentence, "Colorless green ideas sleep furiously" (Chomsky, 1957, p. 15) seems to be more grammatical than sentences without grammatical structure as "Sleep ideas green colorless furiously", but less grammatical than a sentence that actually makes sense)

[106] Albeit note that in the Push Case, not only the Close Contact Harm Principle can be applied, but also the Doctrine of Double-Effect, which might be one of the reasons most people tend to not find it permissible to shove the man.

there was very high agreement amongst the study participants that it is permissible to turn a train rushing towards five people onto a side track where it is only going to kill one. This measure shows how permissible it is to sacrifice one in order to save many combined with the impact of the principle: In Push Cases, people who do not find it permissible to sacrifice anyone at all are not going to vote that it is permissible, just like people who find sacrificing one permissible in principle but who find it impermissible to sacrifice one in order to save five when one person is pushed in front of a train in the process. Furthermore, people who judge according to the Close Contact Harm Principle as well as people who judge according to the Doctrine of Double-Effect will not find killing one in Push Type scenarios permissible. Hence, the 11% who found it permissible to kill the one in the Push scenario in Cushman 2007 were the ones who find it permissible to kill someone in principle AND who do NOT judge according to the Close Contact Harm Principle AND who do NOT judge according to the Doctrine of Double-Effect. Because if either of those components would have had veto power, they would have judged the action impermissible.

To further disentangle the effects of each of those components, I will introduce a third measure:

5.5 A different angle: How many people judge two dilemmas equivalently?

Because it is difficult to compare yes/no(/I don't know) reply options to Likert scales and Likert scales with different amounts of points to each other, earlier in this chapter I decided to only compare the order of preference between experiments with different answer options. This is reflected in table X as well, where I have a section with binary (yes/no) and three-option (yes/no/I don't know) experiments and one with Likert scales.

Now, however, I will draw a quantitative comparison between inner-cultural agreement rates for different dilemma pairs, hence comparing which ratio of each population sample (in the same study) judged two dilemmas the same/differently.

To find out what the impact of a principle is as opposed to the sacrificing part, one could, for instance, use the Switch Type Dilemma as a base line: The most obvious reason why people seem to disagree with sacrificing one person is that they would not be willing to kill to save five people. So if we are looking at a Push Type Dilemma, we can assume that the people who found it impermissible to kill one in order to save five are going to vote for 'impermissible' anyway. Those people who additionally vote for 'impermissible' are likely going to be the people who find it permissible, in principle, to sacrifice one to save five, but were repelled by the additional feature that someone pushed the victim off a bridge. [107]

[107] Or, but less likely, the loss of the calming fact that the victim was standing on a sidetrack. All of this, of course, only holds if we assume that the other factors such as wording effects are sufficiently controlled for between scenarios. In experiments with binary answering scales, we might not catch the people who

Hence, the second measure for the homogeneity in a population's opinions about Trolley Dilemmas and the third measure for reviewing evidence concerning universality claims is the percentage of people who judge two dilemmas differently: It shows how many people employ the principle that systematizes the differences between those dilemmas, e.g. the Close Contact Harm Principle in the Switch Type/Push Type pairings. The large difference between 89% permissibility ratings in the Switch Case and 11% permissibility ratings in the Push Case observed in the Cushman et al. 2007 study shows that many people seem to judge in accordance with the Close Contact Harm Principle, if we assume that the discrepancy between permissibility rates in the two dilemmas is an effect of the "Close Contact Harm" features which the latter displays. [108] Hence, only 11% of each population sample in Cushman et al. 2007 judged both scenarios equivalently permissible while 11% of each population sample did not judge them permissible; if we assume that they judged on a binary scale with "permissible" and "impermissible" as judgment options, the total agreement rate would be 22% (11% who judged the Push Case action permissible and 11% out of the 89% who judged the Switch Case action permissible plus 11% who did not judge the Switch Case action permissible and 11% out of the 89% who did not judge the Push Case action permissible).

These results, however, might be less stable than they seem to be. In a different experiment with a 7-point scale ranging from "(1) 'not at all morally blameworthy' to (7) 'extremely morally blameworthy' with the midpoint labelled 'substantially morally blameworthy'" (Schwitzgebel and Cushman, 2012, p. 139), the results for each scenario were much less homogeneous: When subjects were able to compare both scenarios, depending on order effects and familiarity with moral dilemmas of the tested population, the number of people who judged the Switch and the Push Type case equivalently ranged from 50% to 78% (Schwitzgebel and Cushman, table on p. 14). This is quite a high level compared to the 89% vs. 11% in Cushman et al. 2007 who said it was permissible to lead the train to a side track and kill someone (Switch Case) vs. shoving someone in front of a train and killing them (Push Case): Although Cushman et al. 2007 did not publish any within-subject comparisons for the Switch vs. Push Type Cases, the results give reason to

would not sacrifice anyone at all to save five but find it even more impermissible to sacrifice them when close contact harm is involved. In experiments with, for instance, 7-point scales, those people are more likely to give Push and Switch Type Dilemmas unequal permissibility ratings.

[108] Note that, unlike the studies I cite later in this text, these figures stem from an inter-subject design. Each subject, however, was presented with four dilemmas in randomized order (Cushman et al. 2007, p. 5). This means that the subjects did encounter several dilemmas and order effects were controlled for, as in the studies I am comparing it to. When I write that 11% found both the Push and the Switch Case actions permissible, I am assuming that both subject groups are homogeneous in the relevant features (which, considering the large sample, I will do here), and that the 11% who judged the Push Type action permissible would have been among the 89% of those who judged the Switch Type action permissible. I think that this is adequate because authors as Schwitzgebel and Cushman 2012 routinely exclude participants from their analysis who judge Push Type actions better than Switch Type actions, probably assuming that they did not properly understand the task, and the rates are around 5/5 to 8%; hence, judging that way is not regarded as the norm.

predict that most people find the action in the Switch Type case more permissible than the action in the Push Type case, hence judge according to the Close Contact Harm Principle. In the Schwitzgebel and Cushman experiment, however, 50-78% did not differentiate between the dilemmas and assigned them the same values on a 7-point scale. Those 50-78%, the majority of people who judged those dilemmas, apparently did not factor the Close Contact Harm Principle into their decision and, additionally, the population decided very inhomogeneously, with about 50% agreeing to one option (that both are equally permissible), leaving about 50% at the most to agree with a different option.

Hence, neither did an overwhelming majority of people judge according to the Close Contact Harm Principle (at least 50% did not judge as if it mattered to them), nor did the population agree on whether they found both dilemmas equally permissible or not (50-78% found that it does not matter whether someone gets killed to save five through close contact harm or without it, while the rest, 22-50%, disagreed).

The results might differ from Cushman et al. 2007's results because of the wording ("morally blameworthy" vs. "morally permissible") or because of the scale that offered more choices; whether order effects play a role is uncertain as Cushman et al. 2007 did not indicate unambiguously whether they observed an order effect. [109]

Whatever the reasons may be, Schwitzgebel and Cushman did not replicate Cushman et al. 2007's previous overwhelming majority that seemed to judge along the lines of the Close Contact Harm Principle. Although the former gave more answering options, people tended to assign the Switch and Push Type Cases the same values more often than the study with the binary choice would have predicted.

In the Ahlenius & Tännsjö 2012 study (Ahlenius and Tännsjö, 2012), in which the inhabitants of Russia and China also disagreed strongly about their judgements inner-culturally (63% (Russia) and 52% (China) thought they should flip the switch in the Switch Type case and 36% and 32% thought they should push the man in the Push Type case), they did not only disagree more strongly on how they rated single dilemmas (ratio of people who think they should flip the switch as opposed to people who think they should not flip the switch in the Switch Type case), but also on how much they tended to judge along the lines of the Close Contact Harm Principle (percentage of people who think they should flip the switch in the Switch Type Dilemma AND not flip the switch in the Push Type Dilemma or, if no within-subject data are available to see how the respective subjects judge both dilemmas, how many people in the population think they should act in a Switch Type case but not act in a Push Type case): Asked whether sacrificing one

[109] Cushman et al. 2007 did not publish whether and how presentation order influenced the judgements but stated rather ambiguously: "In order to eliminate the possibility of order effects, we restricted our analyses to first-trial responses, with comparisons made between-subjects." but later: "Permissibility judgments for these cases were widely shared, independently of the order in which they were presented in the session." (Cushman et al., 2007, p. 7); each subject was presented with four scenarios in random order.

to save five is permissible, the subjects from China judged with yes in at least 22% of both cases, and no in at least 36% of both cases, meaning that, according to our rationale, 58% gave equivalent judgements to both Switch and Push Type case. The subjects from Russia considered acting to be obligatory in at least 36% and non-obligatory in at least 20% of both cases, making their equivalency rating similar 56%. Both of these numbers would fall within the scope of Schwitzgebel and Cushman's 2012 results.

To sum up the review testing whether people universally judge according to the Close Contact Harm Principle: In Cushman et al.'s 2007 study, a great majority of people seemed to agree that they found action in the Switch Case permissible and action in the Push Case impermissible. This seemed to indicate that

- most people agree that it is permissible to sacrifice one in order to save five in Switch Type cases and
- most people agree that it is not permissible to sacrifice one in order to save five if this implies imposing close-up, personal contact harm on the person that is sacrificed in the Push Type Case.

As the difference between the Switch Type and the Push Type case is the kind of harm imposed, the results seem to show that most people judge in accordance with the Close Contact Harm Principle.

In the last section, I have shown that different studies indicate other distributions of judgement patterns in the tested populations, all of them less homogeneous than in the Cushman et al. 2007 experiment:

First of all, one study with Mayan subjects from a "rural" area did not find any preference for actions that did not involve close-up, personal contact harm over those that did (Abarbanell & Hauser 2010). Although the data are not strong enough to suggest any conclusions without follow-up experiments, the results contradict the assumption that the Close Contact Harm Principle is upheld universally.

Secondly, in many of the reviewed studies, the percentage of people who found it morally permissible (or a similar formulation) to sacrifice one in the Switch Type case was considerably smaller than in Cushman et al. 2007 (and differed cross-culturally), indicating that many people did not find it morally permissible to sacrifice one to save five even if no close-up, personal contact harm is involved (and the person is not sacrificed as a means but merely as a foreseen side-effect). And there was no unifying opinion in this matter, no overwhelming majority in most of the populations that agreed about judging the Switch Type case permissible or impermissible. But, more importantly for our consideration whether those principles are universal: If the percentage of people who agree to sacrificing one person in the Switch Type case is not much greater than the percentage of people who agree to sacrificing one in the Push Type case, or a majority of the subjects judges both types of cases equivalently, the principle might not play as big a role in those populations as in those with a bigger difference in ratings between the two dilemmas. This means that people do not judge according to the Close Contact Harm Principle very universally. Only parts of the population (and, in some cases, only minorities) do so. This probably shows that the principle is not universal; alternatively, the principle is universal but people do not judge according to it for different performance reasons: In Noam Chomsky's terms, they have the competence to judge according to the principle, and they would judge according to it if there were no hindering reasons that are not connected to the principle, e.g. that they did not understand the situation presented to them, that they could not read it, or that they wanted to judge 'impermissible' but they ticked the wrong box. I will come back to this at the end of this chapter.

I have shown that even among single cultures (or, in some papers, nations), people do not judge homogeneously in cases that consider the Close Contact Harm Principle.

But what about the famous Doctrine of Double-Effect that is so often cited as potential principle of Universal Moral Grammar? Is it universal?

5.6 Comparing judgements in Switch and Loop Type cases: Is the Doctrine of Double-Effect a universal principle?

Here, again, I will start with Cushman et al.'s 2007 experiment (Cushman et al., 2007, p. 6): They test two dilemma case types that are more similar and whose factors are therefore more controlled for than the usual Switch Type and Loop Type Dilemma pairing. I have described both the Loop and Loop with Heavy Object Type case above. Those dilemmas are designed to only differ in terms of the means/side-effect structure: In both cases, one person is on a loop track that loops back to the five people the train is headed towards; in one case, the weight of the person stops the train if a switch is thrown; in the other case, the person would not have been needed to stop the train as behind them is a heavy object. Hence, in one case, their death must have been intended to stop the train, in the other case, their death may be a foreseen side-effect. And, as already mentioned, 56% found throwing the switch permissible in the "intended" death case (Loop case) and 72% in the "foreseen" case (Loop with Heavy Object Case) (Cushman et al., 2007, p. 8). These results show a clear preference for actions that follow the Doctrine of Double-Effect code.

The only study directly comparable to this one is Cushman et al. 2006 (Cushman et al., 2006), who did not publish the percentages of equivalent ratings (they used a 7-point scale from 1, "forbidden" to 7, "obligatory" (Cushman et al., 2006, p. 1084)). They reported that they found the same, significant preference effect as Cushman et al. 2007 in all 6 different dilemma pairs of the same (Loop vs. Loop with Heavy Object) structure. This is all the more remarkable as their scenarios included different kinds of settings (boxcars, boats etc.) (Cushman et al., 2006, p. 1084). Hence, we can say that they replicated

Cushman et al. 2007's effect insofar as their subjects, too, rated using someone's dying as a means to saving five significantly worse than killing someone as a side-effect to saving five.

The studies above are the ones that show the 'purest' impact of the Doctrine of Double-Effect, as the used scenario pairs differ very marginally and one of their biggest differences is the 'intended/foreseen' death structure. Their results were stable in all six nations that Cushman et al. 2007 tested and in all the very different kinds of Loop/Loop with Heavy Object Type scenarios Cushman et al.'s 2006 study tested: People tended to judge the 'intended' death case as less permissible than the 'foreseen' death case. The tested subjects and their nationalities (5.2, p. 111) were not representative enough to show that a principle is 'universal', as I have let Abarbanell and Hauser argue above, because they were mostly "from English-speaking countries, most were formally educated, and many had read books discussing moral issues" and "all were technologically savvy [...]" (Abarbanell and Hauser, 2010, p. 208) in both studies (although Abarbanell and Hauser were referring only to the Cushman et al. 2007 study). But even if we take both studies to be representative for a diverse subject mix, they show a trend towards judging along the lines of the Doctrine of Double-Effect but, at least in Cushman et al. 2007, it is far from universal: Even though only 56% found it morally permissible to sacrifice someone as a means, they still constitute the majority of the subjects. And even though as many as 72% find it permissible to sacrifice someone under the same circumstances, but as a foreseen side-effect, on average only 16% seemed to care about the distinction (and in a within-subject design, when one subject was presented with both scenarios in one single session, only 5.8% judged the Loop and the Loop with Heavy Object scenario differently (Cushman et al., 2007, p. 14): The principle does not even seem to be universal (or at least, people do not judge according to it universally) amongst one otherwise quite homogeneous group of subjects. This is not even a case of inner-cultural disagreement: A plain majority does not seem to judge according to the Doctrine of Double-Effect.

All other experiments I have reviewed in this chapter compare Switch Type Dilemmas (classical Side-Track Dilemmas) and Loop Type Dilemmas. The difference between those cases is less subtle, as we can see in Cushman et al. 2007: 89% found it permissible to kill someone who is standing on a side-track that does not lead back to the main track with the five people by redirecting a train towards that track (Cushman et al., 2007, p. 7). That is a significant difference to the 56% who found it morally permissible to sacrifice one person in the Loop Case. I will stay within a similar population with the next study to be reviewed, Schwitzgebel and Cushman 2012 (Schwitzgebel and Cushman, 2012). They tested whether academics trained in moral philosophy judged Loop and Switch Cases differently from other subjects. As mentioned before, Eric Schwitzgebel was kind enough to send me some unpublished results of their study (E. Schwitzgebel, personal communication, January 17, 2015) and it turns out that about 83% of all philosophers and

80% of all non-philosophers rated Loop and Switch Type Cases equivalently, while only 10% of all philosophers and 12% of all non-philosophers rated throwing the switch in Loop Type cases as worse than in Switch Type cases. As subjects were tested on both dilemmas in one trial (hence, it was a within-subject design), this tendency was to be expected in the light of Cushman et al. 2007's findings where 5.8% of subjects rated Loop and Loop with Heavy Object cases differently in one session (and 33% of all who either judged the Loop Case impermissible or the Loop with Heavy Object case permissible when judging the respective other scenario on average 20 weeks later) (Cushman 2007, p. 14/15). The data between both studies, however, differ in two important aspects: Cushman et al.'s subjects only had a binary choice (the action was either permissible or impermissible), whereas Schwitzgebel and Cushman's subjects could choose on a 7-point-scale from "not at all morally blameworthy" (1) to "extremely morally blameworthy" (7) (Schwitzgebel and Cushman, 2012). And Cushman et al. 2007 compared Loop with Heavy Object Dilemmas to Loop Dilemmas, whereas Schwitzgebel and Cushman 2012 compared Switch Dilemmas to Loop Dilemmas. This makes for a significant difference because subjects usually tend, on average, to more seldom judge Switch Dilemmas equally to Loop Dilemmas than they judge Loop with Heavy Object Dilemmas equally to Loop Dilemmas. The latter might explain the tendency to more disparate ratings in Schwitzgebel and Cushman 2012 while the former might explain why the this tendency is not stronger.

Interestingly, 7% of philosophers and 8% of non-philosophers rated the Switch Case worse than the Loop Case, hence found killing someone as a means more permissible than killing someone as a side-effect (E. Schwitzgebel, personal communication, January 17, 2015). This is not so different from the 10% philosophers and 12% non-philosophers who judged along the lines of the Doctrine of Double-Effect. Formulated differently: Almost as many subjects judged against the Doctrine of Double-Effect as did people in accordance with the Doctrine of Double-Effect, rendering hypotheses questionable that human moralities against our three principles will not develop naturally (Harman, 2011, p. 18). Beyond that, far from homogeneously judging according to the Doctrine of Double-Effect, only 10-12% of all subjects judged according to the doctrine.

Schwitzgebel and Cushman pursued another experiment with Drop Case scenarios. Drop Cases involve manipulating the victim by letting them drop onto the rails through a trap door (Schwitzgebel and Cushman 2015, p. 130). In this scenario, the harm is not brought about in a close-up manner and does not involve physical contact with the sacrificed person, but the person is dropped in front of a train to stop it and hence used as a means. Many of the subjects were philosophically trained graduates; 98% of the subjects were from the US (Schwitzgebel and Cushman 2015, p. 129). 46-70% of all subjects judged the two dilemma cases equivalently depending on the order of presentation (Schwitzgebel and Cushman 2015, p. 131); half of them were instructed to make considered judgements: The instructions said that "we are particularly interested in your reflective, considered responses"(Schwitzgebel and Cushman 2015, p. 129). One reason for the generally lower equivalency ratings compared to the Loop Type Cases may be that in Drop Type Cases, the agent has to let the victim drop onto the rails through a trap door to save the five. This may be "constituting additional intentional harm" (Greene et al. 2009, p. 187) as compared to just making a train hit the victim. [110] Another cause might have been the "consideredness" of half of the responses. "Philosopher respondent" (as compared to "non-philosopher respondent") was not amongst the listed predictors for equivalency ratings (Schwitzgebel and Cushman 2015, p. 132).

Speculations about the cause aside, the study shows that here a smaller percentage of subjects rated these two dilemmas equivalently than in (regarding most aspects) comparable studies that test for the Doctrine of Double-Effect and that I have reviewed. This means that in this study, more judgements were in line with the Doctrine of Double-Effect than in the previously cited studies that involve Loop rather than Drop scenarios. Schwitzgebel et al. used the same 7-point scale as in their 2012 experiment: From 1 (Extremely Morally Good) to 7 (Extremely Morally Bad) with a midpoint labelled "Neither Good Nor Bad" (Schwitzgebel and Cushman 2015, p. 130; Schwitzgebel 2012, p. 138). This makes the two studies quite comparable and the results of the more recent study are more favorable toward the universality thesis. But even in the latter study, preferring the Switch scenarios was far from universal amongst the participants. In a nutshell:

Only 2% rated Drop or Push Cases better than Switch Cases (and were excluded from the following analysis). But whereas the rest either rated in accordance with the Doctrine of Double-Effect or found both scenarios equally permissible, with a significantly differing average rating of 3.7 for Switch and 4.5 for Drop Cases, only very roughly half of the population in both the considerate and the average judgements judged in accordance with the Doctrine of Double-Effect (Schwitzgebel and Cushman 2015, p. 131). This means that half of the sampled subjects judged in accordance with the Doctrine of Double-Effect and half of the sampled subjects ignored it; with only two possibilities to judge (according with and ignoring it), there could be no bigger disagreement in a population; the (implicit) endorsement of the Doctrine of Double-Effect is, as we can see, not universal in this study either. On average (with half the subjects presented with the Switch Case first and half of them presented with the Drop Case first), fewer subjects judged in accordance with the Doctrine than stayed unaffected by it and, on average, they only judged the Switch Case about 1 point on a 7-point scale towards "extremely morally good" in comparison to the Drop Case.

So much for inner-cultural universality of the principle in the most frequently tested

[110] Greene et al. 2009 write about trap door dilemmas: "We speculate that [the moral acceptability ratings being relatively low, almost as low as a Push Type Dilemma they tested for] may be due to the fact that the actions in these dilemmas involve dropping the victim onto the tracks, constituting an additional intentional harm (Mikhail, 2007)." (Greene et al. 2009, p. 187 in "A Companion to Experimental Philosophy"), citing (Mikhail 2007).

countries; what is the situation inter-culturally?

In their study about Mayan morality, part of which I have reviewed in the previous sub-chapter, Abarbanell and Hauser tested how Mayan subjects who grew up without much contact to city life ("rural group") would judge Switch Type in comparison to Loop/Drop Type Dilemmas. They used a 5-point scale that ranged from the Ts'eltal translation of 'very impermissible' (1) to the translation of 'very good' (5) (Abarbanell and Hauser, 2010, p. 211) and found that the participants gave the Switch Type Dilemma a rating of 4.07 and the Loop Type Dilemma a rating of 1.90 (Abarbanell and Hauser, 2010, p. 211).

Now this is a rather strong (and significant) difference in permissibility ratings; but in a later study with different scenario settings, subjects from a similar population group assigned an unintentional killing action where the five are saved as a side-effect the score of 3.44 on average and an intentional killing action where the five are saved as a means the score of 2.67 (values that, however, still significantly differ) [111] (Abarbanell and Hauser, 2010, p. 213). The fact that these values are closer to each other than the ones with the initial scenario settings may be due to the very unconventional settings of the first dilemma pair:

Both initial dilemmas start with the same text: "A man is sitting near the side of the road when he sees a truck speeding along. It is headed towards a group of five men, who do not hear or see it, and if nothing appears in the road, it will certainly hit and kill them."

The Switch Type Dilemma continues as follows: "Across the road is another man sitting in front of his house. If the man who is sitting by the road calls out to the truck driver and says 'watch out,' the truck driver will swerve away from the five men, saving them, but into the path of the one man in front of his house, killing him. If the man sitting by the road says nothing the truck will travel on and kill the five, and the one man in front of his house will be safe. The man decides to call out, so the one man is killed and the five men are saved."

And the Drop Type Dilemma continues: "Across the road is another man sitting in front of his house. If the man who is sitting by the road calls out to the man by his house and says 'come here,' the man will walk into the road in the path of the truck, be killed, and stop it from continuing on toward the five, saving them. If the man sitting by the road says nothing, the truck will travel on and kill the five. The man decides to call out so the one man is killed and the five men are saved." (both Abarbanell and Hauser, 2010, p. 212).

[111] Also note that, if our previous assumptions hold and subjects would tend to rate Drop Type Cases worse than Loop Type Cases because of the additional harm caused by the drop, we would expect the ratings for the Drop Type scenario to be closer to 1 as compared to the Loop Case tested before rather than higher, as is the case, especially considering the rather drastic setting with a baby being dropped in front of an elephant herd.

The scenarios, hence, not only differ in terms of their 'means/side-effect' structure; by shouting 'Come here' in the Loop Type Case, the actor might be considered to be abusing the soon-to-be-victim's trust (or compliance) and has a direct verbal interaction with him in order to sacrifice him. Hence, other factors besides the 'means/side-effect' distinction come into play that have not been controlled for.

The second dilemma pair that people assigned closer values to respectively involves throwing vs. dropping a baby from a tree to stop a stampede of peccaries that would otherwise kill five babies. Having the baby fall as a means vs. as a side-effect did not make as big a difference to the subjects in this case with more similar respective settings (Abarbanell and Hauser, 2010, p. 213). Although this dilemma pair, too, tests for the impact of killing as a side-effect vs. killing as a means of saving five, it is for several reasons not comparable to any of our previous scenario pairs: The person performs a close-up, personal killing in both cases (in our other pairs, the killing was neither close-up nor personal nor did it involve a drop in the side-effect scenario) and, unlike most other dilemma pairs, involves killing babies.

Notably, a population with more contact to cities rated the first (unconventional) dilemma pair with 2.83 (Drop Type Case) and 2.3 (Switch Type case) (Abarbanell and Hauser, 2010, p. 216). On average, they judged the side-effect harm significantly better than the means harm, but the ratings were much closer to each other than the "rural" comparison group's ratings.

Even with the inconsistencies I have mentioned above and although Abarbanell and Hauser have not published the percentage of equal ratings, meaning that we cannot say much about the homogeneity inside the subject group or about ratings that favored harms as a means over harms as a side-effect, the results suggest that most of the Mayans seemed to consistently make the distinction along the lines of the Doctrine of Double-Effect with a tendency of the group that had more contact to cities to draw a smaller distinction in favor of the Doctrine.

The last study that tested for the universality of the Doctrine of Double-Effect (and that I have partially reviewed above, too (namely, Ahlenius and Tännsjö 2012), see beginning of Chapter 5.4, p. 122) was conducted with people from the USA, Russia and China. The subjects from the two former countries were interviewed via phone while the interviewers sat opposite the subjects from China. The participants were asked whether they should flip the switch to kill one and save five in a Switch and a Loop Dilemma respectively and could answer with 'Yes', 'No' and 'I don't know'. 81% of all subjects from the US thought they should flip the switch and lead the trolley on a side-track with one person (Switch Case), while only 60% thought they should flip the switch and lead the trolley on a loop track where the person on the track would stop the train (Loop Case). 63% of subjects from Russia judged that they should flip the switch in the Switch Case and 54% in the Loop Case while 52% from China said 'Yes' to the Switch Case and 34% to the Loop Case

(Ahlenius and Tännsjö, 2012, p. 4). Again, the experiment is not necessarily comparable to other (written form) experiments as subjects from China could see (and be seen by) their interviewers and the others were on the phone (which makes a comparison between groups more difficult as well) and because participants were asked what they would do and not how permissible they found an action, except for the Chinese who were asked "Would it be morally permissible to [flip the switch and have one person killed in order to save five]?" (Ahlenius and Tännsjö, 2012, p. 4). However, we can see here that the inner-cultural agreement is high about the Switch Case in the US, but only 21% of the subjects' judgements on average could be attributed to the difference between Switch and Loop, hence, possibly to the Doctrine of Double-Effect (the percentage of people who judge the two respective dilemmas differently): The discrepancy between the ratings for the Switch Type and the Loop Type action amounts to 21% of US participants (81% vs. 60% found the Switch Type vs. the Loop Type action obligatory); 18% for all subjects from China (52% vs. 34%) and 9% for the subjects from Russia (63% vs. 54%) (Ahlenius and Tännsjö, 2012, p. 4). Although all three populations showed a tendency towards finding harms done as a side-effect more permissible than harms done as a means, it is hard to attest universality when we consider that, on average, only a small percentage of the judgments seems to be influenced by the means vs. side-effect structure of the scenarios.

To sum up, in all reviewed experiments that cited percentages, a rather small percentage of the population judged along the lines of the Doctrine of Double-Effect at all (judging actions that involve harm as a side-effect as more permissible than actions that involve harm as a means). In Schwitzgebel and Cushman's 2012 experiment,(Schwitzgebel and Cushman, 2012) a significant amount of subjects even judged in the opposite direction, judging the dilemmas that involved harms as a means better than those that involved harms as a side-effect. Although the trend is to more frequently judge according to the Doctrine than against it, and this trend is universal, the judgements themselves and thereby the implicit endorsement of the Doctrine are far from universal even amongst rather homogeneous populations.

5.7 How universal is the Action/Omission Principle?

Last but not least, I wish to analyze how universal the Action/Omission Principle is. Many articles explore the Action/Omission bias, but only few use Trolley Dilemma Type Cases to do so. The majority of those articles, many of them concerned with vaccination, shows that subjects tend to evaluate harm caused by actions as less permissible than harm caused by omissions with few exceptions such as cases when an action would have been expected by a person, e.g. because they were in a responsible position. [112] Cushman

[112] In a case with a train where injuries would occur when it wasn't stopped but also when it was stopped, people judged the appropriate compensation higher when the engineer did not stop the train than when

et al. 2006 (Cushman et al., 2006) do indeed show that subjects significantly judge in accordance with the Action/Omission Principle, hence in favor of harm by omission, in a plethora of Trolley Type Cases but do not provide any data on whether people with varying backgrounds use the Action/Omission Principle to the same degree. Neither do they mention the overall percentage to which subjects agree more with an active harming case or agree more with a case of harm by omission, which makes it difficult to assess how homogeneously people judged.

Instead of testing the effects of nationality, Hauser and Banerjee tested the influence of religion on Trolley Dilemma judgements. Unfortunately, they did not publish the results for single types of principles they had tested for, but conflated them to "3p" ("three principles") or represented them with labels as "Trolley-Scenario 6", [113] so that I could not discover which data belonged to action/omission contrast dilemma pairs. They found significant effects of gender and religiosity for the "3p" dilemmas but classified them as very small and attested that the pattern of the effects was inconsistent (Banerjee et al., 2010, p. 273).

Banerjee et al. conclude that "[...] our results suggest that human minds rely on a default set of moral principles that are robustly present across a wide array of demographic and cultural differences; and this holds true for a number of types of moral scenarios, including violations of conventions, moral transgressions and moral dilemmas." (Banerjee et al., 2010, p. 273)

An experiment with Mayans living in a rural area by Abarbanell and Hauser (Abarbanell and Hauser, 2010), however, reveals a different picture: "Rurally living" Mayans consistently did not judge Trolley Type scenarios differently when there was an action involved than when there was an omission involved. Abarbarnell and Hauser tested five different action-omission dilemma pairs using a five-point-scale ranging from the Ts'eltal word for "very impermissible" to the Ts'eltal word for "very good". On average, the rural Mayan participants gave the same values to the omission scenarios as to the action scenarios in all five cases. More educated, younger Mayan participants who spent more time in cities did not show this effect.

Several explanations may account for this seeming absence of the Action/Omission Principle in this rural Mayan population sample: Previous studies have shown that subjects perceive the moral difference between actions that cause harm to another person and omissions that cause harm to another person as smaller when the acting/omitting person and the harmed person were "significantly related to each other, either in terms of 'hierarchy' or 'solidarity'." ((Fraser and Hauser, 2010, p. 24), referring to (Haidt & Baron

he tried to stop the train and failed and or when he "actually stopped (with the sudden stop causing identical injuries)."((Baron and Ritov, 2004, p. 76), citing (Ritov & Baron 1994))

[113] The texts belonging to those labels were originally disclosed on the Moral Sense website when the paper was published but were not available under the online address provided in the paper at this point (28.4.15).

1996)).

In the small-scale community which the subjects were part of, Fraser and Hauser propose that what made the crucial difference might not have been the degree to which those subjects endorsed the Action/Omission Principle, but who they imagined would be the protagonists in those dilemmas and how they were related to each other: If they imagined that the person who made the sacrifice and the person who was sacrificed were in a relatively close personal relationship (which, Fraser and Hauser suggest, applies to most people in the tested small-scale community), the difference between actions and omissions would vanish. In Haidt & Baron's work, however, the omission bias did not completely vanish; people only tended to perceive a smaller moral difference between actions and omissions that harm people closely related to the actor/omitter. I endorse Fraser and Hauser's suggestion that further empirical work is needed to identify the causes for these unexpected results but I tend to view them as, possibly preliminary, counter-evidence to the universality thesis.

5.8 "Universal Moral Grammar" or just "Moral Grammar?"

In this chapter, I have shown that none of the three principles is as universal as proponents of the Linguistic Analogy claim it is. Most of them are not cross-culturally stable and, more strikingly, even within comparably homogeneous population samples, at least in some experiments, subjects showed large disagreements; when the choice was dichotomous (such as permissible/impermissible), in many cases the distribution was close to 50% vs. 50% (50% of the population thought an action was permissible while 50% thought it was impermissible). We can examine some arguments in favor of the proponents of the Linguistic Analogy, such as the following: Natural moralities may be constrained. Maybe it is not necessary for everyone to judge in accordance with the same principles to make moral grammar universal, but its universal part may consist of universal constraints to certain types of moralities.

But when proponents of the Linguistic Analogy write about constraints, they usually argue that most people employ the same principles (see, for instance, Dwyer et al., 2009, p. 502), (Harman 2011, p. 18) and (Dwyer, 2006, p. 249) and that therefore a natural morality will not emerge that goes against those principles, for instance where actively harming someone is regarded as better than harming someone by omission or where harming someone as a means is regarded as more permissible than harming them as a foreseen side-effect. We can find some counter-evidence in the form of studies that feature experiments where most people judge harm by actions as more permissible than harm by omissions in certain situations (for an extensive and critical review, see (Baron and Ritov, 2004, p. 76 ff.)). Beyond those studies, I have shown in this chapter that the three principles are not upheld as homogeneously in all moralities as it may seem at the first glance at Cushman et al.'s 2007 data (Cushman et al., 2007). Admittedly, however, the percentage of people who judge that harm as a means is more permissible than harm as a side-effect is so small that Schwitzgebel et al. excluded the subjects that did so (5% of double effect cases (and 8% of action-omission cases) (Schwitzgebel and Cushman, 2012, p. 140). Maybe, however, those were not people who did not understand the dilemmas or did not seriously answer them and deserved to be excluded, but instead people who represented the part of a society that did not judge in accordance with those principles: When subjects only compared Loop and Switch Cases, 8% of non-ethics PhDs and 7% of philosophers judged harm as a means more permissible than harm as a side-effect, whereas only relatively few 11% of non-ethics PhDs and 10% of philosophers judged in the opposite direction that conforms with the Doctrine of Double-Effect [114] (Schwitzgebel and Cushman, 2012; E. Schwitzgebel, personal communication, January 17, 2015). We should shift the burden of evidence when almost as many people judge contrary to the principle we take to be universal, or at least constraint-setting, as judge in accordance with it. The ratio of exception by rule should not be 7/10.

Or we might not even need universality to make claims about innate features of morality. The principles-and-parameters account holds that we might have an innate set of principles which, when triggered by our environment, sets certain parameters differently. This leads to moralities that have common features but also display different features in different societies due to different upbringings of the 'carriers'. Here, the proposed parameter is often the group of people to whom one applies a principle (such as "Do not harm people of your own group" or "Do not kill people unless they break certain laws") (Harman, 2011, p. 20), (Dwyer, 2006, p. 249). See also: (Roedder and Harman 2010)

But, again, our three principles are mostly considered to be the universal principles and not the culturally adjustable parameters:

> "We do not expect universality across the board. Rather, we expect something much more like linguistic variation: systematic differences between cultures, based on parametric settings. Thus [...] we expect differences between cultures with respect to how they set the parameters associated with principles for harming and helping others. [...] I believe that some aspects of the computation are universal, part of our moral instinct. I doubt very much that people will differ in their ability to extract the causes and consequences of an action. Everyone will perceive, unconsciously, the importance between intended and foreseen consequences, intended and accidental actions, actions and omissions, and introducing a threat as opposed to redirecting one. The central issue in thinking about cross-cultural variation is to figure out how different societies build from these universal factors to generate differences in moral judgments."

[114] Subjects more proficient in ethics, however (ethics PhDs) completely avoided this trend with only 1% judging actions that harm as a means better than actions that harm as a side-effect.

(Hauser, 2006, p. 129/130)

But: If the principles are universal and the parameters depend on your environment while you grow up, why is there so much inner-cultural disagreement on the seemingly universal principles? This would mean that the percentage of people who do not judge according to the principles had different influences in their lives than the people who judged according to them. It is, of course, an empirical question as to what those influences are. They seem to be mostly detached from gender, exposure to moral philosophy, education and other factors we have been reviewing here (see Poverty of Stimulus chapter). Hauser et al. write: "[...][T]his view does not deny cultural variation. Rather, it predicts variation based on how each culture switches on or off particular parameters. An individual's moral grammar enables him to unconsciously generate a limitless range of moral judgments within the native culture." (Hauser et al., 2008, p. 266). But as long as we have not found out what makes parts, and in some cases half of, a population not judge according to the principles, or, speaking to Hauser et al., what constitutes a moral culture, the inner-cultural variation in the experiments I reviewed here does not speak for a universality that varies with cultural parameters but always falls within the scope of those principles. Instead, it seems to constitute counter-evidence for the universality of those principles.

A third reason why our three principles could be universally present and innate, although the behavioral data do not reflect their universal presence, could be the difference between competence and performance: The principles might be present, but not expressed.

This distinction, like many others in the Linguistic Analogy, goes back to Chomsky. Competence, according to Chomsky, is the mental (but not necessarily conscious) presence of those rules that can generate a language (Chomsky and Hornstein, 1980, p. 201). Competence is, as Sag and Wasow state, "idealized linguistic knowledge" (Sag and Wasow, 2011, p. 316). Someone who is grammatically competent in a language can, in principle, understand and produce sentences in that language as long as they do not lack performance factors, thus rendering it impossible for them. Performance, on the other hand, are factors that enable people to produce and understand utterances independently of this system of rules (competence), such as a functioning memory and, in the case of spoken languages, the motor and auditory system. If someone is not able to memorize the beginning of a sentence when they perceive the end of the sentence, it will be hard for them to understand the whole sentence. So even if they were, in principle, able to recognize the rules and derive the grammatical structure (and even meaning) of the sentence, they would fail to do so because of their memory constraints. This means, for research on linguistic competence, that you always have to abstract from performance factors if you want to infer something about the linguistic competence system per se from your behavioral data. Transferred to moral grammar, it means that people might have the proper three principles in their minds but lack performance factors. As Erica Roedder and Gilbert Harman 2010 have argued, it is not quite clear in the moral domain what is part of the moral competence and what is not (consider emotions: are they part of the moral grammar that can generate new judgements, or are they distorting factors? It depends on our theory of moral processing) (Roedder and Harman, 2010).

If we transfer this linguistic theory to our moral case, we could try to interpret the divergent moral judgements in light of the competence-performance distinction: Although in principle everyone would judge according to our three principles, a number of people is influenced/distracted/lacks some resources to arrive at the 'correct' judgement. Asking what might be the reason that keeps their 'pure' moral grammar from generating the correct judgements is, of course, once again an empirical question, albeit very difficult to answer if we have not individuated the mechanism that is the core of moral grammar. The latter, however, might only make sense if we assume that there is something 'genuinely moral-grammatical'. Additionally, the large proportion of people who do not judge in accordance with the three principles would mean that people making 'faulty' moral judgements (not executing their moral grammar correctly) would be the rule rather than the exception.

I conclude that although proponents of the Linguistic Analogy take the universality of moral judgement patterns as evidence that the psychological system that generates them has developed evolutionarily, they rest their paradigm on the wrong assumption: Experiments show that often subjects neither judge homogeneously inner-culturally, nor do they show the same judgement patterns cross-culturally. Although I have named several reasons why the moral grammar mechanism could be universal although people do not judge universally, I do not believe that the data warrant an interpretation in favor of a congenital, universally present moral grammar at this point and they are especially unfit to be used as main reason for believing believe in the latter.

But maybe the Poverty of Stimulus arguments are more apt to give us good reasons for an innate Trolley Dilemma judgement mechanism.

6 Poverty of Stimulus

Poverty of Stimulus gives us reasons that make it more probable for a mechanism to be adapted. If Poverty of Stimulus is given for a behavior, it is not explainable purely by learning. Applied to our snake-fearing mechanism: If people can learn to be afraid of snakes by imitating someone who is scared of snakes, why do they not learn to be scared of flowers just as quickly by imitating people who are scared of flowers?

Because they are not explainable solely by the environmental input of an individual, features like that are commonly taken as evidence that some information is ‘built in’ in the individual (for a review on an early argument of this kind see also: (Lehrman, 1953)). Behavior that we cannot account for by pure experience must be based on a mechanism that has somehow ‘been there’ previously.

One particular instance of this kind of argument, and the most widely used one, is the Poverty Of Stimulus argument. It originally stems from Noam Chomsky, who states the following: "When we turn to language, many examples have been studied of shared knowledge [115] that appears to have no shaping stimulation - knowledge without grounds, from another point of view - and that seems to be based on principles with only the most superficial resemblance to those operative in other cognitive domains." (Chomsky, 1980, p. 41/42). "In this case too, it can hardly be maintained that children learning English receive specific instruction about these matters or even that they are provided with *relevant experience* that informs them that they should not make the obvious inductive generalization [...]." ((Chomsky, 1980, p. 43); in both quotes: emphasis added by me).

Noam Chomsky was referring to language when he introduced this concept; in the tradition of the Linguistic Analogy, we will transfer this concept to the realm of Trolley Dilemma Cases. When Noam Chomsky speaks of "shared knowledge", he is talking about a pattern of behavior (in his case, linguistic behavior) that all or a great percentage of people show consistently when they are put in a certain situation. I have discussed this in more detail in the part about universality (Chapter 5, p. 105.

The main point here is that he argues that some part of linguistic behavior he observed is based on "knowledge without grounds" and that people show some kind of performance that they have not had the "relevant experience" to show.

As I will argue against the claim that Trolley Dilemma judgement mechanisms are adapted functions, I will make his (counter-)argument as strong as possible and refrain from claiming that everyone or almost everyone must show a behavior or its prerequisites ("shared knowledge", as Chomsky calls it) [116] for it to be a case of POS. I will start

[115] Chomsky has later relativized his concept of "knowledge" and created a neologism, "cognizing"; this concept, however, is not important for my argumentation at this time and I will therefore not discuss it here.

[116] I will show that the "shared knowledge" that seems to underlie Trolley Dilemma judgements does not seem to be common to everyone in a population in the "Universality" part of the argumentation,

with some clarifications regarding concepts that are important in the context of Poverty of Stimulus arguments, elaborate on one concern that questions the whole Poverty of Stimulus concept and, after this concern is dispelled, progress with a portrayal of the most relevant Poverty of Stimulus argument, the Logical Problem of Language Acquisition, and some strings of arguments related to it. I will apply those to Trolley Dilemma Cases by examining the effect of explicit instructions. To do this, I will resort to empirical studies, transfer conceptual arguments that stem from the realm of linguistics to moral principles and develop some own conceptual arguments.

I have not yet provided a clear formulation of the Poverty of Stimulus argument.

6.1 Different facets of the Poverty of Stimulus argument

In the linguistic literature, different sub-arguments regarding the Poverty of Stimulus argument are cited frequently. The authors adduce different reasons why the linguistic input is not sufficient to lead to the kind of linguistic (behavioral) output we can observe in almost all adults. Pullum and Scholz (Pullum and Scholz, 2002) have discerned 13 arguments of this kind after analyzing literature on the Poverty of Stimulus. [117] I will, however, only discuss the arguments that people have used to postulate that Trolley Type Dilemma judgement mechanisms have not been learned.

6.1.1 Definitions of "Learned"

It is difficult to find a good definition of "not learned". The next two citations, along with a (rather vague) version of the Poverty of Stimulus argument as applied to the moral [118]

depending on which part of "knowledge" is defined to be the "shared" part. If I posit the preconditions for something to fall under the category of Poverty of Stimulus as a conjunction including universality, the claim that Trolley Dilemma judgement mechanisms fall under Poverty of Stimulus would have already been refuted together with their universality. I believe that the fact that something cannot be learned is a very strong argument for the underlying information to be more or less indirectly provided by the genetic endowment.

[117] "(1) Properties of the child's accomplishment a. SPEED: Children learn language so fast. b. RELIABILITY: Children always succeed at language learning. c. PRODUCTIVITY: Children acquire an ability to produce or understand any of an essentially unbounded number of sentences. d. SELECTIVITY: Children pick their grammar from among an enormous number of seductive but incorrect alternatives. e. UNDER-DETERMINATION: Children arrive at theories (grammars) that are highly under-determined by the data. f. CONVERGENCE: Children end up with systems that are so similar to those of others in the same speech community. g. UNIVERSALITY: Children acquire systems that display unexplained universal similarities that link all human languages. (2) Properties of the child's environment a. INGRATITUDE: Children are not specifically or directly rewarded for their advances in language learning. b. FINITENESS: Children's data-exposure histories are purely finite. c. IDIOSYNCRASY: Children's data-exposure histories are highly diverse. d. INCOMPLETENESS: Children's data-exposure histories are incomplete (there are many sentences they never hear). e. POSITIVITY: Children's data-exposure histories are solely positive (they are not given negative data, i.e. details of what is ungrammatical). f. DEGENERACY: Children's data-exposure histories include numerous errors (slips of the tongue, false starts, etc.)." (Pullum and Scholz, 2002, p. 12/13)

[118] I will use 'moral' in the following text although I am aware that it is arguable whether all Trolley Type Dilemma situations are truly moral and whether all Trolley Type Dilemma judgements are truly

domain, contain definitions for ”non-learnedness”: ”Since the emergence of this knowledge [some features of moral competence] cannot be explained by appeals to explicit instruction, or to any known process of imitation, internalization, socialization, and the like, there are grounds for concluding it may be innate (Dwyer, 1999; Mikhail, 2000).” (Mikhail, 2008, p. 354), citing (Dwyer, 1999) and (Mikhail, 2000) As you can see, this version contains the words ”and the like”, which makes it a piece of guesswork to find out what else exactly can be ways to learn something. The following version contains an even vaguer account for non-learnedness:

> ”If the relevant principles can be shown to emerge and become operative in the course of normal development, but to be neither explicitly taught nor derivable in any obvious way from the data of experience, then there would appear to be at least some evidence supporting an argument from the poverty of the stimulus in the moral domain.[...]” (Mikhail, 2011, p. 82)

But what does “not derivable in any obvious way” from experiential data actually mean? I will not present an exhaustive definition of non-learnedness here. For an account of non-learnedness that would be as exact as possible at this point in time, one would need to examine a compilation of all processes of learning that have been observed/researched to date or at least of those that have been observed frequently or researched successfully. Those would have to be excluded in order to see whether something has not been learned in general (according to our knowledge).

I will instead formulate exact versions of the arguments that proponents of the Linguistic Analogy have stated (or, taking the often-fuzzy formulations of those arguments into account, could have had in mind when they stated their versions) in connection to Poverty of Stimulus and see what kind of experiential data would have been necessary to invalidate those single arguments. As you will see, this will significantly reduce the scope of ”non-learnedness” in the respective cases.

6.1.2 Adaptedness, Evolutionary Adaptedness and Inheritedness

One thing should be noted about the structure of Poverty of Stimulus arguments before I start expounding those arguments: Poverty of Stimulus arguments put their focus on developmental issues rather than on the match between ancestral problems and ”cognitive design” as do classical Evolutionary Psychologist arguments. The former are used to argue for a trait to be genetically determined/implemented, the latter for something to be an adaptation. Both properties are not mutually exclusive in the same trait. Although they are compatible, they are not co-extensive. The following section aims to show that if proponents of Universal Moral Grammar argue for the non-learnedness of moral principles, they want to show that they are somehow genetically determined or, at least, some kind of

moral. Most texts that I will cite use this word and I adopt it for the length of this chapter.

information is ‘built in’ that people require to produce moral judgements as they do. I will show that genetic determinedness is necessary, but not sufficient for the moral principles to be adaptations in the Darwinian or, more narrowly, in the Evolutionary Psychologist sense.

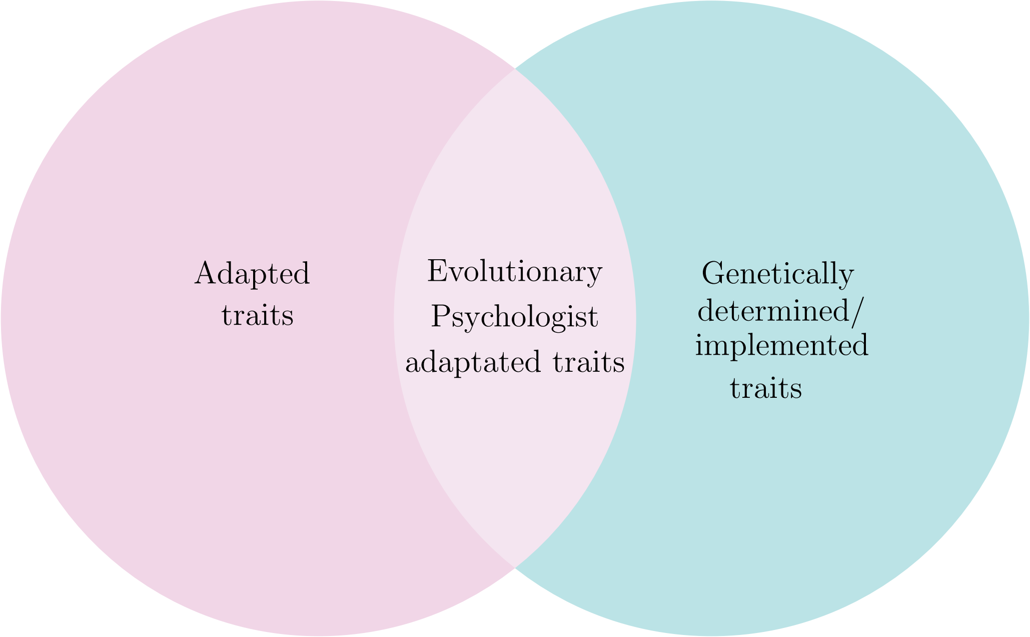

Adapted traits
Evolutionary Psychologist adaptated traits
Genetically determined/ implemented traits

Figure 8: Adaptedness, Evolutionary Adaptedness and Inheritedness

Adaptive traits in the Evolutionary Psychologist sense are a subclass of all adaptive traits. They are those adaptive traits that are genetically determined/implemented. All adaptive traits in the Evolutionary Psychologist sense are genetically determined/implemented. Not all genetically determined/implemented traits, however, are adaptations in the Evolutionary Psychologist sense (or adaptations at all, for that matter).

In the next section, I will argue that genetically determined traits and Darwinian/Evolutionary Psychologically adapted traits (one sub-group of Darwinian adaptations) are not co-extensive but that if a trait is genetically based, this provides evidence for it to be adapted in the Evolutionary Psychologist sense. The argument has the following form:

1. Being genetically based and being an adaptation are two different things.

2. Being genetically based and being a Darwinian/Evolutionary Psychologist adaptation are two different things.

3. Every Evolutionary Psychologist adaptation is genetically based.

4. If you can show that something is genetically based, you can show that one precondition for it to be an Evolutionary Psychologist adaptation is fulfilled.

5. If one precondition for something to be an Evolutionary Psychologist adaptation is fulfilled, it becomes more plausible that something is an Evolutionary Psychologist

adaptation. This is why people who believe that there is a ‘moral faculty’ (an adapted organ that is specialized in making moral judgements) use Poverty of Stimulus arguments.

6. If you can show that something is not genetically based, you can show that it is not an Evolutionary Psychologist adaptation.

I have remained vague as to what exactly those traits are, what “genetically based” means and what “non-learned” means as this is not relevant to this argument, and I will return to “non-learnedness” and apply the Poverty of Stimulus arguments to concrete examples in Chapter 6.2.3, p. 160

1. As Richard Joyce has noted (Joyce, 2013), the development of a trait that is based on genes and the development of a trait that is an adaptation are not the same thing. One example Joyce cites for this is trisomy 21. Some of the traits that have been associated with this genetic particularity develop relatively reliably under a variety of environmental conditions. Trisomy 21 is improbable to be an adaptation as in most environments, it does not increase chances of survival, survival of kin or reproduction rate. Hence, it probably is a case where traits are genetically based and develop relatively independent of the environment, [119] but are not adaptations. Joyce also gives an example for adaptations that are not genetically based: A particular stone-knapping technique may ”satisfy the criteria for being an adaptation (it may be transmitted from parent to offspring and may owe its existence to the fact that it enhanced reproductive fitness), but is neither non-learned nor developmentally robust”; Joyce associates being non-learned and developmentally robust with being based on genes and not being contingent on environmental factors (both examples: (Joyce 2013, p. 3)).

   This shows that the concepts of being a genetically based trait and being an adaptation are not co-extensive.

   An adaptation such as being able to knap stones in a certain way is, of course, not an adaptation in the Darwinian or Evolutionary Psychologist sense but rather a classic case of a learned skill. In the next paragraph, I will expound why being a genetically based trait and being a Darwinian or Evolutionary Psychologist adaptation are not the same.

[119] In people with trisomy 21, there is a great range for each trait to be developed: For example, there is a general tendency for people with trisomy 21 to have a lower intelligence level; the ’impairments’, however, range from mild to severe. 7% of a Dutch group of children with Down Syndrome stayed in mainstream school throughout secondary education (van Wouwe et al., 2014). Environmental factors such as caretakers’ education play a role in developing cognitive abilities (de Graaf et al. 2013), (Couzens et al., 2012) and there are cases of people with Down Syndrome who graduated from university (“Pablo Pineda Ferrer,” 2014).

2. Darwinian adaptations, according to Joyce, may differ from genetically based traits in that the latter are to a great degree invariant in different environments and that the former are not necessarily developmentally fixed; the former may be dependent on environmental triggers or there might be ‘switches’ that lead to very different kinds of development depending on the environment. There could, [120] for example, be grammatical constraints that allow only two variants of building sentences; depending on the linguistic inputs in someone’s childhood, the child may find only one of the two variants grammatical. The two possible outcomes of what the child finds grammatical may even be mutually exclusive; they are still genetically based in the sense that the space of grammatical options is constrained to two options. The child may even need less information to learn the complete set of grammatical rules than, for example, a computer without any knowledge about grammar would need to learn them, even if the computer has a much better memory. We could, for example, imagine that as soon as the child hears a certain word, one of the two sets of grammar is activated, even though this single word does not carry any grammatical information relevant to the grammatical rule that is activated. According to Poverty of Stimulus arguments, this would mean that both complete grammars (or the information about the grammars not given in the environment) were somehow ‘there’ before; the trigger would only have ‘activated’ one of them.

3. Darwinian adaptations, according to Joyce, may differ from genetically based traits in that the latter are to a great degree invariant in different environments and that the former are not necessarily developmentally fixed; the former may be dependent on environmental triggers or there might be ‘switches’ that lead to very different kinds of development depending on the environment. There could, [121] for example, be grammatical constraints that allow only two variants of building sentences; depending on the linguistic inputs in someone’s childhood, the child may find only one of the two variants grammatical. The two possible outcomes of what the child finds grammatical may even be mutually exclusive; they are still genetically based in the sense that the space of grammatical options is constrained to two options. The child may even need less information to learn the complete set of grammatical rules than, for example, a computer without any knowledge about grammar would need to learn

[120] This is an example I have made up.

[121] If we want this problem pair to have a similar structure as the Trolley Dilemma pair in terms of the Doctrine of Double-Effect conditions, we must assume the following: Telling the five people how to treat their disease has the bad effect of occupying the line and thereby keeping the one person from talking to a doctor about their disease, thus letting them die. If we do not consider keeping the one person from talking to a doctor as a bad effect (e.g. because it is an omission rather than an action), the act of helping the five people only has a good and no bad effect. This means that the action would not be within the scope of Doctrine of Double-Effect considerations at all: ”A person may licitly perform an action that she foresees will produce a good and a bad effect provided that four conditions are verified at one and the same time.” If there is no bad effect produced, there is no reason to apply the Doctrine of Double-Effect to check whether an action is permissible.

them, even if the computer has a much better memory. We could, for example, imagine that as soon as the child hears a certain word, one of the two sets of grammar is activated, even though this single word does not carry any grammatical information relevant to the grammatical rule that is activated. According to Poverty of Stimulus arguments, this would mean that both complete grammars (or the information about the grammars not given in the environment) were somehow 'there' before; the trigger would only have 'activated' one of them.

4. This information would be genetically determined, too; the (two possible) outcomes, however, would depend on the kind of trigger and vary greatly.

   I disagree with Joyce in that I believe that in our world, there is no trait that is not contingent on environmental factors at all (be the contingency as trivial as the fact that genetically based facial features can be changed by injuries and those injuries are the environmental factors) and that the difference between developmentally determined traits and traits that are contingent on environmental changes is gradual. I believe that the difference between the two concepts of genetically determined/implemented traits and Darwinian/evolutionarily adapted traits is rather that traits that are Darwinian/Evolutionary Psychologist adaptations, in addition to being inheritable, need to have been selected for (and that there are genetically determined/implemented traits that are not adaptations).

5. We have seen that both, 'determined' and 'Darwinian/Evolutionary Psychologist adapted' traits are based on genetic factors. If you can show that something is genetically based as in non-learned, you can thereby show that one precondition for it to be an evolutionary adaptation is fulfilled. Therefore, people who argue that certain psychological traits are evolutionary adaptations often argue for their non-learnedness, too.

6. If they can show that one condition for that trait to be an evolutionary adaptation is fulfilled, namely, that it is genetically inheritable, this makes it more plausible for that trait to be an evolutionary adaptation; especially, if it fulfils other criteria of evolutionary adaptations as well.

7. As being inherited is a necessary condition for a trait to be an evolutionary adaptation, in order to show that something is not an Evolutionary Psychologist adaptation, it is sufficient to show that it is not inheritable/genetically based/determined/non-learned.

The following Poverty of Stimulus arguments are intended to show that the Trolley Dilemma judgement mechanisms are non-learned. I will show that those arguments do not fulfil their purposes for different reasons. I will do so to argue against the claim that Trolley Dilemma judgement mechanisms have developed as Evolutionary Psychologist adaptations.

6.1.3 A general objection against Poverty of Stimulus arguments

Let me start with some remarks about my methodology. Some arguments I will discuss, especially by Susan Dwyer, do not directly refer to Trolley Dilemma mechanisms but, more broadly, to the moral domain. This is because Susan Dwyer did not work with Trolley Dilemmas systematically [122] when she started comparing the moral to the linguistic domain and looking for similarities between linguistic rules and rule acquisition and moral rules and rule acquisition, neither was there a 'Universal Moral Grammar community'. Just as John Mikhail does in his 2000 dissertation, she refers to Rawls and his proposal to compare linguistic and moral rules [123] in her early works about the Linguistic Analogy in which she extensively discusses Poverty of Stimulus arguments. But, unlike Mikhail, she does not use Trolley Dilemma Type Cases to identify possible principles that might underlie 'moral' judgements. Some of her arguments, however, can be transferred to the domain of Trolley Dilemmas and this is what I will do.

This means that I will not stick to Poverty of Stimulus arguments and objections to them made in connection with Trolley Dilemma Cases, but I will transfer different Poverty of Stimulus arguments and objections to the realm of Trolley Dilemma Cases.

A precondition for Poverty of Stimulus arguments in the moral realm is what John Mikhail, following Chomsky's theory of Universal Grammar, calls "descriptive adequacy". This, quite trivially, means that it is necessary to know what the abilities of someone are in order to propose a story about Poverty of Stimulus in their development. Put differently: If I want to tell you that you cannot have learned what you know because you have not had the experiences to do so, I need to first be able to denominate what exactly it is you know. So, you need to examine the state of my knowledge before you present a nativist or developmental explanation about how I acquired this knowledge. The ability we want to assess in this case is what people need to be capable of in order to judge Trolley Dilemmas in the way they judge them (see, for instance, Mikhail 2000, p. 93).

[122] In her 1999 essay, she writes: "Since, at least to my knowledge, no-one has yet undertaken the search for moral parameters, I am left to speculate." (Dwyer, 1999, p. 177). This is one year before John Mikhail published his dissertation on his variant of Universal Moral Grammar which already uses empirical research on Trolley Dilemmas as source for his principles-and-parameters theory (Mikhail, 2000, p. 95 Chapter 3.1). A large part of Dwyer's thoughts on moral parameters relies on the distinction between conventional and moral rules that children, according to Elliott Turiel (Turiel, 1983), make. Despite criticism of the research underlying this (see, for instance, (Haidt et al. 1993), (Kelly et al., 2007) and, for a review and partial refutation of both and further empirical research (Sousa et al., 2009)), it is a good idea to examine whether there is an innate faculty that enables children to make this distinction, especially with the tools the Poverty of Stimulus argument provides. I will, however, restrict myself to the parts of her theory in which she discusses the Poverty of Stimulus argument in connection to Trolley Dilemmas or she discusses the Poverty of Stimulus argument in a way that could be transferred to Trolley Dilemma Type structures as this is the topic of this dissertation.

[123] "We do not understand our sense of justice until we know in some systematic way covering a wide range of cases what these principles are. A useful comparison here is with the problem of describing the sense of grammaticalness that we have for the sentences of our naive language." (Rawls, 1971, p. 41), referring to (Chomsky, 1965, pp. 3–9). Dwyer, however, points out clearly how her account of moral principles and Rawls' differ (Dwyer, 2006, p. 253/254).

There is a general objection to arguments that require ruling out all kinds of experiential learning that people may have undergone. Noam Chomsky has, charmingly, put this objection the following way:

> ”Of course, in order to demonstrate that there is no relevant experience with respect to some property of language, we really would have to have a complete record of a person’s experience - a job that would be totally boring; there is no empirical problem in getting this information, but nobody in his right mind would try to do it. So what we can do is to find properties for which it is very implausible to assume that everyone has had relevant experience.” (Piattelli-Palmarini, 1980, p. 113), cited after (Pullum and Scholz, 2002, p. 21)

So, we either need to come up with a scientist ‘out of his right mind’ who is willing to do this vexing work, or else we rely on ‘normal’ environments about which we make our claims. This means that empirical and, with a grain of salt, anecdotal evidence that shows that some experience is rarely acquired in most environments is admitted as evidence that it is likely not to be part of the developmental learning process most people go through. If, for example, some grammatical construct comes up only once in 10,000 texts that are examined and those texts represent everyday language, we can argue by induction that this grammatical construct is not part of the normal learning (linguistic) environment. As Pullum & Scholz (2002, p. 21) put it: ”*Our default assumption will be this: a construction inaccessible to infants during the language acquisition process must be rare enough that it will be almost entirely absent from corpora of text quite generally.*” (Emphasis by authors).

To determine whether something is part of a normal learning environment in Trolley Dilemma Cases we can, for example, use psychological studies about how often something is said to children or, less scientifically controllable and probably less generalizable, use our own direct and indirect experiences to at least transport an idea of a ‘normal’ learning environment; both methods come into use in Susan Dwyer [124] and Jesse Prinz [125] (Prinz, 2007a, p. 393).

6.2 The Logical Problem of Language Acquisition

6.2.1 The Logical Problem of Language Acquisition and generativity

Next, I will outline a specific version of the Poverty of Stimulus argument, namely what Kent Johnson calls ”Logical Problem of Language Acquisition” (LPLA).(Johnson, 2004, p.

[124] One example of everyday observations Dwyer makes is the following: ”First, consider the usefulness of the positive moral instruction we offer to children. Usually this takes two forms: either post-hoc evaluations (’You ought not to have broken your sister’s train.’), or unexplained imperatives (’Keep your promises.’).” (Dwyer, 1999, p. 172).

[125] He cites some “casual observations” he made watching adults react to children who harm each other without showing remorse to solidify his point.

571). It requires us to look at structural properties shared by both language and morality. I will start with some general remarks about generativity and subsequently show how this concept can be applied to Trolley Dilemma principles.

Everyone who is a proponent of the Linguistic Analogy points to the similar structure of language and grammar: In both realms, people show rule-governed behavior. This means that they seem to follow rules when they build sentences or judge the permissibility of actions in Trolley Dilemma cases. There seems to be some kind of systematicity when it comes to building sentences and judging Trolley Dilemma cases.

In many studies (e.g. Cushman et al. 2006a; Mikhail 2002; Greene et al. 2009), the subjects' judgements as to whether it is permissible to sacrifice the one person showed a systematic pattern which complied with principles such as

“(a) harm caused by action is worse than an equivalent harm caused by omission;

(b) harms caused as a means to some greater good are worse than equivalent harms caused as a foreseen side-effect of an action”14 (Dwyer et al., 2009, p. 497), based on (Cushman et al., 2006) and

(c) harms caused by a combination of personal force (direct muscle impact, as in pushing someone with a pole) and spatial closeness (as in standing next to someone) are worse than harms caused by a combination of impersonal force (as in dropping someone through a switch-operated trap door) and spatial distance.” (Greene, 2014, p. 23) [126]

Proponents of the so-called Linguistic Analogy think that this net of principles constitutes a moral grammar like the one John Rawls proposed in 1971 in his Theory of Justice (Rawls 1971). [127] They argue that both grammatical sentences and Trolley Dilemma judgements have a law-like structure. This means that we can compress them because they include a regular pattern.[128] We can observe those kinds of compression, e.g. when we tell people to note down telephone numbers like 1010101010 by telling them to write 10 five times (this may also be one reason we can remember numbers like this so easily (Compare Schmidhuber, 2009)): That way, we can express the number much more briefly than if we spell it out in its entirety.

Other instances where we can find rules and therefore compress [129] information are natural laws: Hudson writes that scientific theories are successful when they ”compress

[126] For a review of all principles in question see (Bruers and Braeckman, 2014).

[127] ”A conception of justice characterizes our moral sensibility when the everyday judgments we do make are in accordance with its principles. [...] A useful comparison here is with the problem of describing the sense of grammaticalness that we have for the sentences of our native language.” (Rawls, 1971), referring to (Chomsky, 1965, pp. 3–9); (see also Mikhail, 2000).

[128] I agree with Nicholas Hudson about this concept: ”By compression I refer to the information theoretic concept of reducing the number of bits needed to encode a given representation.” (Hudson, 2011, p. 2).

[129] Here, I am following what Wheeler calls the ”Algorithmic Compression Theory of Laws”: ”The concept of compression is especially important in the computer sciences where, since the dawn of the digital

vast numbers of apparently diverse environmental observations into concise Laws that can sometimes be expressed using nothing more than a handful of symbols." (Hudson, 2011, p. 7), referring to (Schmidhuber 2009). [130]

If we, for example, know the starting conditions of a system we want to know more about (e.g. the height we are dropping a body from) and apply natural laws to them (on Earth, for approximations, Newton's laws) we can predict observations (if we disregard friction: How long it will take until the body hits the ground). We can even mentally simulate situations and their outcome without measuring the starting conditions first: Ernst Mach, for instance, writes that if we know the laws of gravity, we can reproduce "in thought all possible motions of falling bodies". (Mach, 1894, p. 193). Hence, instead of measuring all instances of falling bodies, we can fold the information about their falls into physical laws and unpack this to predict every fall trajectory if we know the starting conditions.

I will transfer this theory about law-likeness and compression to sentences and moral Trolley Dilemma cases.

When sentences follow the grammar of their language (the set of rules of their language), most people will judge that they are grammatically correct. In Trolley Dilemma cases, when the one person that needs to be sacrificed is used as a means to save the five and is killed by a close-up, personal action, most people will judge this action impermissible.

If we want to know whether most people find some sentence grammatically correct or some action permissible, we do not need to ask people to judge every single case; to know how they will judge, we only need to find out whether the sentences comply with the rules of their language or whether the action that is to be judged has properties that make it impermissible. If, for example, we know that in English there is a rule saying that verbs in the third-person-singular form have an 's' at their end, we can imagine any English sentence containing a verb in the third person singular form; if the verb lacks an 's' in the end, we know that most people will judge that sentence as grammatically incorrect. For Trolley Dilemmas, we already know the three laws mentioned above; in the "Footbridge Problem", someone is pushed from a bridge. This is an action, and therefore worse than an omission according to Principle a). The big man is used as a means to stop the train;

age, it has been subject to rigorous mathematical study. In the language of computation, a compression of data is usually described as consisting of two parts: (i) an unstructured string of symbols called the 'input data', and (ii) a list of instructions for interpreting the input data known as a 'program'. The essential idea being that any simple machine which can read and write symbols in accordance with rules can produce (iii) a structured string of symbols called the 'output data' when given (i) and (ii). If the combined length of the program and the input data (measured as number of symbols in some particular computing language) is shorter than the length of the output data, then compression has been achieved."

[130] See also Murray Gell-Mann's formulation: "The best way to compress an account of large numbers of facts in nature is to find a correct scientific theory, which we may regard as a way of writing down in a concise form a rule that describes all the cases of a phenomena that can be observed in nature...A scientific theory thus compresses a huge number of relationships among data into some very short statement." (Gell-Mann, 1988, p. 4), cited after (Wheeler, 2014, p. 13).

this is worse than if his death was merely a side-effect according to Principle b). And in order to use him as a means to stop the train, someone has to touch him, applying personal force from a close distance, which is worse than impersonal force from a large distance according to Principle c). So, we can predict that many more people will find this action impermissible than a situation which involves an omission, where no-one is killed as a means and no-one is killed by application of close-up, personal force.

Trolley Dilemma judgements differ from processes governed by natural laws in that they are decisions made by people (not planets or inanimate falling bodies). They are not only observed, they are produced. Because of this and because Trolley Dilemma cases have a law-like structure, proponents of the Linguistic Analogy assume that people have a cognitive mechanism which enables them to judge the permissibility of actions in Trolley Dilemma cases according to principles, just as many linguists think that people have a cognitive mechanism that enables them to judge the grammaticality of sentences: something like a judgement-factory.

As John Mikhail puts it:

> "The argument for moral grammar holds that the properties of moral judgement imply that the mind contains a moral grammar: a complex and possibly domain-specific set of rules, concepts and principles that generates and relates mental representations of various types. Among other things, this system enables individuals to determine the deontic status of an infinite variety of acts and omissions." (John Mikhail, 2007a, p. 144)

Susan Dwyer et al. (Dwyer et al., 2009, p. 489) say about generativity: "Moral competence requires the capacity to produce and understand an unbounded range of morally significant judgments and actions."

Linguists have various definitions for the term 'generativity'. [131] To me, a set of rules

[131] This poses a problem insofar as most proponents of the Linguistic Analogy cite Noam Chomsky, but not all of them define 'generativity' as Chomsky did (see Dwyer's use). According to (at least the later) Chomsky, a grammar is generative when "its principles can be explicitly stated by the linguist in a format suitable for functioning as the premises in a derivation", a definition used by Linguistic Analogy proponent (Mikhail, 2000, p. 211), referring to (Chomsky, 1965, pp. 3–9). This interpretation the late Chomsky insisted on, however, is hardly applicable to a cognitive system as in "the system generates sentences" (as Mikhail does without further defining the term); it can only be applied to the grammar produced by the cognitive system. Another meaning that has been ascribed to "generate" by Chomsky is to "'map' or 'recursively map'" ((Ney, 1992, p. 447), referring to (Fodor and Pylyshyn, 1988) and (Chomsky, 1991, p. 430)). If this implies that an infinite set of sentences is being specified (the correct set of sentences) and assuming that the verb "to map" reflects a more process-oriented way (hinted at by the term 'recursive') to specify sentences, this notion is closer to the way Susan Dwyer is using the term in the citation above: Namely in a way that includes the process of the production of judgements, not only the systematic structure that can be ascribed to the patterns of those judgements. This also holds for the everyday language use of "to generate", as "to produce", which is also used by some linguists ((Ney, 1992, p. 447), referring to (Broderick, 1975, p. 14)). In the source cited by Ney, Broderick writes: "... the user of a language [...] can produce an infinite number of sentences. Hence the term *generative*." (emphasis by author) and Cattell's interpretation of Chomsky from 1972: "Chomsky's grammar is '...not only transformational but also GENERATIVE. That means it consists of a set of rules for generating the

is generative if it can be systematically applied to previously unknown situations and, in the case of moral grammar, if it can produce judgements about the moral permissibility of actions depending on their features.

6.2.2 Generativity in the moral domain

I will illustrate this property more extensively using the “Doctrine of Double-Effect”. Joseph T. Mangan formulated the Doctrine in its ”full modern dress” back in 1949:

> ”A person may licitly perform an action that he foresees will produce a good and a bad effect provided that four conditions are verified at one and the same time:
>
> 1) that the action in itself from its very object be good or at least indifferent;
> 2) that the good effect and not the evil effect be intended;
> 3) that the good effect be not produced by means of the evil effect;
> 4) that there be a proportionately grave reason for permitting the evil effect.”
>
> (Mangan, 1949, p. 43)

In this formulation however, the Doctrine of Double-Effect is hard to operationalize for the purpose of psychological research. I will find a formulation for the Doctrine that, firstly, unifies the different accounts of the Doctrine of Double-Effect in the Trolley literature and secondly, makes it easier to decide whether subjects judged in accordance with the Doctrine of Double-Effect based on properties of the Dilemma. This will not change any of the results in the cited literature (what they call judgements in accordance with the Doctrine of Double-Effect, unless explicitly stated otherwise by me, remain judgements in accordance with the Doctrine of Double-Effect). But we will have some common ground and will be referring to the same Doctrine when we talk about it. The Doctrine of Double-Effect as defined above is hard to operationalize for our Dilemmas for two reasons.

The first reason has to do with Principle 2). This principle is at the core of the Doctrine of Double-Effect and mirrors many people’s intuitions: [132] Many would find it permissible to kill people who are standing next to a military target as a side-effect when bombing the military target itself. If the same situation occurred with the same bombing, but the intention of the bomber was to kill the people standing next to the military target and the destruction of the target was only a side-effect, by far most people would find the bombing unacceptable. It is, however, not obvious in the dilemmas presented whether the agent intended an effect or not. And it is even harder to determine whether the judging person

sentences of the language. [...] Chomsky tried, successfully, to create rules which would enumerate an infinite number of sentences’.” (Cattell, 1972, pp. 28–29), cited after (Ney, 1992, p. 447).

[132] Although it has been controversially discussed: See (Scanlon, 2010), (Thomson, 1999), (Holton 2017).

thinks that the agent intended an effect. This is especially so because subjects who judged Trolley Dilemmas often provided unreliable reasons for their judgements: Those reasons often did not explain why they judged one action permissible and another impermissible (Cushman et al. 2007). This means that we cannot always trust their reports based on introspection and they themselves might not know whether they ascribed intentions to an agent either.

The second problem has to do with Principle 4): The concept of a "proportionately grave reason" for permitting an evil effect is not very clear. What is a "proportionately grave" reason to kill someone? John Mikhail solves the first problem by deducing whether the judging person perceives an action as intended. To him, the default intention the judging person projects on the agent is good; they presuppose that the agent has only good intentions and this only changes when "evidence to the contrary is presented" (Mikhail 2002, p. 76): This is the case when the death of the person is necessary to save the persons. In the "Footbridge Problem", for instance, we cannot really believe that someone who pushes the man from the bridge did not intend that the person be run over: Even if we believe that their main intention was to save the five persons, these would not have been saved had the train not hit the big man. If the big man had not existed, the train would have run over the five people. Even if the overall intention of the agent was to save the five persons, part of their plan must have been for the train to hit the big man; otherwise, the five persons would not have survived and there would have been no point in pushing the man from the bridge. In the "Side-track Problem", the one person does not need to die in order to stop the train: It would be sufficient to divert the train to the side-track to save the five people on the main track, regardless of whether someone is standing on the side-track. If we assume that the agent's intentions are good and the agent's intention in diverting the train to the side track is not to kill the one person but to save the five, running over the one person is not even a sub-intention. Hence, in Trolley Dilemmas, the agent may intend two things, the overall outcome (the good effect, here: saving the five) and something that leads to it (in the "Footbridge Problem": The train hitting the big man and killing him [133]). I will proceed to call the second one 'sub-intention'. In this respect, Trolley Dilemmas are more complicated than the bombing dilemma I have discussed before: In the latter, the agent either intends the good or the bad effect. In Trolley Dilemma cases, the agent can intend the overall good effect (saving the five) and at the same time intend the bad effect that comes with it (killing the one). We will not even question whether the judging person thinks that the agent intends the overall good effect. We will assume that the judgement is affected only by the issue of whether the agent intended the bad effect that comes with the good overall effect and we will follow Mikhail in assuming that people ascribe bad intentions when the bad effect is necessary to cause the good effect. If the bad effect is necessary to achieve the good effect, we could

[133] I will assume here for the sake of simplicity that hitting the man and killing him is one event.

say in these Trolley Dilemma cases that the bad effect is a means [134] to achieving the good effect. So, according to Mikhail's operationalization, if the bad effect is a means to achieving the good effect, the bad sub-effect (and not only the good overall effect) is intended.

A slightly different version of the Doctrine of Double-Effect as Dwyer et al. state it in 2009 does more justice to this operationalization than Mangan's formulation: "The Doctrine of Double-Effect holds that an action that is known to have a bad consequence is permissible only if (a) a good consequence is *intended*; (b) the bad consequence is merely a *foreseen* consequence of the intended action; (c) the bad consequence is not a necessary means to bringing about the good effect; and (d) the good consequence is sufficiently important to justify the foreseen bad consequence." (Dwyer et al., 2009, p. 505, emphasis in the original). The good consequence in a) is the overall consequence; the "bad consequence" (killing one person) and the sub-(non-)intention that belongs to it are conflated into the term "foreseen": In our definition, if the bad consequence is a side-effect and not a necessary means to the good overall consequence, it is merely "foreseen" and not intended. If the bad consequence is a means to the good overall consequence, it is intended. The term 'bad consequence' might be a little confusing here because the bad consequence, if it is a means to achieving the good overall consequence, comes causally before the good overall consequence (the end). The bad consequence (e.g. the big man's death) of the initial action (e.g., pushing the big man) entails the good consequence (e.g., stopping the train and saving the five). The good consequence is a consequence of the bad consequence. Dwyer et al.'s version, however, does not solve the second problem: c), "the good consequence is sufficiently important to justify the foreseen bad consequence", is similarly vague as Mangan's "4) that there be a proportionately grave reason for permitting the evil effect." This is why I will use a conglomerate of Mangan's (Mangan, 1949, p. 43), Mikhail's (Mikhail, 2002, p. 11), referring to (Mikhail, 2000) and (Fischer and Ravizza, 1992) and Dwyer et al's (Dwyer et al., 2009, p. 505) versions and replace Dwyer's last formulation with Mikhail's formulation: "The good effects outweigh the bad effects." From now on, I will use the following, optimized version of the Doctrine of Double-Effect:

> A person may licitly perform an action that they foresee will produce a good and a bad effect provided that four conditions are verified at one and the same time:
>
> 1. A good overall consequence is intended;
> 2. The bad consequence is merely a foreseen consequence of the intended action; [135]

[134] If we define 'means' independently from 'intention'.

[135] The intended action here is the initial act with the intention to achieve the overall good effect.

3. The bad consequence is not a necessary means to bringing about the good effect; and
4. The good effects outweigh the bad effects. [136]

These are the properties that make an action permissible according to a version of the Doctrine of Double-Effect. This Doctrine (or something similar) seems to be underlying judgements in Trolley Dilemma cases (Cushman et al., 2007), (Mikhail, 2002), (Mikhail, 2011). [137] Hence, an action is permissible to more subjects in psychological studies if it fulfils the conditions cited above than if it does not. In the "Footbridge Problem", those conditions are clearly not fulfilled; however, the big man also experiences close-range, personal-force harm when someone pushes him from the bridge. So this Dilemma is not perfect for observing the relevance of the Doctrine of Double-Effect: We do not know how many people think pushing the man from the bridge is impermissible because of the close-range personal harm and how many think it is impermissible because it does not meet the criteria of the Doctrine of Double-Effect or, if both principles interact, how they do so. Cushman et al. tailor-made two Trolley Dilemmas that isolated the conditions of the Doctrine of Double-Effect from those other principles. They constructed two cases, one of which fulfils the conditions of the Doctrine of Double-Effect, the so-called "Oscar" case, and one of which does not fulfil them, the so-called "Ned" case. In the "Oscar" case, "Oscar is walking near the train tracks when he notices a train approaching out of control. Up ahead on the track are 5 people. Oscar is standing next to a switch, which he can throw to turn the train onto a side track. There is a heavy object on the side track. if the train hits the object, the object will slow the train down, giving the 5 people time to escape. There is 1 man standing on the side track in front of the heavy object. Oscar can throw the switch, preventing the train from killing the 5 people, but killing the 1 man. Or he can refrain from doing this, letting the five die." (Cushman et al., 2007, p. 7). In the "Ned" case, "Ned is walking near the train tracks when he notices a train approaching out of control. Up ahead on the track are 5 people. Ned is standing next to a switch, which he can throw to turn the train onto a side track. There is a heavy object on the side track. If the train hits the object, the object will slow the train down, giving the men time to escape. The heavy object is 1man, standing on the side track. Ned can throw the switch,

[136] I am aware that by replacing Mangan's and Dwyer's fourth rule with Mikhail's, I give the entire Principle a push towards utilitarianism. I am also aware that many would not agree that saving five is a proportionately grave reason to kill one. The results of the Trolley Dilemma studies reviewed here show that most people would judge whether saving five is a proportionally grave reason to kill one depending on other properties of the situation (amongst which is whether the Doctrine of Double-Effect is met). Due to constraints of space I cannot argue about normative theories here, and taking into account that this part of the thesis is only supposed to make structural properties of Trolley Dilemma Judgements more easily accountable for, I will just assume here that, nothing else considered, saving five outweighs one person's death.

[137] For an extended discussion of the empirically most adequate interpretation of the Doctrine of Double-Effect, see Fitzpatrick 2014.

preventing the train from killing the 5 people, but killing the 1 man. Or he can refrain from doing this, letting the 5 die." (Cushman et al., 2007, p. 6).

We can assume that both cases fulfil Condition 1 of the Doctrine of Double-Effect: Oscar and Ned will probably throw the switch to save the five people, not to kill the one person. Condition 4 is fulfilled in all Trolley Dilemmas (if we judge the survival of five persons to outweigh the survival of one person). Conditions 2 and 3, however, are met in the "Oscar" case, whereas they are not met in the "Ned" case: In the "Oscar" case, 2. The bad consequence is merely a foreseen consequence of the intended action and 3. The bad consequence is not a necessary means to bring about the good effect. In the "Oscar" case, the train hits the person on the track and kills him. Oscar can foresee this effect. But the train would slow down even if there was no person on the track because the heavy object would do the job. The person on the track is not necessary to slow the train down. This is why the bad consequence (the person on the track getting killed) is not a necessary means for bringing about the good effect, saving the five people on the main track. And now we can see why we needed the term "necessary" in our formulation of the Doctrine of Double-Effect: When a train hits one person in front of a heavy object and slows down, the person in front of the heavy object is a means to stopping the train. He is, however, no necessary means: The train would have hit the heavy object behind him and stopped, had he not been there. So if we only assume that agents have bad intentions when we have evidence for them, Oscar would not have bad intentions when he throws the switch: He foresees that the man on the side-track is going to get killed, but as his death is not necessary for the train to stop, he does not need to use him as a means to stop the train and therefore, according to our prior working definition, not intend him to die. In the "Ned" case, however, it is necessary for the train to hit and therefore kill the person on the side-track to slow the train down. While it is only a side-effect in the "Oscar" case that a person gets hit, in the "Ned" case, the person is used as a necessary means to slow down the train. This is why Condition 2 is not met; the bad consequence is necessary to bring about the good effect. If the train did not hit the man, the five persons would not be saved. The train hitting the man must be part of "Ned's" plan if he throws the switch to save five persons; the bad action is not merely a foreseen consequence of the intended action. It is a wanted, an intended consequence. This is why Condition 3, "the good effect is not produced by means of the bad effect", is met in the "Oscar" case, but not the "Ned" case.

I have illustrated one way to apply the Doctrine of Double-Effect to a situation. The "Oscar" case fulfils all conditions of the Doctrine of Double-Effect, whereas the "Ned" case only fulfils two of them. According to the Doctrine, throwing the switch in the "Oscar" case is permitted while in the "Ned" case, it is not. But what does this have to do with generativity? We can apply the Doctrine of Double-Effect to many situations. The generativity lies in the list of situations to which the Doctrine can be applied: You can apply the Doctrine of Double-Effect to an unbounded range of situations and judge whether they are permissible.

Likewise, you can create an unbounded number of actions in a way that they meet the conditions of the Doctrine of Double- Effect and are therefore permissible to people who base their judgements on it.

The parallel to language is as follows: You can assess whether an unbounded number of sentences is grammatically permissible if you use a set of grammatical rules: If all grammatical criteria are met in a sentence, it is grammatically correct. And you can generate an unbounded number of grammatically permissible sentences using a set of grammatical rules. [138]

Let us now consider a different setting to which we can apply the Doctrine of Double-Effect: The following scenarios are two imaginary situations with a structure similar to the two Trolley Dilemmas described above:

Scenario 1

Assume there is a space station with six persons. All of them are suffering from a deadly illness. They only have one radio line to earth to talk to a doctor. Five of the persons suffer from the same disease; one of the persons, however, suffers from a different disease. All of them need help quickly or they will die. The only way they can be helped is if the doctor tells them how to treat their disease. There is only time to tell the people in the station how to treat one disease before they die. Is it permissible to call them and tell them how to treat the disease afflicting the five persons, thereby occupying the line and letting the person with the different disease die because this person cannot talk to a doctor quickly enough to get help?

Scenario 2

Now let us assume a space station with just the same situation, but in addition the station is running out of oxygen and can only supply five persons with enough oxygen until they can land on earth; if there were six persons on board, the five with one disease would die of oxygen deprivation before the station landed because the effects of their disease would lead to a temporarily deteriorated ability to store oxygen, and the one with the other disease would survive. Is it permissible for the doctor to call them and tell the people on the station how to treat the disease afflicting five persons, thereby occupying the line and letting the person with the different disease die because this person cannot

[138] While it would not only make sense to judge the grammaticality of sentences but also to create grammatical sentences themselves, you can act according to your permissibility judgements in reality but in most situations it would make no sense to create moral dilemma situations. The Linguistic Analogy is not quite accurate here.

talk to a doctor quickly enough to get help? In that case, the one person's death is not only unavoidable but also necessary to save the five others' lives. [139]

As before, we assume that the overall good effect that is intended is to keep as many people as possible alive until they return to Earth. I will now check the two situations for their accordance with Doctrine of Double-Effect conditions. [140] As a brief reminder, the conditions are the following:

> A person may licitly perform an action that they foresee will produce a good and a bad effect provided that four conditions are verified at one and the same time:
>
> 1. A good consequence is intended;
> 2. The bad consequence is merely a foreseen consequence of the intended action;
> 3. The bad consequence is not a necessary means to bringing about the good effect; and
> 4. The good effects outweigh the bad effects.

In both cases, we have decided to assume that Condition 1 is fulfilled: A good consequence is intended, namely to keep five persons alive. In the first scenario, Condition 2 is fulfilled in addition: The bad consequence, namely the line being busy and the one person dying of their disease because they cannot get instructions how to treat it, is only a foreseen consequence of the action with the good effect, namely telling the five persons how they can treat their disease. If the good effect, however, is to save the five persons' lives (at least until they get back to Earth), then Condition 2 is only fulfilled in the first scenario, but not in the second: In the first scenario, the one person with the different disease does not need to die for the others to survive. In the second scenario, by contrast, the five

[139] This scenario may seem a little forced or far-fetched. It has the space station component because I wanted to create a scenario where the actual death of the one person was necessary to save the others. Having them stop breathing seemed one of the things closest linked to death to me. There is, however, a different (but similarly complex) scenario in (Waldmann and Wiegmann 2012, A Double Causal Contrast Theory of Moral Intuitions in Trolley Dilemmas) where someone's heart has to be stopped to save the other five. Stopping a person's heart could be considered even closer to their death than making them stop breathing. See Fisher, Ravizza and Copp 1993 on the Problem of "Closeness".

[140] If we want this problem pair to have a similar structure as the Trolley Dilemma pair in terms of the Doctrine of Double-Effect conditions, we must assume the following: Telling the five people how to treat their disease has the bad effect of occupying the line and thereby keeping the one person from talking to a doctor about their disease, thus letting them die. If we do not consider keeping the one person from talking to a doctor as a bad effect (e.g. because it is an omission rather than an action), the act of helping the five people only has a good and no bad effect. This means that the action would not be within the scope of Doctrine of Double-Effect considerations at all: "A person may licitly perform an action that she foresees will produce a good and a bad effect provided that four conditions are verified at one and the same time." If there is no bad effect produced, there is no reason to apply the Doctrine of Double-Effect to check whether an action is permissible.

persons will die even if they are treated if the one person does not die. The one person's death, then, is necessary to bring about the good effect, the five persons' survival. This means that their death is part of the plan to save the five people which makes it more than a foreseen side-effect: Their death, in Scenario 2, is a means to saving the five because otherwise they will run out of oxygen. This makes the one person's death a necessary means to bring about the good effect, the survival of the five others. [141]

In Scenario 1, by contrast, the others would not die if the one person survived as well; the one person's death is not a means to saving the five. This means that Condition 3 is fulfilled in Scenario 1, but not in Scenario 2. In both scenarios, as in the Trolley Dilemma cases, the good effects outweigh the bad effects if what counts is the number of lives saved. This means that Condition 4 is fulfilled in both cases.

We can now judge that if we interpret the scenarios as I proposed, according to the conditions of the Doctrine of Double-Effect, the action in the first scenario is permissible and the action in the second scenario is not.

What I have shown by now is that the Doctrine of Double-Effect as a set of rules/conditions can be applied to different situations. I have created scenarios that are similar in structure to the "Ned" and "Oscar" Trolley Dilemma scenarios to show how situations with very different settings (train tracks vs. space) can have similar properties with regard to the Doctrine of Double-Effect. Many pairs of scenarios are cited in the literature as examples for Doctrine of Double-Effect decisions, for example the (not-always-)hypothetical scenario where someone who is in pain and terminally ill dies more quickly when they are treated with pain killers. According to the Doctrine of Double-Effect, if the intended effect is the person's death, this treatment is not permissible; if the intended effect is easing their pain, the treatment is permissible. Those situations, however, differ only regarding the overall intentions (Condition 1). Therefore, I have chosen scenarios that differ in Conditions 2 and 3 just as the "Ned"/"Oscar" cases: If the one person survived, the five persons would die. The one person's death is a necessary means to saving the five.

The unbounded space of situations to which we can apply the Doctrine of Double-Effect contains situations like these and situations with similar properties. I have shown what "generativity" means using one of the principles underlying the Trolley Dilemma judgements as example. We can similarly apply principles such as "inflicting close-up personal force is worse than inflicting long-distance impersonal force" to situations in a generative manner: There can be various situations that fit either the criteria that define "personal harm" (close-up, direct muscular force) or the criteria that define "impersonal harm" (distant, no direct muscular force). Depending on whether they fit these criteria, we will be able to judge killing a person to save five either rather impermissible if “personal

[141] This, in turn, is only true if the ground personnel does not operate by a step-by-step structure but is aware of the inevitability of the one person's death to save oxygen. If the ground personnel operates on something like "we will first save as many people as possible" and then see whether there is enough oxygen for all of them, Conditions 2 and 3 are fulfilled and the person's death is not a means to saving the five.

harm" is involved, or rather permissible if no "personal harm" is involved.

Generativity enables us to make (regularized) judgements about "actions in actual and hypothetical, novel and familiar cases" (Dwyer, 2006, p. 237). Remember John Mikhail's quote:

> "The argument for moral grammar holds that the properties of moral judgement imply that the mind contains a moral grammar: a complex and possibly domain-specific set of rules, concepts and principles that generates and relates mental representations of various types. Among other things, this system enables individuals to determine the deontic status of an infinite variety of acts and omissions." (John Mikhail, 2007a, p. 144)

This generativity is important for an argument that comes originally from linguistics and seems [142] to be one of the arguments that proponents of the Linguistic Analogy use frequently. I will spell out this argument first and detect and cite cases where proponents of the Linguistic Analogy seem to refer to it. Afterwards, I will expound why I think this argument is not apt to show that something is not learned and therefore not suitable as a Poverty of Stimulus argument. For these first arguments, I will assume that children learn things in the way the Logical Problem of Language Acquisition challenges: Like little scientists, acquiring very precise rules (see next Chapter 6.2.3). This means I will make the stance of the proponents of the Logical Problem of Language Acquisition as strong as possible and argue against them in their own framework. After that, I will make a proposal about how the learner might acquire the principles in question that, in my opinion, is closer to scientific evidence about the way people learn. To do this, I will rely on Fiona Cowie's and Steven Pinker's proposal how linguistic principles might be acquired as well as Jesse Prinz's sentimentalist framework.

6.2.3 The Logical Problem of Language Acquisition: The issue with "Learning like a little scientist"

In languages, there are rules that determine how correct sentences are built. They can also be used to assess whether a sentence has been built correctly; they are correctness criteria. We have learned that these rules are generative: They can be used to build an unbounded number of sentences or be applied to them.

One early theory about how those rules are learned is that they are acquired empirically, by hypothesis building: Children learn languages like 'little scientists'. They experience a set of sentences and build hypotheses about the regularities that describe those sentences correctly. Using induction, they build new sentences using those rule-related hypotheses.

[142] They do not refer to it by name but I will later cite sources where proponents of the Linguistic Analogy make arguments that are most likely based on the Logical Problem of Language Acquisition.

When they come across sentences with rules that do not seem to converge with their set of rules, they change their respective hypothesis.

I am not claiming here that it is a wide-spread assumption amongst linguists nowadays that children learn this way. I consider this example because in the literature about the Linguistic Analogy, there is discussion of negative evidence which is clearly rooted in the tradition of Logical Problem of Language Acquisition arguments. To be fair, however, this logical problem has still been discussed by well-known (and some not-yet-well-known) linguists during the past fifteen years (see, for instance, (Pinker, 2004), (Cowie, 2010), (MacWhinney, 2004), (Johnson, 2004), (Finley, 2012)).

Presuming that children learn this way, the following problem (LPLA) might apply: If learners happen to have a hypothesis that includes the correctness of too many sentences, they will without negative evidence not be able to correct this hypothesis. Say, the learner is hypothesizing a rule that includes all sentences of the actual language that is spoken in their environment. The set of sentences they are hypothesizing and that would be generated by their rule, however, includes some additional sentences that the actual language does not include. This may be because the rule they have hypothesized is too general in scope, such as, for instance, that past tenses are formed by attaching ”-ed” to the present tense (”count” - ”counted”) for every verb. This would then lead to forms like ”she goed”. If they came across someone forming the past tense ”she went”, they would hypothesize a grammar in which both sentences, ”she goed” and ”she went”, are correct. The rule they would apply then would be something like ”the past tense of ’to go’ can formed both by adding ’-ed’ to the present tense, and by building a new form such as ’went’, and both forms are correct.” If no-one told them that ”goed” is not a correct past tense, hence, if they were not provided with negative evidence against their over-generalized rule, they would keep thinking it is a correct form. And, as Cowie rightly remarks, the mere non-occurrence of sentences does not itself constitute negative evidence as many sentences are never uttered, both correct as well as incorrect ones (Cowie, 1997, p. 21). Kent Johnson presents a nice formulation of this argument which is often termed the ”Logical Problem of Language Acquisition”: ”If a child learning a given language doesn’t use negative evidence, then it is logically possible [143] that she would begin to acquire a language whose grammatical sentences included all those of the target language, plus some more.” (Johnson, 2004, p. 571)

An example that Johnson gives is the following:

> ”[...] if a child was learning English, she could begin to acquire a language

[143] And, indeed, this is not only logically possible, which is a very careful formulation: It is, to many, an open question as to how language acquisition can happen in those circumstances. As Fiona Cowie summarizes their view: ”The challenge posed by the Logical Problem is thus to explain why people do not make these sorts of errors [certain sorts of plausible but false grammatical hypotheses - in particular over-general hypotheses], given that there is nothing in their experience to prevent their doing so. How could language acquisition occur successfully in the face of such massive evidential deprivation?” (Cowie, 1997, p. 18).

> that included all the grammatical sentences of English, plus sentences of V[erb]S[ubject]O[bject] clausal order, like **Kicked Mary the boy* and **Sang Susan*. If the child received no information that these extra sentences aren't part of English, then she would have no way of knowing that she was not learning English, but a syntactically more expressive language instead." (Johnson, 2004, p. 571, emphasis by author)

This means that, without negative evidence, children would judge all sentences of English grammar as correct, plus sentences that are not considered to be correct English. The set of rules that the child would assume to underlie English grammar would be too unspecific, or maybe not contain enough exceptions. One example for such a form that is erroneously produced without considering exceptions is the regular version of an irregular plural form such as "mouse"/"mouses" (Ramscar and Yarlett, 2007, p. 927). Here, the rule for regular plural forms is applied although the correct plural of "mouse" is the irregular form "mice", which makes it an exception to the rule. Another example for a rule that is applied too broadly is, for example, the assumption that "John is possible to leave" is a correct English construction because "*It is likely that John will leave, John is likely to leave* and *It is possible that John will leave*" are correct constructions (Cowie, 1997, p. 18; emphasis by author).

6.2.4 LPLA arguments applied to Trolley Dilemmas

What makes the Logical Problem of Language Acquisition a version of a Poverty of Stimulus argument? It is not in itself, but it is used as part of a Poverty of Stimulus argument. The argument goes something like this: 1. If their environment provides no frequent negative evidence and no explicit instruction (in the form of grammatical rules), we would expect children to learn overgeneralized rules that apply to too many cases. 2. There is no or very little negative evidence provided ((Cowie, 1997, p. 21), referring to (Chomsky et al., 1959), (Brown et al., 1969) and (Brown and Hanlon, 1970)) and if so, children tend to ignore it ((MacWhinney, 2004), referring to (McNeill, 1966)). 3. Adults do normally not use overgeneralized rules to produce language. 4. It is likely that there is something different from negative evidence that keeps children from learning overgeneralized versions of languages. 5. This something involves constraints or some other information (that may be inaccessible to consciousness) about language that is genetically implemented and reduces the space of learnable languages for humans as compared to the space of logically possible languages. [144] Those constraints thereby keep children from learning certain rules that are too broad, such as that "possible" can be used in the same manner as "likely" and therefore "John is possible to leave" is a correct English sentence.

[144] Of course, this space is already limited in humans by things like memory and time constraints.

[145]

How can we transfer this argument to the realm of morality? As I just have shown, generativity is something that grammars and Trolley Dilemma principles have in common. And to be able to generate new pairs of actions and judgements - as in the situations with the Trolleys and the space stations and the matching (im)permissibility judgements - it is necessary to acquire the rules or principles that those judgements follow. In our example cases, this principle was the "Doctrine of Double-Effect". Do children acquire those moral principles as they acquire their linguistic competence? Can we parallel the two acquisition processes?

6.2.4.1 Is explicit exposure to moral principles necessary or sufficient for acquiring them? Proponents of the Linguistic Analogy have proposed that deducing such moral principles from a 'normal' environment is not trivial: They hold that children cannot extrapolate principles simply from encountering moral judgements. According to them, even if the children receive some explicit instructions, this will not suffice:

Firstly, Susan Dwyer presents an argument why children cannot learn the principles merely from examples they encounter. This is not exactly a case of the Logical Problem of Language Acquisition, but as it also refers to the general impossibility to deduce the moral principles underlying moral judgements, I will discuss it here. It is an even more basic form of the Poverty of Stimulus argument than the Logical Problem of Language Acquisition, questioning whether it is possible to extrapolate any regularities from behavior at all, may they be over-generalized or not. She says that "there is the matter of telling the difference between rule-governed behavior and accidentally-regular behavior." (Dwyer, 1999, p. 172). [146] She introduces the following example:

In a girl's household, there is a rule "that glass containers go in the right-hand side of the recycling box and plastic containers go in the left-hand side of the box." So, the girl observes that every time someone throws away a glass container, it is thrown in the right-hand side of the recycling box. But there is also a cereals box which is placed on the breakfast table in a particular orientation, because the person in charge of setting the breakfast table is left-handed (Dwyer, 1999, p. 172). So how does the girl distinguish between the random orientation of the cereals box and the rule-governed placing of the glass containers?

I do not think this example is optimal as in the recycling case, *everybody always* deposits their bottles this way while *only one person* sets the breakfast table, albeit repeatedly. The

[145] Children actually do make plural mistakes like "mouse"/"mouses", so that those constraints only account for those over-generalizations that children do not make (see Ramscar and Yarlett, 2007). This reminds me of internet memes that are supposed to mimic toddler speech, such as "I can has cheezburger [sic]", that over-generalize the exception, 'has' as third person singular form only, to the first person form although 'have' would be correct for most forms (including first person)

[146] A further issue, to her, is the distinction between moral and conventional behavior. As I have argued before, this distinction has by now been widely criticized.

first case seems to be more systematic and therefore more bound to be rule-governed. We could, however, imagine that everyone always puts the breakfast cereals on the left-hand side of the table merely because it is more practical.

One could argue whether placement that is more practical is merely accidental or whether a regularity that emerges out of convenience is not somehow governed by the rules of convenience (that may be derived by physical laws or psychological effects such as sticking to an old habit that used to be energy-saving). However, the same could be asked about the orientation of the box that is only the way it is because the person setting the table is left-handed: This regularity seems to be due to convenience, too. I think, in Dwyer's case, "accidental" means "not governed by normative rules [147]": This is what captures the difference in her examples best. I came to that conviction because, amongst other indicators, she mentions "rule governed regularities and merely accidental regularities", where rules seem to be human-made rules. Instances of rule-governed regularities are, according to her, that "forks go on the left for right-handed diners" and "promises ought to be kept" (all citations in the last sentence: Dwyer 2006, p. 240).

I will discern the two types of recurring behavioral patterns, calling the one "rules" (what Dwyer calls "rule-governed regularities") and the second "regularities" (what she calls "accidental regularities").

The question Dwyer has posed and poses again in 2006 (Dwyer, 2006, p. 240) is "how, absent explicit instruction, will she learn to discriminate between the rule-governed behavior concerning recyclables and the merely accidental but regular placement of the cereal box?"

This distinction, however, only makes sense if we believe that children treat normative rules unlike regularities in behavior that are not based on rules by humans. Whether children behave differently when they learn rules as compared to when they learn regularities (where both are not explicitly stated) is an empirical question that still needs to be investigated. [148] We could, for example, assume that children just imitate all kinds of behavioral patterns or that they sometimes try to deviate and see whether they get corrected. Then, they would not need to discern between those two kinds of regularities in acquiring them, only later in case people correct them explicitly. They might however be more careful not to diverge from regular behavior that they think is according to normative rules than from regular behavior that is due to practical reasons or habits without any

[147] Whereby I understand "normative rules" in a very wide sense that refers to moral as well as social norms or other norms, everything that implies an 'ought'.

[148] Note that this is a different distinction than the distinction between moral and conventional rule-following which, as I have mentioned above, has been investigated extensively. Of course, we could hypothesize that children treat conventional rules just as they treat regularities and that they treat moral rules differently or that there is a continuum from regularities over conventional rules to moral rules in terms of behavior. However, what we want to explain here is how children are able to acquire the content of moral principles, not how they come to behave differently towards moral principles and whether they stick to them independent of authorities. This might be another Poverty of Stimulus argument, but it does not refer to the principles underlying judgements in Trolley Dilemmas.

reason (e.g. the habit of getting with the left foot first). We do not have the means to decide this question here and the aim of this text is to assess how children acquire the content of the principles underlying Trolley Dilemma judgements, not how they might treat those principles, once acquired, differently than regularities in behaviors that they have learned to copy. Hence, I will answer the question "how, absent explicit instruction, will she learn to discriminate between the rule-governed behavior concerning recyclables and the merely accidental but regular placement of the cereal box?" with "Are you sure she does discriminate between rule-governed behavior and regular placement?". Shifting the burden back seems to be a trivial move here. But Poverty of Stimulus arguments rest on the assumption that we are not able to explain abilities except by means of domain-specific learning mechanisms with built-in information. If we might not even have those abilities, there is nothing to explain.

There is, besides Dwyer's claim about rules and regularities, something even more interesting to what she said about the topic. The reference to "explicit instruction" she makes in the quote above could remind us of the Logical Problem of Language acquisition which has the following presupposition: If we teach people explicit rules, they will not overgeneralize because they do not need to infer the rules underlying their linguistic input anymore. In Dwyer's case, however, what Dwyer's learner gains from "explicit instruction" is something more modest:

In the text cited above, Dwyer argues that explicit instructions might show children that there is a rule they should indeed comply to or, at least, that without explicit instruction they probably would not be able to individuate rule-bound behavior; she leaves open whether explicit instruction is necessary or sufficient for children to learn that some behavior involves following rules. Anyway, according to her line of argumentation, if someone told them that they should put the plastic bottles here and the glass bottles there, that would be a safe indication that this behavior is not random and that they act regularly in this way because they comply with rules that matter to their society (Dwyer 2006, p. 240). If this interpretation is right, children would not have problems distinguishing regular and rule-governed behavior (assuming they do that) if they had explicit instructions. And according to the definition of the Logical Problem of Language Acquisition, as I have mentioned above, over-generalization would be no problem either: If children had the actual rules that are correct for the generation of all their linguistic/moral actions or judgements, they would not need to learn them by extrapolating them from everyday examples and hence have no need to try out rules, including over-generalized rules.

6.2.4.2 Is explicit exposure to moral principles necessary or sufficient for acquiring them? If this is correct, it would be sufficient for children to have explicit instructions of any kind to learn that there are principles underlying moral judgements (and they are not just random regular behaviors) and instructions in form of the actual correct principles to learn the actual correct principles of their target morality.

The principles we are talking about here are the principles stated in "The Linguistic Analogy: Motivations, Results and Speculations" (Dwyer et al., 2009).

"(a) harm caused by action is worse than an equivalent harm caused by omission;

(b) Harms caused as a means to some greater good are worse than equivalent harms caused as a foreseen side-effect of an action" [149] (Dwyer et al., 2009, p. 497), based on (Cushman et al., 2006) or our improved version of the Doctrine of Double-Effect: "A person may licitly perform an action that they foresee will produce a good and a bad effect provided that four conditions are verified at one and the same time:

1. A good overall consequence is intended;
2. The bad consequence is merely a foreseen consequence of the intended action;
3. The bad consequence is not a necessary means to bringing about the good effect; and
4. The good effects outweigh the bad effects."

and Joshua Greene's refined formulation of the personal/impersonal principle:

(c) harms caused by a combination of personal force (direct muscle impact, as in pushing someone with a pole) and closeness are worse than harms caused by a combination of impersonal force (as in dropping someone through a switch-operated trap door) and spatial distance. (Greene, 2014, p. 23) [150]

And not only are the explicit rules sufficient to learn a language. According to one of our Poverty of Stimulus arguments, the Logical Problem of Language Acquisition, either explicit instructions or negative evidence are *necessary* to learn languages by induction.

If we follow these arguments and want to find out whether children have a chance to learn behavior that follows one of the three principles above (in the case above, the Doctrine of Double-Effect), we need to examine whether, in normal environments, children explicitly hear the principles above (here: The parts of the Doctrine of Double-Effect). If we want to know whether they can learn that a behavior follows principles at all (according to Dwyer 2006), we need to find out whether children normally encounter instructions regulating behavior with which the aforementioned principles are concerned (e.g. "don't push her between those two fighting persons to make them stop hurting each other", hence,

[149] Or a similar but more complex principle, as we shall see later.

[150] Dwyer et al.'s original formulation was the following: "(c) harms that rely on physical contact are worse than equivalent harms, that are brought about by a nonhuman causal intermediary." (Dwyer et al., 2009, p. 497). I chose Green's formulation because he scrutinized the features of physical contact that make the difference for Trolley Dilemma judgements.

“don’t use her as a means to do good”) or even the principles themselves (“you are only allowed to do something with a good and a bad effect if...”).

Those are empirical questions, and I will examine literature about them. I will start with the Doctrine of Double-Effect, because it has a special status: Most participants of Trolley Dilemma studies explicitly stated that Principles (a) (action/omission) [151] and (c) (personal-close-up/impersonal-distant) [152] (both: Cushman et al., 2006, p. 1086) were the reasons why they found only some sacrifices in Trolley Dilemmas permissible. Harman (Harman 2011, p. 18) argues that in the case of the Doctrine of Double-Effect, however, the learner cannot receive *any* explicit instructions or principles as input because most people are not aware of this principle.

In the paper ”A Dissociation Between Moral Judgments and Justifications” (Cushman et al., 2007) only 3 of 23 participants [153] came up with a response that explained why they had judged the “Oscar” scenario but not the “Ned” scenario permissible: In their justification, they were the only participants that ”correctly identified any factual difference between the two scenarios and claimed the difference to be the basis of moral judgment.” (Cushman et al., 2007, p. 13).

This means that only three had identified that in the “Ned” case, sacrificing a man was a means and in the “Oscar” case, it was a foreseen side-effect to saving five. Thus, Cushman et al. conclude that ”under the conditions employed, intuition drives subjects’ judgments, and with little or no conscious access to the principles that distinguish between particular moral dilemmas.” (Cushman et al. 2007, p. 17). ”Though we sometimes deliver moral judgments based on consciously accessed principles, often we fail to account for our judgments. When we fail, it appears that operative, but not expressed, principles drive our moral judgments.” (Cushman et al., 2007, p. 18).

Mikhail refers to his own research (Mikhail, 2006, 2002) about the same topic [154] when

[151] ”In the case of the action principle, a large majority of subjects were able to provide sufficient justifications for their judgments, whereas relatively few provided failed justifications, denied any moral difference between the scenarios, or expressly doubted their ability to justify their responses.” (Cushman et al., 2006, p. 1086)

[152] ”Subjects’ justifications of their responses to contact-principle cases occupied an intermediate position between justifications for action-principle and intention-principle cases. Subjects were typically able to articulate the relevant principle used, but were relatively unwilling to endorse it as morally valid.” (Cushman et al., 2006, p. 1086)

[153] There are a few reasons why this pool of participants is so small although the experiment was part of Cushman et al.’s large online study with 5000 subjects (Cushman et al. 2007, p. 5): Only 5.8% of people who judged both scenarios in the same session had judged one permissible and the other impermissible. This is why they used participants that had only been presented with either the “Ned” or the “Oscar” case in their first session and who had judged the “Ned” case impermissible or the “Oscar” case permissible. They re-contacted them and presented them with the scenario they had not judged yet. Of those 207 who participated in the ”second round”, on average 20 weeks after the ”first round”, 33% judged the “Oscar” case permissible and the “Ned” case impermissible. Of those 68 subjects’ responses on why they judged that way, 45 were either blank or included ”added assumptions” such as ”the five men will be able to hear the train approaching and escape in time”. This makes for the 23 remaining ones (Cushman et al., 2007, p. 14/15).

[154] In Experiment 4 in (Mikhail, 2011, p. 336) and Experiment 4 in (Mikhail, 2002, p. 48), John Mikhail

he states that

> "[i]n the case of trolley problems, for example, children must represent and evaluate these novel fact patterns in terms of properties like ends, means, side effects, and prima facie wrongs such as battery, even where the stimulus contains no evidence of these properties [...]. These concepts and the principles which underlie them are as far removed from experience as the hierarchical tree structures and recursive rules of linguistic grammars. It is implausible to think they are acquired by means of explicit verbal instruction or examples in the child's environment". (Mikhail 2008, p. 355; Mikhail cites Harman, 2000a; Mikhail, 2000 for this paragraph).

As Chandra Sripada notes in his reply to Mikhail, [155] we could take the last sentence either as statement about the properties Mikhail takes to be the preconditions for the moral judgement pattern (hence, about the ability to have concepts of ends, means etc.) or as a statement about the principles themselves that, according to him, underlie them. Harman, however, is less ambiguous about the non-learnability of some moral principles:

> "[...] whether the child has sufficient learning resources for acquiring certain moral principles may depend on whether those principles are explicitly invoked by others or are merely implicit in the judgments of others. Comparable linguistic principles are highly arcane and are not explicitly known to speakers of the language. The same might be true of relevant moral principles, if, for example, those principles are like double effect in not being something that ordinary people are aware of and able to teach to children acquiring a morality." (Harman, 2008, p. 347)

Harman mentions the Doctrine of Double-Effect research I have summarized above before he concludes that "it is unlikely that children require explicit instruction in order to acquire a first morality, any more than they require explicit instruction in order to acquire

describes how he conducted the "Ned"/"Oscar" experiment with 309 adult volunteers aged 18-35 from the MIT community without asking for justifications; 49 of them gave some kind of verbalized justification anyway. 30 of them were logically adequate and 19 were logically inadequate. 48% of the ones presented with the "Ned" (no heavy object) scenario, and 62% of the ones presented with the "Oscar" (with heavy object) scenario judged them permissible. The difference was significant. He, however, used a between-subject design: Each person was only presented with one scenario. Before, he had tested other scenarios in a within-subject design (Experiment 2, Mikhail, 2002, p. 25) and, in this experiment, asked people for a justification of their judgements. He resumes that the judgement pattern in those cases is explainable by the Doctrine of Double-Effect as well; it is, however, also explainable by principles such as the Personal-Impersonal distinction (Greene et al., 2009). I therefore (unlike John Mikhail) do not want to take the results from this experiment as evidence that people judge according to an explicit version of the Doctrine of Double-Effect.

[155] "The second somewhat stronger innateness claim is that in addition to the constituent concepts, the content of the Doctrine of Double-Effect, that is, the actual proposition expressed by the doctrine, is innate as well. It is unclear whether Harman and Mikhail endorse both these claims or just the former, but I'll address them both." (Sripada, 2008, p. 365)

language, although interaction with others may be necessary in both cases." (Harman, 2008, p. 348). He later states that "It is unclear how such principles might be learned; one possibility is that they are built in ahead of time, perhaps in a 'moral faculty'." (Harman, 2008, p. 349). So the argument, made more explicit by Harman in 2011, goes like this:

> "[T]hese principles are not generally recognized and it is therefore unlikely that they are explicitly taught to children. If children do acquire moralities containing such principles, it would seem that they must be somehow predisposed to do so. If so, we might expect such principles to be found in all natural moralities that children naturally acquire." (Harman 2011, p. 18)

And here is the entire line of thought condensed, although formulated carefully with many subjunctive verbs:

> "[...] certain underlying norms may be implicit in people's moral judgments without themselves being explicitly known to the people whose moralities reflect those norms, just as there are underlying linguistic principles that are not known to speakers of a given language. For example, some theorists have suggested that our moral judgments reflect an implicit acceptance of the principle of double effect. Supposing there is widespread implicit acceptance of such a principle, it would seem that ordinary people have no explicit knowledge of it, and it would appear not to be transmitted by explicit instruction. If there is no other obvious way for the principle to be learned, the hypothesis suggests itself that the principle is somehow innate and should be universal." (Harman, 2008, p. 346)

In those two text passages, when he talks about moral principles, Harman refers to the Doctrine of Double-Effect and to the Deflection of Harm Principle. [156] Based on the citations above we might interpret Mikhail and can certainly interpret Harman as having stated the following:

1. The Doctrine of Double-Effect (and, in Harman's case, the Deflection of Harm Principle), is a principle that people are not aware of.

2. Therefore, it cannot be taught explicitly.

3. Therefore, if children acquire moralities that are generated according to the Doctrine of Double-Effect, they must be somehow predisposed to do so. We can with good

[156] "It is better to save a person X by deflecting a harmful process onto another person Y than by originating a process that harms Y." (Harman, 2011, p. 18)

> conscience read this as a claim that there are genetically 'built in', hence inheritable, tendencies to judge according to principles such as the Doctrine of Double-Effect. [157]

I doubt whether people who cannot account for their intuitions in Trolley Dilemmas are not able to teach those principles explicitly at all (whether 1 really entails 2). Even if they cannot explicitly state the principles that caused their judgements in Trolley Dilemma cases and suchlike, they might be able to state them in different (e.g. less artificial) cases. Maybe we cannot generalize from the Trolley Dilemma research to everyday situations.

Or they might know they hold a principle like the Doctrine of Double-Effect and know to which cases to apply it (its application conditions), but not be able to state that they have applied it when they made an intuitive permissibility judgement:

They could in principle be able to give you the rule including its application conditions, but not be aware of it every time they apply it: They might have stopped applying it consciously and internalized it so much they by now just apply it (similarly to grammatical rules in a second language we have learnt, for instance, at school). This does not mean they have forgotten the original rule or are not applying the 'textbook' rule to other situations.

But even if people who cannot account for the reasons or causes for their judgements in Trolley Dilemmas cannot teach the principles that would lead to the respective judgements explicitly, this may not be the end of the road for criticism of LPLA arguments: I think there might be ways to learn moral principles without explicit instruction that are not subject to LPLA critique. I will come to this later (Chapter 7, p. 200)

However, concerning the lack of explicit instruction, instead of speculating we can empirically test whether there actually are explicit formulations of the principle and whether the degree to which people have been exposed to them might influence their moral judgements: Do people normally encounter explicit formulations of the principles we are talking about, and do those influence the way they judge? If they do, this is evidence that people learn those principles from explicit formulations that they hear in their environment.

If exposure to explicit versions of the Doctrine of Double-Effect has no impact on the degree people explicitly endorse it or the rate they judge according to it, this could either mean that we do not learn those principles at all and that, instead, they are inheritable (and explicit exposure would not even help to learn them) or else that we learn those principles in a non-explicit way and the Logical Problem of Language Acquisition does either not exist or is not extendible to the moral realm (or, at least, to the Doctrine of

[157] This is the claim that the linguists make about grammatical features and the linguistic analogists parallel those features with the principles underlying moral judgements, as e.g. John Mikhail writes: "Thus far I have [...] defended the existence of a 'Universal Moral Grammar' analogous to the linguist's notion of 'Universal Grammar' (UG), that is, an innate function or morality acquisition device [...]"(Mikhail, 2008, p. 355); Chandra Sripada reads both Mikhail and Harman as endorsing nativist claims about the Doctrine of Double-Effect ("Thus, here again, the example of sophisticated innate structure proposed by Harman and Mikhail [...]"(Sripada 2008, p. 365)); the term "moral faculty" has traditionally been connected to nativist claims (see, for instance, abstract of "There is no moral faculty" (Johnson, 2011)).

Double-Effect).

The two significant personal background factors that correlated with the likeliness to distinguish between the “Ned”/“Oscar” cases according to Cushman et al. were

1. age, which significantly predicted whether someone would judge the “Oscar” scenario permissible and the “Oscar” scenario impermissible. Age, however, ”accounted for only 1.4% of the variance” (Cushman et al., 2007, p. 11). Although it is easily arguable that the older people are, the more likely they have heard a version of the Doctrine of Double-Effect explicitly, I will dismiss this predictor because of the small effect size.

2. religion: 5.6% of Catholics, 2.0% of Protestants and 7.2% of Atheists judged the “Oscar” scenario permissible AND the “Ned” scenario impermissible (Cushman et al., 2007, p. 11).

6.2.4.2.1 Catholics and the Doctrine of Double-Effect: Not innate, but indoctrinated? Remember, Cushman et al. designed the “Ned”/“Oscar” cases to see whether people decided in accordance with the Doctrine of Double-Effect. Did more Catholics judge the cases that did not fulfil the Doctrine of Double-Effect conditions as impermissible because Catholics explicitly teach different principles than Protestants, including the Doctrine of Double-Effect? If they do, it is not too hard to argue why Catholics may be more prone than Protestants to stick to the Doctrine of Double-Effect as a principle without having to resort to inherited genetic knowledge/information.

Endorsing the Doctrine of Double-Effect follows a strong Catholic tradition [158] that starts with Thomas of Aquinas (Aquinas, 1265, p. 41,42/ II–II, q. 64, a. 7.).

The Doctrine of Double-Effect is often mentioned as a ”Catholic theory” (e.g. citing from a book about the familiarity of Protestant ethicists with Catholic ethical tradition: ”Paul Ramsey, another leading Protestant ethicist, in his 1961 book on war depended heavily on the Catholic notion of double effect[...].” (Curran, 2008, p. 146/147), and Protestant theoretician James M. Gustafson writes the following:

> ”Fifth, the traditional Catholic arguments are rationalistic. [...] One often finds brief assertions of ”fundamental truths” which include definitions of terms used in these truths or in subsequent arguments. This might be followed by ”basic principles” which will include distinctions between the kinds of law, principles pertaining to conscience, principles of action, a definition of the principle of double effect, and others.[...] But the rationalistic character of the arguments

[158] Mangan writes: ”[...] the principle is perfectly valid and justifiable by reason and Catholic tradition.”, and lists six theological texts from 1755 to 1944 as references (Mangan, 1949, p. 41) and, later: ”Then, as moral theology gradually developed, the principle took a more and more prominent place, until today in all the manuals of moral theology we find a special section devoted to it.” (Mangan, 1949, p. 42).

> seems to reduce spiritual and personal individuality to abstract cases. The learning from historical experiences with their personal nuances seems to be squeezed out of the timeless abstractions. The sense of human compassion for suffering and the profound tragedy which is built into any situation in which the taking of life is morally plausible are gone. Individual instances must be typified in order to find what rubric they come under in the manual. While it is eminently clear that any discussion must abstract facts and principles from the vitality and complexity of live-experience, the degree of abstraction and the deductive reasoning of the traditional Catholic arguments remove the issues far from life. The alternative is not to wallow in feeling and visceral responses, nor is it to assume that one's deep involvement with the persons in a situation and one's awareness of / the inexorable concreteness of their lives are sufficient to resolve the issues. But an approach which is more personal and experientially oriented is another possibility." (Gustafson, 1998, p. 602)

I have chosen such a long quote to show that Gustafson considers it a decidedly Catholic approach to ethics to have basic principles and explicitly mentions the Doctrine of Double-Effect among them.

This is evidence that at least some Christian theoreticians see the Doctrine of Double-Effect (or basic moral principles at all, for that matter) as an inherently Catholic principle. It is canonical for Catholic, but not for Protestant ethics. Although the Protestant Kant famously upheld the principle that you should never treat a person merely as a means (Kant, 1957, p. 53/ AA IV, 429), we can therefore assume that most Protestants did not hear as much about the Doctrine of Double-Effect as most Catholics did. This would explain why 5.6% of Catholics' judgements but only 2.0% of Protestants' judgements conformed with the Doctrine of Double-Effect: They might have heard the principle in its explicit formulation more often and therefore been able to learn it, avoiding the Logical Problem of Language Acquisition.

This, however, does not explain why 7.2% of Atheists judged the "Oscar" scenario permissible but the "Ned" scenario impermissible. There are different explanations for this fact, e.g. that many Atheists have been raised in a Catholic tradition and might hence have heard of the Doctrine of Double-Effect (but this would not account for the higher number of judgements according to the Doctrine of Double-Effect in Atheists compared to Catholics) or that Protestants have heard of the Doctrine but are deliberately not following it. A study on converts in the Netherlands makes the first explanation even more unlikely: In a review of studies from between 1966 and 2003, the authors found that 61.1% of former Protestants [159] had shifted to 'no religion', whereas only 29.7% of former Catholics had done the same. Given that there were only slightly less Protestants

[159] I have, possibly illegitimately, added followers of the Protestant Church in the Netherlands and other Protestants and made them the total number of Protestants.

before this shift and assuming that people mostly adapt to their parents' religions first, this means that among Atheists in the Netherlands were more people with Protestant upbringing than with Catholic upbringing (Need and Graaf, 2005, p. 295).

This is confirmed by a study by the PEW Forum on Religion and Public Life relating to US inhabitants who found that 7% of people raised Protestant, but only 4% of people raised Catholic are now unaffiliated (Street et al. 2009, p. 5), with a slightly higher absolute number of former Catholics that, however, could not account for the difference in Trolley Dilemma judgements compared to Protestants. Interestingly, however, roughly 70% of all people raised Catholic in this study stated that they had visited a Sunday school, hence about 70% had an intense Catholic education that might have implied learning the Doctrine of Double-Effect (Street et al. 2009, p. 5).

I have not found any satisfactory explanations for the correlation between confession and Trolley Dilemma judgements in terms of accordance to the Doctrine of Double-Effect. It seems, however, hard for me to present any explanations apart from ones that take into account the religious tradition people were raised in . To really test this hypothesis, I suggest more experimental work explicitly investigating familiarity with the Doctrine of Double-Effect and correlating it to judgements in Trolley cases with Loop vs. Switch structure.

A study by Julia Christensen et al. seems to confirm the relation between religion and moral judgements in Trolley Dilemmas, and specifically so regarding the Close Contact Harm Principle and the Doctrine of Double-Effect. It compares Roman Catholics to Atheists using Close Contact Harm and a few Doctrine of Double-Effect Trolley Type Dilemmas and found that Atheists, on average, judged significantly more harmful actions "appropriate" in comparison to Catholics in all dilemmas.

This trend was even more significant when it came to what they called "Personal" harm, hence, harm that was inflicted with the agent close to and in physical contact with the victim (Christensen et al. 2014, p. 244). Hence, Roman Catholics judged less "Impersonal" harm scenarios "appropriate" in comparison to Atheists, and the gap is even wider for "Personal" harm scenarios. Put differently, although less Roman Catholics than Atheists tended to judge all kinds of harm appropriate overall, the ratio of Roman Catholics to Atheists who found "Personal" harm appropriate was smaller than the ratio of Roman Catholics to Atheists who found "Impersonal" harm appropriate: The tendency for Roman Catholics to find "Personal" harm more inappropriate than Atheists was even stronger than the former's tendency to find "Impersonal" harm more inappropriate than Atheists.

The Doctrine of Double-Effect had, in combination with Personal force, an impact on Roman Catholics', but not Atheists', reaction times (Christensen et al. 2014, Supplementary Online Material, p. 14). The authors have a paper under review with more behavioral results, likely including more data about the Doctrine of Double-Effect dilemmas, but it does not seem to have been published yet.

All the results from the Christensen paper, paired with their fMRI data (the latter only refer to the "personal/impersonal harm" distinction) which suggest that Roman Catholics processed those dilemmas in a manner quite unlike Atheists, could be read as evidence that the Doctrine of Double-Effect and the Close Contact Harm Principle play a larger role for Roman Catholics than for Atheists and generally, harming someone to save others is less permissible to Roman Catholics than to Atheists.

The authors explain the effects as follows:

> "Our results indicate that moral judgment can be influenced by an acquired set of norms and conventions transmitted through religious indoctrination and practice. Catholic individuals may hold enhanced awareness of the incommensurability between two unequivocal doctrines of the Catholic belief set, triggered explicitly in a moral dilemma: help and care in all circumstances - but thou shalt not kill." (Christensen et al. 2014, p. 240)

Hence, the authors of our study think that the principles (they call them norms) that fit their subjects' judgements were acquired socially, "through religious indoctrination and practice". I look forward to the publication of the behavioral data that should give us even more evidence to answer whether being exposed to Catholic principles changes the way we judge moral dilemmas and hence might be one of the factors that lead to our pattern.

We have learned in the previous paragraph that Catholics might actually be exposed to explicit versions of the Doctrine of Double-Effect more than others (e.g. in religious coursework) and this might be the reason why Catholics judge according to the Doctrine more than Protestants do, but it cannot explain why even more Atheists judge according to it in our first study. The second study, however, is far more conclusive: It shows that Roman Catholics and Atheists make different judgements (and may engage in a different judgement process) and the authors think that this might be due to their religious socialization. Remember, at the beginning of this argument we assumed that you need to be exposed to explicit versions of the Doctrine of Double-Effect to learn it. I have shown that this, at least with Catholics, could well be the case and that it might at least have an impact on the degree to which people judge according to it, hence that people learn and use the principle if they come across explicit versions of the Doctrine of Double-Effect.

We have seen that at least some people (Catholic scholars) have the opportunity to read those explicit versions of the Doctrine of Double-Effect. Sentence 2 of this argument, and Sentence 3 as well, do not apply to those people, at least. The argument that people cannot have acquired the principle explicitly is not true, because at least some may have acquired it explicitly and the fact that some of those who might have been exposed to it judged according to it is evidence that they did acquire it explicitly. This means that, at least in those cases, we do not require anything inheritable or pre-given to explain why children learned those principles.

The question remains whether the baseline rest (i.e. the Protestants in the first study and the Atheists in the second one who judged without violating the Doctrine of Double-Effect) acquired the principle explicitly too, and hence did not need any congenital additional constraints or negative evidence to avoid acquiring an over-generalized rule. Alternatively, a homogeneous part of the entire population had some innate principles that needed triggers to develop and then encountered these and beyond that, the additional fraction of Catholics (and Atheists in the first study) who judged in accordance with the Doctrine of Double-Effect learned the principle explicitly in addition, or more of them were exposed to those triggers. And why does the broad rest of the population not judge according to the Doctrine?

If the Doctrine of Double-Effect was inheritable and unlearnable in a normal environment, either all populations should behave according to it to the same degree or it should depend on the environmental triggers they came across. It is, however, very unlikely that Catholics encounter those triggers more often than Protestants or Atheists if the triggers are not connected with their religious education. Of course, the triggers might be something correlated to Catholicism and Atheism. Put bluntly, I have no idea what that might be.

To further test whether exposure to explicit versions of the Doctrine of Double-Effect leads to judging Trolley Dilemma as morally worse or less permissible when someone is sacrificed as a means, we could assess whether experiments show a correlation between philosophical expertise or familiarity with the Doctrine of Double-Effect and negative judgements in those cases.

Cushman et al.'s 2007 study seems to disconfirm this: People who had exposure to "formal moral philosophical coursework" did not judge Trolley cases differently overall when compared to people without moral philosophical pre-knowledge in their study (Cushman et al., 2007): People who had done "formal philosophical coursework" judged just as many Doctrine of Double-Effect cases permissible as other subjects. So people who might have heard explicit versions of the Doctrine of Double-Effect during their studies because of their experience with moral philosophy did not judge significantly differently from people who had not attended moral philosophical courses.

But does "exposure to formal moral philosophical coursework" constitute enough expertise to show a significant difference in judgements? This especially in view of such a small effect as the difference between the "Ned"/"Oscar" cases was in the same study (only 5.8% of all subjects judged the cases differently, hence "Ned" impermissible and "Oscar" permissible within a single session (Cushman et al., 2007, p. 14)). Maybe so few of the people who had done formal philosophical coursework had come across the Doctrine of Double-Effect that the effect on their judgements was not significant?

Schwitzgebel et al. have collected data on judgements made by philosophers, ethics PhDs and non-philosophers in Trolley Dilemmas to research order effects (Schwitzgebel and Cushman, 2012).

If only exposure to explicit versions of the Doctrine of Double-Effect makes it possible for people to learn it, we would predict that with growing expertise, people's likeliness to judge in line with the Doctrine of Double-Effect in Trolley Dilemma cases would increase with their exposure to ethics theories because while studying ethics, they would likely encounter explicit Doctrine of Double-Effect versions, learn them and apply them.

Schwitzgebel et al. have until now mainly published data on judgements made by expert and less-expert persons of cases with Push vs. Switch structure (e.g. a Trolley Footbridge and a Trolley Switch scenario). [160] The result was that non-philosopher academics and Ethics PhDs made very similar judgements, with a tendency of the latter to rate both scenario types equivalently more often: In those cases, they gave both dilemma types the same ratings on a 7-point scale, e.g. "Pushing the one person is Extremely Morally Bad" and "Flipping the switch is Extremely Morally Bad" (Schwitzgebel & Cushman 2013). [161] This means that Ethics PhDs who should have encountered the Doctrine of Double-Effect more often do not judge according to it more often. And this is not a question of significance either as, if there is a tendency at all, they seem to conform to the Doctrine less often than non-philosopher academics.

If we accept the presupposition that many more Ethics PhDs know the Doctrine of Double-Effect than non-philosopher academics, this could mean either of the following two things:

[160] The wording of the Footbridge Scenario in Schwitzgebel & Cushman was: "Jane is standing on a footbridge over the railroad tracks when she notices an empty boxcar rolling out of control. It is moving so fast that anyone it hits will die. Ahead on the track are five people. There is a person standing near Jane on the footbridge, and he weighs enough that the boxcar would slow down if it hit him. (Jane does not weigh enough to slow down the boxcar.) The footbridge spans the main track. If Jane does nothing, the boxcar will hit the five people on the track. If Jane pushes the one person, that one person will fall onto the track, where the boxcar will hit the one person, stop because of the one person, and not hit the five people on the track. Pushing the one person is:" and a scale with points from 1-7 representing the value judgements from "Extremely Morally Good" (1) over "Neither Good Nor Bad" (4) to "Extremely Morally Bad" (7).

[161] Subjects who found it morally better to push the person than to flip the switch were excluded from the analysis by the authors (Schwitzgebel and Cushman, 2012, p. 140). I calculated these data without considering order effects because the scenarios were presented in a random order. I wanted to annihilate the order effects to make the study more similar to the original Cushman Trolley Dilemma study where scenarios were presented in a random order as well (Cushman et al., 2007, p. 5). This should make both studies more comparable, regardless of the different wording (permissible vs. morally good) and measure (7 point scale vs. permissible/impermissible). For effects of scales vs. binary choices in Trolley Dilemmas, see (Kaufmann, 2015). Schwitzgebel et al. presented the data split up: They presented how many people with different degrees of philosophical education gave the two dilemmas equivalent ratings when a Push-Type case was presented first and how many when a Switch Type case was presented first (Schwitzgebel and Cushman, 2012, p. 142). I merged those data as follows: I added the percentage of people who rated equivalently in the Push-first case to the percentage of people who rated equivalently in the Switch-first case and divided them by two, each for non-philosopher academics and Ethics PhDs independently. As the cases were presented randomly and only two cases at a time (meaning, either Switch first or Push first), these should be the averages for the respective groups.

1. Everyone has an ”inheritable” structure that makes it possible for them to learn the Doctrine of Double-Effect (including people who never heard it explicitly) but not everyone subsequently judges according to it (which may be due to the competence-performance distinction (see Chapter 5, p. 105) or other reasons (see beginning of this chapter)). Therefore, it is not necessary to be familiar with explicit versions of the Doctrine of Double-Effect and knowing them does not change anything about your abilities to judge according to these.

2. The presuppositions of the Poverty of Stimulus proponents are faulty and there are other ways to acquire those principles apart from exposure to explicit ('formal') versions of the Doctrine of Double-Effect: For example, imitating people who act according to the principle. This would explain why people who had exposure to explicit versions of the Doctrine of Double-Effect do not judge according to it more often than other people: Everyone would have had the same kind of 'training' and already 'cognized' the Doctrine of Double-Effect principle before.

Of course, many other factors could explain these results, too, e.g. that philosophers (including Ethics PhDs) would be more prone to apply the Doctrine of Double-Effect but, because they are more motivated to make consistent decisions than non-philosophers, annihilate this effect by rating both dilemma types equivalently more often than non-philosopher subjects in the Push-first cases.

Indeed, philosophers, including Ethics PhDs, ”trended marginally higher than the comparison groups” when it came to order effects in those dilemmas (Schwitzgebel and Cushman, 2012, p. 148), making them more prone to rate Push and Switch Type cases equivalently when they read the Push Type case first. This would annihilate larger numbers of differential ratings in the Switch-first cases, an effect that would go against a bias to rate both cases equally; maybe the philosophical difference between the cases is more apparent with Switch-first; Ethics PhDs, in Schwitzgebel and Cushman's study, tend to make such differential ratings in some cases but not in others, depending on the overall order of presentation. [162]

Schwitzgebel and Cushman write:

> ”However, it is likely that order effects between closely-matched pairs of hypothetical scenarios reflect a general desire to maintain consistency in judgment (see also Lombrozo, 2009). For example, having judged that it is morally bad to push a man in front of a train to save five others, some participants may resist

[162] In the dilemmas that were presented later in the questionnaire (and which, in some cases, were mixed with other dilemma types in between), the Ethics PhD group gave more equivalency ratings (70%) than the non-philosopher academics group. In the Push Type and Switch Type dilemma pair at the beginning, they assigned less equivalent ratings (50%) to the two Push and Switch Type dilemmas than did non-philosopher academics.

> the apparent inconsistency in judging that it is permissible to flip a switch that produces the same consequences. Accordingly, we suggest that order effects arise from an interaction between intuitive judgment and subsequent explicit reasoning: The intuition elicited by the first case becomes the basis for imposed consistency in the second case (Lombrozo, 2009). When the intuition elicited by one case is 'stronger' - that is, more resistant to revision by explicit reasoning - than the intuition elicited by the complementary scenario, this would lead to the asymmetric equivalency effects that we report here. When the stronger case comes first, it would exert a relatively larger influence on the subsequent judgment of the weaker case, making it more likely for the cases to be judged equivalently; but when the weaker case comes first, it would exert a lesser influence on the stronger case, leading to more inequivalent judgments. To take the familiar example of the trolley problem, it has been proposed that the 'push' version engages an automatic, affective response that the 'switch' case does not (Cushman, Young and Greene, 2010; Greene et al., 2001; Greene et al. 2009). This may explain why judgments of the switch case are apparently more malleable under the influence of prior push judgments, whereas push judgments are comparatively stable." (Schwitzgebel and Cushman, 2012, p. 148, emphasis in the original), citing (Greene et al., 2009), (Greene et al., 2001), (Cushman et al., 2010)).

If the philosophically salient feature of a situation is the number of lives you save, and you have rated the first dilemma one way, you would also tend to rate the second dilemma similarly because the same number of lives is saved, an effect that may concern Ethics PhDs more than others (but did not always in the Switch-first presentation order); on the other hand, people had a seven-point scale to choose from and if the Doctrine of Double-Effect mattered to them, they could have at least shifted one point on the scale instead of rating both dilemmas entirely equally, presupposing that they detected the Doctrine of Double-Effect in those dilemmas at all.

Considering these inconclusive effects, those Push Type dilemmas, which according to Greene possibly emotionally charged, [163] in combination with philosopher subjects, might not be ideal to test whether exposure to explicit versions of the Doctrine of Double-Effect makes people act according to it. But there is another, related reason that Push dilemmas are not perfect for testing whether people judge in accordance with the Doctrine of Double-Effect:

One difference between those cases is whether sacrificing a person is a means or a side-effect of saving the five. They, however, also differ in terms of the physical contact

[163] Greene has been challenged for his work (see, for instance, (Kahane, 2012), (Kahane and Shackel, 2010)). Most of this critique, however, does not concern the relative higher fMRI activation of brain areas associated with emotion in Push Type dilemmas compared to Switch Type dilemmas.

involved: In most cases, someone pushed to their death the person who was sacrificed as a means (e.g. from a boat so they drown or from a bridge so they get run over by a trolley). Therefore, it was unclear whether the Close Contact Harm Principle or the Doctrine of Double-Effect Principle was responsible for the pattern of answers and it might be more interesting to have a look at experts' assessments of Loop Type cases (similar in structure to the "Ned" scenario) which test for the Doctrine of Double-Effect without testing for the Close Contact Harm Principle, hence without hands-on contact and the act of dropping someone onto the rails. [164] Schwitzgebel & Cushman tested some of them, but have not published the results in a paper but only in a very short entry on Schwitzgebel's blog which states that philosophers rate Loop scenarios in a similar manner to non-philosophers (83% of them assign them equal ratings). On the website, he does not make any remarks about the effects of growing expertise. [165] After I sent him a request by e-mail, Eric Schwitzgebel was kind enough to run a few analyses which yielded the following results:

Philosophers and even Ethics PhDs did not rate Loop cases significantly worse than Switch cases in comparison to non-philosophers and non-Ethics PhDs.

Even expertise as strong as doing one's PhD in an Ethics-related field did not lead to a more negative rating of Loop cases in comparison to Switch cases. [166] Hence, this study, too, replicates Cushman et al.'s study.

Another 2015 paper by Schwitzgebel and Cushman (Schwitzgebel and Cushman 2015) is even more interesting because they use a Trolley Dilemma called "Drop case" in addition to the "Push case" in which the sacrificed person does not get pushed from the footbridge, but instead a switch-operated trapdoor opens below him and he falls in front of the train. As the Loop case structure, this structure has the advantage of not differing from the Switch Dilemma in terms of physical contact. Hence, the Personal/Closeness principle cannot account for differences in judgements. This makes the Doctrine of Double-Effect the most obvious difference between the dilemmas (although, unlike in the Loop case, in the Drop case the sacrificed person is moved whereas in the Switch case, the train is moved by throwing the switch and in terms of dropping someone on the trails, Drop Type cases are possibly perceived as more cruel than Loop Type cases).

The paper shows a trend for philosophers to give less equivalent ratings to Push/Drop cases and Switch cases (a smaller number of them assigning the same number on the seven-point scale to Push as compared to Switch respective to Drop as compared to Switch)

[164] Although the results of, for instance, Loop Type cases can be attributed to different effects than the Doctrine of Double-Effect, too (see Waldmann and Wiegmann, 2010).

[165] "How about Switch vs. Loop? Again, we found no difference in equivalency ratings between philosophers and non-philosophers: 83% of both groups rated the scenarios equivalently (Z = 0.0, p = .98)." (Schwitzgebel, 2013; bold type in the original)

[166] Only 27 out of 264 (10%) of philosophers rated Loop Type cases worse than Switch Type cases, vs. 210 out of 1801 (12%) non-philosophers (p = .50). And 10 out of 73 (14%) of Ethics PhDs vs. 227 of 1992 (11%) of non-ethics PhDs (p = .54) rated Loop Type cases worse than Switch Type cases (Schwitzgebel, 2015).

with growing expertise. Formulated differently: The more Trolley Dilemma expertise someone has, the more likely they are to differentiate between a Push case and a Switch case or a Drop case and a Switch case. This difference might be because the subject judges according to the Doctrine of Double-Effect (and more certainly so in the Drop/Switch case). But:

This only applies if the Switch case is presented first (Figure 4 in Schwitzgebel and Cushman, 2015, p. 135). This might also be why "philosophers reporting familiarity, expertise, stability, and specialization in ethics trended toward showing *larger* order effects than the remaining philosophers." (Schwitzgebel and Cushman, 2015, p. 134, emphasis by the authors). And an even greater caveat: The paper does not contain any statistical analysis regarding this trend as its research focus are order effects, so we are only talking about trends here. A further analysis is necessary, but the data give a faint hint that people who occupy themselves more with the Doctrine of Double-Effect for professional reasons might tend to differentiate more between Trolley Dilemmas if the Switch Type case is presented first.

In the light of the uncertainty and unpredictability of those results, I would suggest more research on whether people who are likely to have encountered explicit formulations of the Doctrine of Double-Effect and Trolley Dilemma problems (in the 2015 study and according to themselves: 77% of all philosophers (Schwitzgebel and Cushman, 2015, p. 134)) in their lives are more likely to judge according to the Doctrine.

All in all, these data are not quite conclusive, but the trend goes towards stronger ordering effects for Trolley Dilemma experts but no large difference in their judgements as compared to non-experts.

This indicates that people who are (or should be) more familiar with the Doctrine of Double-Effect do not automatically endorse it more, neither explicitly nor when they are judging scenarios. [167]

However, there was one significant difference between Ethics PhDs and PhDs in other disciplines: Only 1% of Ethics PhDs judged Switch worse than Loop, whereas 8% of the other PhDs judged Switch worse than Loop (once again, personal correspondence with Eric Schwitzgebel about unpublished (Schwitzgebel and Cushman, 2012) data). Hence, being familiar with the Doctrine might not make people judge in accordance with it, but nonetheless keep them from judging against it. Do we learn constraints that keep us from judging against the Doctrine if we have explicitly encountered it, even if they do not make us judge in accordance with it?

[167] They do, however, uphold principles more in accordance with their arguments (post-hoc rationalization): When asked after they judged two Trolley Type cases, they explicitly agree to the Doctrine of Double-Effect more often when they had judged both dilemma types differently and less often when they had given them the same rating. This shows that, although Ethics PhDs might not have learned to apply the Doctrine of Double-Effect more than others, they have at least learned to uphold principles consistently to their judgements during their studies (Schwitzgebel and Cushman, 2012, p. 147).

The following argument rests on the assumption by linguistic analogy proponents (for morality) and by UMG proponents (for language) that acquiring moral principles/linguistic grammars without explicit teaching and/or negative evidence is impossible, the Logical Problem of Language (or, in our case, morality) Acquisition. Here is a recapitulation of the argument at the beginning of this chapter that expounds the role of explicit teaching for acquiring moral principles, followed by a summary of what we have learned about the exposure to explicit versions of our principles and whether it influences the judgements made by people who have encountered them:

1. The Doctrine of Double-Effect (and, in Harman's case, the Deflection of Harm Principle) is a principle that people are not aware of. Therefore, people cannot explicitly uphold it.

2. Therefore, it cannot be taught explicitly.

3. Therefore, if children acquire moralities that are generated in accordance with the Doctrine of Double-Effect, they must somehow be predisposed to do so. We can with good conscience read this as a claim that there are genetically 'built in', hence inheritable, tendencies (or at least necessary information) to judge according to principles such as the Doctrine of Double-Effect.

We have partially answered three questions:

1. Do at least some people get taught explicit versions of the Doctrine of Double-Effect? This would be enough to make Claim 2 of the argument above fail for at least those people.

2. If people get taught explicit versions of the Doctrine of Double-Effect, do those versions have an impact on whether they learn those principles? Do more people who came across explicit versions of the Doctrine consciously know it and do they know how to apply it? If the answer to the latter two questions is negative, this would raise the further question of whether this is because the inherited preconditions for the principle are sufficient for learning exactly this principle with non-explicit input or whether we can acquire those principles in an alternative way.

3. And does explicitly knowing those principles also mean that people uphold and apply them? They could, for example, be able to explicitly recite them, know how to apply them but not be inclined to apply them. This would lead us to further questions.

I will answer those questions in the order I asked them.

1. Is it possible for people to come across explicit versions of the Doctrine of Double-Effect? Do some people get to learn the Doctrine of Double-Effect in its explicit form? Do some people teach the Doctrine of Double-Effect explicitly?

Well, the obvious and simple answer is: Yes, for example in this text. But what matters, of course, is not whether people who have specialized in scholarship about the Doctrine of Double-Effect can find explicit formulations if they look for them or whether people specialized in research about morality might stumble upon them. The question is whether people come across this principle 'normally', whether adult persons are likely to have encountered it in their environment. In the beginning of this chapter, I presented a few examples of texts about Catholicism in which the Doctrine of Double-Effect was treated as one of the canonical principles in Catholicism. As many Catholics attend religion classes in school and thereby might occupy themselves theoretically with those principles, it might well be that a certain percentage has come across them. Further research is, of course, necessary. And in the Schwitzgebel et al. study from 2015, 77% percent of philosophers and 9% of non-philosophers claimed to be familiar with the Doctrine of Double-Effect (Schwitzgebel and Cushman, 2015, p. 134). If this is true, and we presuppose that "claiming to be familiar with" the Doctrine of Double-Effect means having come across explicit versions of it, Premise 1 of the argument would not be true: People would have ways of coming across explicit formulations of the Doctrine of Double-Effect, either in their religion courses or, apart from that, the way (whatever it was) in which 9% of Schwitzgebel and Cushman's subjects have learned about the Doctrine of Double-Effect. So, we can say that the argument is unsound because one of its premises is false.

I do, however, not wish to discard the Poverty of Stimulus argument too hastily. What if people have come across (or had the chance to come across) those explicit versions of the Doctrine, but this did not influence whether they learned it or not or their behavior in general. 9% of 'normal' people have encountered the Doctrine of Double-Effect in its explicit form. Hence, we have opportunities to learn it apart from having inheritable constraints that enable us to learn just the right version of the Doctrine. And those 9% of Schwitzgebel and Cushman's subjects are a big proportion considering the small overall percentage of people who differentiate according to the Doctrine of Double-Effect in their judgements (e.g. 5.8% in Cushman et al. 2017 (Cushman et al., 2007, p. 14)). But do we really acquire it that way? If so, being familiar with the Doctrine should have an impact on knowing it, being able to apply it correctly and applying it. To test this, I want to answer the following Questions 2 and 3:

2. Do people know the principles after they have encountered explicit versions of them? Do they know when they most certainly have applied the Doctrine of Double-Effect? Do they know how to apply the Doctrine of Double-Effect when they claim to be familiar with it?

3. Do people who have come across explicit versions of the Doctrine of Double-Effect tend to judge according to it? Are those people the ones that differentiate between Loop and Switch Type cases and are they the only ones?

The answer to whether people know the Doctrine of Double-Effect after they have encountered explicit versions depends on what “knowing” means. I do not know of any studies where people were asked to recite the Doctrine of Double-Effect, so even if in some studies they claimed to be familiar with it, that does not mean that they were able to consciously access or explicitly cite it. If “knowing” means that we consciously uphold principles and are aware of using them when we judge something, the answer is no: Even most people who judged in accordance with the Doctrine of Double-Effect could not explicitly account for the reason why they did so. [168] Philosophers, however, sometimes seem to be able to realize when they have made judgements in accordance with the Doctrine of Double-Effect: When presented in an order where a larger percentage of subjects differentiated between the Trolley cases, more philosophers tended to explicitly endorse the Doctrine of Double-Effect when they were asked afterwards in the 2012 study [169] (Schwitzgebel and Cushman, 2012, p. 145) whereas, in the 2015 study, paradoxically, non-philosophers did the opposite: More of them explicitly endorsed the Doctrine of Double-Effect in cases where less actually judged according to it (Schwitzgebel and Cushman, 2015, p. 133). And in the 2015 study as well, the order of presentation did not correlate with philosophers’ endorsement of the Doctrine of Double-Effect; hence, the findings of the previous study were not replicated in this respect (Schwitzgebel and Cushman, 2015, p. 133).

So, having expertise seemed to help recognize when you already had judged according to the Doctrine of Double-Effect in the 2012 study, where philosophers were asked about their endorsement of the principles after judging the dilemmas, whereas this effect was not present in the 2015 study with the same question order. Hence, we cannot conclusively answer the question, but evidence rather speaks against people’s (and even experts’) ability to explicitly state or even recognize the principle they have just judged in accordance with.

But what about Question 3: Do people who have come across explicit versions of the Doctrine of Double-Effect tend to judge according to it? And are those people the only ones who differentiate between Loop and Switch Type cases? Is it necessary to encounter explicit versions of the Doctrine of Double-Effect to judge according to it? Remember, the percentage of subjects who gave different ratings to Loop and Switch Type (or Loop-With-Heavy-Object Type in the 2007 study) Dilemmas in the same session were 5.8% in the 2007 Cushman study (Cushman et al., 2007, p. 14) when they were asked about the permissibility of throwing the switch and had a binary choice (”Is it permissible for [...] to throw the switch, with the answer options yes and no (Cushman et al., 2007, p. 6)) [170] and 10-14% in the Schwitzgebel & Cushman 2012 study (personal correspondence

[168] Only 3 of 23 persons who judged the Switch Type case permissible and the Loop Type case impermissible could give a sufficient justification that ”identified any factual difference between the two scenarios and claimed the difference to be the basis of moral judgement.” ((Cushman et al., 2007, p. 15) and (Cushman et al., 2007, p. 13).

[169] Whose available version, unfortunately, only cites conflated data for Push/Drop Type cases, so I cannot really say anything about the Drop Type cases in isolation.

[170] In the same study, ”subjects who had been presented with only one of the scenarios and who had

with Eric Schwitzgebel) where subjects could give the respective sacrificing action ratings on a 7-point scale (from "Extremely Morally Good" to "Extremely Morally Bad"). This means, overall, only very few people differentiate between those two types of cases. Could those be exactly the people that have come across explicit versions of the Doctrine of Double-Effect?

And as I already suggested, having done coursework in moral philosophy did not make a difference in the Cushman et al. 2007 study (Cushman et al., 2007), being an Ethics PhD did not make a difference in the Schwitzgebel and Cushman 2012 study (Schwitzgebel and Cushman, 2012), but Schwitzgebel and Cushman 2015 (Schwitzgebel and Cushman, 2015) showed an inconclusive effect: Philosophers only tended to differentiated between the Switch and Drop Type cases (slightly!) more than non-philosophers when the Switch Type cases were presented first; when the Drop Type cases were presented before the Switch Type cases, this trend was not present. And at the beginning, I presented the percentages of Catholics, Protestants and Atheists who judged according to the Doctrine of Double-Effect; the result was that the group who applied the Doctrine of Double-Effect most were neither Catholics nor Protestants but actually Atheists in the first study. Of course, Atheists might have attended religion courses as well; however, the results are still inconclusive and the group who should, according to my research, have had the most contact with the Doctrine of Double-Effect, namely Catholics, were somewhere in between Protestants (who seldom judged according to the Doctrine of Double-Effect) and Atheists. The second study (Christensen 2012) did not test behavior for the Doctrine of Double-Effect specifically, apart from reaction times, but showed behavioral data that were consistent with Catholics' exposure to moral principles: A lower readiness of Roman Catholics to judge the sacrifice of persons as appropriate overall and a lower tendency to find Close Contact Harm appropriate than their fellow Atheist subjects.

We still lack a study correlating familiarity with the Doctrine of Double-Effect directly with the rate that people differentiate between Switch and Drop Type Dilemmas. Ideally, someone would correlate exposure to (instead of familiarity with) explicit formulations of the Doctrine with Switch and Loop Type Dilemma judgements. The results reviewed above are inconclusive.

This means that people, even if they have encountered explicit formulations of the Doctrine of Double-Effect, probably have not acquired it this way.

To sum up, proponents of the Linguistic Analogy had argued that people do not have access to explicit versions of the Doctrine of Double-Effect and without explicit instructions or negative evidence, no-one can learn the right set of morally permissible situations. They

judged scenario 3 [Loop case scenario] as impermissible or scenario 4 [Loop case with Heavy Object scenario, in terms of causality similar to Switch Type scenario] permissible, and asked [...] to make a judgment on the corresponding case" after 20 weeks on average turned out to rate both dilemmas differently in 33% of all cases; this is the percentage only of subjects who responded after they had been contacted by Cushman et al. to come in for a second session (Cushman et al., 2007, p. 14).

had argued that hence learners must have had innate constraints that enabled them to learn the right set of situations. I have shown that this is not quite correct as many people do encounter explicit versions of the Doctrine of Double-Effect. However, whether this influences their judgements as it should if they learned 'like little scientists' remains unclear.

I have presented the following possible interpretations:

A. The proponents of the Linguistic Analogy are right: Everyone has an "inheritable" structure that makes it possible for them to learn the Doctrine of Double-Effect (including people who have never heard it explicitly) and therefore it is not necessary to be familiar with explicit versions of the Doctrine of Double-Effect to learn it. Therefore, knowing the explicit versions (or being familiar with them) might change something about your abilities to judge according to them, but this only concerns part of the subjects who judged in accordance with the Doctrine of Double-Effect.

B. The Logical Problem of Language Acquisition does not apply (or does not apply to morality) and there are other ways to acquire those principles (making it possible to act according to them) apart from exposure to explicit ('formal') versions of the Doctrine of Double-Effect.

A and B might be possible explanations for what I found in the literature review: The at least partial non-relatedness of explicit knowledge about the Doctrine of Double-Effect (and exposure to written versions of it in the context of a philosophical education) and judgements that are consistent with the Doctrine of Double-Effect.

The second interpretation suggests that the Poverty of Stimulus argument might lay too much weight on explicit formulations (which, of course, is a consequence of the Logical Problem of Language Acquisition; but I believe it can be solved.)

In the light of the chapters before, especially the Universality chapter, I am inclined to adopt Explanation B. I believe that people, regardless of their exposure to the Doctrine of Double-Effect, all tend to judge according to it to similar degree because encountering explicit versions of the Doctrine of Double-Effect is neither sufficient nor necessary to judge according to it. It is not necessary, not because everyone 'has it in them' already, but because people do not need to acquire this principle in the way that the proponents of the Poverty of Stimulus argument claim. I will briefly present alternative ways of acquiring the Doctrine of Double-Effect later(7, p 200): these being (more or less) empiricist paradigms.

And coming across explicit versions of the Doctrine of Double-Effect in order to judge (or act) according to it is not sufficient because the Doctrine of Double-Effect is a moral principle. We not only need to implicitly know [171] it to follow it (meaning, to have

[171] To know it, without necessarily being conscious about it or being able to explicitly account for it; Noam Chomsky calls this kind of knowing "cognizing" and in our case, it is a precondition for people to judge dilemmas systematically according to their propositional content, tacitly "cognized" principles,

acquired the ’formula’ of the correct principle, to recognize the rule that is behind people’s judgement patterns), but we also need to approve of it.

I believe that approving of moral principles has a great deal to do with imitating your surroundings. If you ‘cognize’ a principle, this does not mean you think it is right. My guess is that seeing a principle in application is necessary (and in some cases sufficient) to acquire it and act according to it. And learning explicit versions or arguments in text form might not even contribute to approving a principle or internalizing it: If you come across them in a rather dispassionate form, e.g. a text that only states the principle’s existence and how many people judge according to it (say, a psychological Trolley Dilemma study), this does not encourage you to endorse this principle, neither does it even tell you whether it is morally right or wrong. The case is possibly different if you encounter explicit moral principles paired with their valence in Catholic education (simply because one of this education’s purposes is consolidating or forming the right beliefs in learners). I will write more about possibilities of acquiring our three principles (and an inclination to act according to them) at the end of this chapter (7, p. 200).

But there might be yet another reason why stating moral principles explicitly is not sufficient to teach children morality:

Some proponents of the Linguistic Analogy (see (Dwyer, 2006; Harman, 1999)) argue that explicit instruction is neither needed nor sufficient for learning the actual rules (and not only the fact that there are rules), and they use arguments drawn from the Logical Problem of Language Acquisition and transferred to the realm of Trolley Dilemma principles.

Gilbert Harman, who is quite explicit about his moral nativism, writes:

> ”Although many parents explicitly teach morality to their children, such teaching is probably not needed and may not be *particularly effectives*. Just as children acquire a version of a local dialect simply through being exposed to others who already speak it, without any need for explicit instructions, it is likely that this is true for morality also.” (Harman, 2000b, p. 223; my emphasis)

He does not give any reasons for this except for the statement that everyone who has not suffered certain brain damage or whose upbringing is ”unusually deprived” develops a ”moral sense” (Harman, 2000b, p. 223). We can possibly understand the quotes above as follows: Although it is possible that not everyone gets taught ”morality”, only very few people do not develop it, which shows us that explicit instruction is not needed. Many, but possibly not all children get taught morality explicitly but almost all children develop a “moral sense”.

analogue to the way a “cognized” linguistic grammar allows them to judge the correctness of grammatical sentences and generate new grammatically correct sentences (Chomsky, 1965, pp. 3–9); (Chomsky, 1980, p. 69/70).

In any case, Harman argues that it is probably unnecessary for children to be taught morality explicitly. And it might even be ”not particularly effective”; this is rather vague, but if a critical degree of ineffectiveness is present, that means teaching children principles explicitly is not only unnecessary, but insufficient for them to acquire “morality”. Explicit versions of the principles are neither necessary nor sufficient for acquiring a certain kind of morality (e.g. one including the Doctrine of Double-Effect) because we already have innate information that make it possible to acquire that language without overgeneralizations. If we want to stay within the realm of the Poverty of Stimulus argumentation, however, every explanation that does not refer to innate information has to be excluded to infer that acquiring moralities is only possible because of this innate information. As I mentioned, I will present an alternative proposal as to how learners can acquire morality in the end of this chapter and Harman has neither shown that some of the people who have acquired a “moral sense” were not exposed to explicit teaching, nor has he shown that they were not exposed to negative evidence.

But Susan Dwyer has made another case why explicit teaching would be insufficient for children to learn principles. The scope of this argument is less general; it does not say that all kinds of explicit teaching are insufficient for the learner to acquire a morality: She only states that the explicit teaching that actually occurs is not sufficient. She writes that children normally encounter two kinds of corrections: One is in the form of ”post-hoc evaluations” such as ”You ought not to have broken your sister's train”, and “unexplained imperatives” [172] (Dwyer, 1999, p. 172.) She goes on to argue that post-hoc evaluations ”are evaluations of particular instances of morally assessable behavior, and rarely involve appeal to a general rule. The child knows that she ought not to have broken her sister's train, but what of her brother's slot cars, or her sister's teddy bear?”

Let us assume that in her society, the ”target” rule (the one that is commonly followed) is ”you should not break other peoples' toys, but you are permitted to break your own”.

However, if she had before wrongly hypothesized that it is acceptable to break everyone's toys because she had broken her own and no-one had said anything against that, this instruction would be quite helpful: She would now know that at least in her sister's case, breaking the toy was impermissible. She might then hypothesize a different generalization, possibly the correct one: That it is allowed to break your own toys, but not those of others.

Let me explain this more concisely by starting with a linguistic argument:

6.2.4.2.2 Negative evidence through explicit instructions I have mentioned before (6.2.3, p. 160) that negative evidence, according to the Logical Problem of Language Acquisition, can help us abandon hypotheses that contain over-generalized rules. Remember, if the learner hypothesized that something like “goed” instead of “went” was correct (e.g. that all verbs are built by adding a “-ed” to the stem), someone telling

[172] I will return to those later.

them that “goed” is actually wrong would be necessary for them to abandon this hypothesis. Or, if we apply this thought to whole sentences: Fiona Cowie has argued that ”direct negative evidence - explicit information to the effect that string S is not a sentence of the target language” (Cowie, 1997, p. 24) would be the kind of negative evidence that would keep a learner from over-generalization.

As suggested by Dwyer’s bibliographical reference ”Hornstein, N. and D. Lightfoot, editors. 1981. Explanation in Linguistics: The Logical Problem of Language Acquisition. London: Longmans” (which she, admittedly, only quotes for her principles and parameters section in her 1999 paper ”Moral Competence”, and not for her section about deduction of moral principles) as well as by many of her formulations, she refers to the Logical Problem of Language Acquisition (just as does Cowie) when she writes that ”one might try to explain moral development in terms of the explicit moral instruction children typically receive” (Dwyer, 1999, p. 172), before she continues to dismiss this instruction as insufficient for learning moral principles (see above).

Seven years later, Dwyer writes: “[T]here may well be a paucity of negative concerning the distinction between the two types of rules. Very roughly, negative evidence is evidence that the child can use to correct a false assumption she has made[...].” (Dwyer, 2006, p. 241). [173]

Hence, at least in her 2006 paper, she again adopts a part of the LPLA argument and determines that negative evidence can help learning morality.

The connection between negative evidence and our ”post-hoc evaluations” such as ”You ought not to have broken your sister’s train.” (Dwyer, 1999, p. 172) is the following:

If the explicit instruction that a learner gets is an answer to an over-generalization of the kind mentioned above (that breaking everyone’s toys is permissible), ”You ought not to have broken your sister’s train” might play exactly the role of negative evidence: The over-generalized (moral) hypothesis would be revised. The learner would no longer think that it is acceptable to break all toys and might thereby get closer to her target principle (she might, for example, now think that ”It is permissible to break all toys except for my sister’s teddy bear” or even ”Maybe it is permissible to break my own toys but not permissible to break other people’s toys”). If, however, the learner did not start with an overgeneralization like this and the imperative is the only thing the person had known about the permissibility of breaking things or the concept of property, we are back to the start: Dwyer would be right that the information about this particular instance would in itself not be sufficient to acquire moral principles (like, in our fictional case, the principle

[173] She writes this in a part about the moral/conventional distinction, but seems to think it can apply to a broader set of distinctions because she writes “Second, there can be a paucity concerning the distinction between the two types of rules. Very roughly, negative evidence is evidence that the child can use to correct a false assumption she has made or that she can use (in this case) to eliminate a candidate criterion for making the [moral/conventional] distinction.” (Dwyer, 2006, p. 241, emphasis by me), which suggests that there are *other cases* in which negative evidence can be used as well.

that people are allowed to break their own toys but not others' toys). Several of these chunks of information, however, would be helpful in building new hypotheses.

This returns us again to almost all problems associated with deducing rules from particular instances and, with this, to the Logical Problem of Language Acquisition: Even if a child gets explicit instructions about her behavior, she will still have to deduce from patterns of experiences or explicit corrections the scope of the instruction she has just received: Either she repeatedly sees someone getting angry (which is a hint that it is (morally) impermissible) when other people break their siblings' trains, slot cars, and teddy bears, or she gets more and more instructions about (not) breaking people's things. Here looms (along with the problems of finding out the salient features of the situations that make them impermissible) the danger of impermissibility principles that are too shallow:

Remember, in her society, the 'target' principle is "you should not break other peoples' toys, but you are permitted to break your own".

If she is told that she should not have broken her sister's teddy bear, she might, if we follow Susan Dwyer, think it is forbidden to break her sister's teddy bear (or, if she generalizes, her sister's teddy bears.) If she then is told that she should not have broken her sister's slot cars, she might start thinking that she is not allowed to break any of her sister's toys. But, given this input, nothing keeps her from thinking that she is allowed to break her brother's toys. And, assuming she gets those instructions, she does not know yet that it is considered impermissible in her society to break any things that belong to any other people.

In this case, Susan Dwyer is right: The rules she is given are too narrow. They do not give you any information on whether you are allowed to break everything but your sister's toys. This is too specific. The learner will hardly ever get those kinds of specific, situation-bound instructions referring to everyone and all of their goods in her society. Thus, she will not find out about every instance to which the target principle applies. [174]

It is therefore crucial that she generalizes and finds out the principles that underlie the single-case judgements. And this is something empiricists would agree to. Proponents of the Logical Problem of Language Acquisition, however, will say that this is where the danger of over-generalization appears.

Let us assume, applying the Logical Problem of Language Acquisition, that the learner starts hypothesizing principles that are wider than the ones she has explicit evidence for, although the explicit instructions she gets do not even cover the entire scope of the 'target' principle (the principle that is endorsed in her society). If she, for example, has had the input "You should not have broken your sister's train" and, additionally, "You should not have broken your brother's teddy bear", she might as well deduce the following principle:

[174] And if she did, in the case of post-hoc evaluations, it would be too late and she would already have committed every single offense there is in her society before she knew the entire scope of appropriate judgements. And even then, the instances of forbidden behaviors would still be bound to time and she would have to generalize over time to get to the target principle.

"You should never break anything" or, formulated differently, "The following actions are not permissible: To break your sister's train, your brother's teddy bear, your own toy train", etc., listing all things she knows that are breakable.

And until she has received explicit instructions such as "You are allowed to break your own things" or seen people approve of (or at least not disapprove of) other people breaking their own things OR seen people breaking their own things at all, there is no way she should deduce that the actual rule in her society is "You should not break things that belong to other people".

In this case, negative evidence would be to see people do something and others not disapproving of them or others explicitly stating that what the people are doing is permissible. She could, for example, see people break their own things and other people watch them and not look disappointed, or someone could tell her that she can do whatever she wants to her own toys, or she could simply see someone break their own toys (which would be evidence that it is acceptable behavior in her society). In other words, negative evidence in cases where the scope of the hypothesized rule is too broad would be something that tells the learner that something is not impermissible, thus permissible. Actions that are permissible probably take place more often than actions that are impermissible so that the learner can see people reacting to them. Therefore negative evidence, in this case, probably is not rare. We are learning actions, paired with permissibility judgements, which is exactly what enables us to tell permissible from impermissible actions. To sum this up: If we over-generalize from explicit rules that tell us what we should not have done (the kind of rules Dwyer deems insufficient), our hypotheses will contain too many prohibitions. And counter-evidence to prohibitions is just the prohibited action taking place or, more strongly, the action taking place without surrounding people being alarmed.

I suggest that this part of the Logical Problem of Language Acquisition, raised by Susan Dwyer, is settled for moral cases. Post-hoc evaluations of the kind she criticizes are helpful and might be sufficient to acquire moral principles.

I have shown that, firstly, in case the learner was 'cognizing' some overgeneralized rule like "I am allowed to break everyone's toys" (or had the default assumption that she was allowed to break everyone's toys), explicit post-hoc instructions to the contrary indeed are helpful, which would be an argument against Dwyer's assumption that such instructions are not very helpful.

Secondly, for post-hoc evaluations just as for other cases of rule learning by evaluation of single instances, the learner must generalize and come up with the principles that generate those evaluations. If this is the case, the problem of over-generalization arises just as with normal learning. Then, Susan Dwyer is right that post-hoc instructions are not more helpful than mere observation in learning principles.

But, thirdly, post-hoc evaluations are mostly about the impermissibility of something (we rarely come across post-hoc evaluations as "you could have drunk a glass of water"). If we derive an over-generalized principle from explicit statements about specific impermissible actions, this means that the scope of the impermissibility principles will be too broad. Hence, too many actions would be hypothesized to be impermissible. Negative evidence would then be permissibility statements and observations of permissible actions or (non-verbal) approval of permissible actions by third persons.

As permissible actions are much more frequent than impermissible ones, there should be sufficient negative evidence to save the learner from over-generalization or at least correct her when she over-generalizes. This means that post-hoc evaluations do provide us with useful information to deduce moral rules.

Susan Dwyer, however, has mentioned other kinds of explicit instructions (beyond post-hoc evaluations) that she takes to be insufficient for the child to learn moral principles and I promised to come back to those. They are what she calls "unexplained imperatives" such as "'Keep your promises'" or "'You ought to tell the truth'" (all three citations Dwyer, 1999, p. 172).

She says that these imperatives do "convey more generality, but when we impart them to children we do not ordinarily draw attention to the fact that moral generalizations are *ceteris paribus* generalizations. [...] [W]e do not state and explain the several exceptions. Thus, the moral generalizations we offer to children are fairly coarse-grained, offering only limited guidance to children in their future actions." (Dwyer, 1999, p. 172, emphasis by the author).

While she is very careful in her formulation "offering only limited guidance", we can take this as a challenge and show how we can employ those unexplained imperatives to get rid of over-generalizations: To make the case clearer, we can parallel the principles that we get by unexplained imperatives to the linguistic case where the child thinks that the appropriate principle for past tenses is attaching "-ed" to the end of every verb. Here, too, the child thinks that there is no exception to the rule: They think that forms like "I goed" are correct, too. We could also think of this kind of principle as over-generalized: The exceptions, if stated explicitly, would make the principle more extensive; it would contain more rules. [175] It would then be something like "Form past tenses by adding an -ed to the present tense, except for the words: 'to go', 'to find', ...". Therefore, the number of cases our basic principle is applicable to would become smaller: Less verbs would fall under the scope of "add an -ed to it to make it a past tense". And, in our analogy, less actions would fall under the scope of, for instance, "Keep your promises", as the exceptions would be formulated something like: "Keep your promises, unless you expose someone to danger if you keep them or unless you cannot do so without endangering yourself (or...).". [176] So

[175] I define "principle" as several rules that tell you which action would be permissible in a situation with certain features: The Doctrine of Double-Effect is a principle whereas its parts are rules.

[176] This example could actually be close to a real-life target principle: Thomas Scanlon, for example, writes in his book "Moral Dimensions": "So, for example, according to the principle of fidelity to promises, the fact that one promised to do a certain thing is normally a conclusive reason for doing it - a reason

if you promised to visit your friend the next day, but there is a tornado, [177] you are now officially allowed to not visit your friend whereas without exceptions, the rule would have told you to do so.

The actions that would fall under the scope of the principle, thus the actions that you should do or you are told to do would be "everything you promised" in the case without exceptions and "everything you promised except for those actions that expose someone else or yourself to danger (or...)" in the case with exceptions.

In the case of "Keep your promises" and "You ought to tell the truth", if we understand the imperative "keep" and the formulation "ought to" as statements that something is obligatory, the instructions can be translated into "It is not permissible not to keep your promises" and "It is not permissible not to tell the truth" (J Mikhail, 2007, p. 144). When the number of cases to which they are applicable grows less, as they are prohibitions, the cases that are permissible become more: "It is not permissible not to keep your promises except you expose someone else or yourself to danger (or...)". When we take those "coarse-grained", "unexplained imperatives" without the exceptions at face value, we get over-generalized negative principles: "It is never permissible not to keep your promises" or "it is never permissible not to speak the truth".

Now we are in the same position as in the case with the post-hoc evaluations: If impermissibility judgements lead to over-generalized impermissibility principles, negative evidence comes in the form of observing permissible cases that are wrongly over-generalized by a too-general impermissibility principle, hence cases that are believed to be impermissible but are in fact permissible. And we can assume that those actions are performed and therefore can be observed far more often than actions that the learner judges to be permissible although they are not, because people tend to avoid impermissible actions. (Those latter actions, admittedly, might be conducted by the learner themselves and proven to be impermissible by reactions. This, however, would be the hard way of learning as morally impermissible behavior usually gets socially punished).

I have exercised this transfer from the linguistic input pairs (sentence - correct or incorrect) to the moral input pairs (action - permissible or impermissible) using Susan Dwyer's own examples. She had, at the stage where she undertook very thorough comparisons of moral and linguistic arguments, not yet started working with Trolley Dilemmas. Her arguments, however, are easily applicable to different principles such as the Doctrine of Double-Effect.

that determines what one ought to do even if it would be more convenient or more advantageous to do something else. But there are exceptions. For example, one need not, and indeed should not, fulfil a promise to one person to do something fairly trivial if doing so would cause great harm to someone else. Fully understanding the morality of promising involves being able to recognize the considerations that do, and those that do not, justify such exceptions." (Scanlon, 2009, p. 22).

[177] Ideally assuming that you did not know this would arise when you made the promise, but that would be the result of another rule, something like "Do not make any promises you know you cannot/are not going to keep".

I have shown that explicit instructions in the form of post-hoc evaluations and unexplained imperatives can be very helpful indeed, even if we start with the perspective of people who think that this kind of explicit instructions does not help much in the acquisition of moral principles and hence that we would need explicit formulations or negative evidence to avoid the Logical Problem of Language Acquisition.

If we concede that explicit instructions in the form of post-hoc evaluations or unexplained imperatives in the form Dwyer mentioned can be helpful in learning principles and that the learner can apprehend the specific target principles or narrow down over-generalized principles this way, we have solved a few of the most salient issues that the LPLA raises and Dwyer problematizes especially for the moral realm: We have shown that certain forms of explicit teaching and negative evidence can provide us with important correctives against over-generalizing principles, and, more importantly, these are forms that Dwyer as a LPLA proponent has suggested are found in children's education.

In the next section, I will argue that apart from this explicit (or observed) negative evidence we could find other ways to adjust our hypotheses about moral principles, using just the kind of empiricist learning that is contested by the LPLA.

Susan Dwyer points to another issue that makes it difficult for children to learn morality. It is connected to the problem I have just discussed. She assumes that firstly, children do not get the feedback they need from their caretakers (i.e. 'caregivers') to learn all relevant principles and secondly, children have to extrapolate moral principles in order to generate new judgements/actions that are consistent with the moral code of their community because they will not encounter all instances or even classes of morally relevant situations: "It is also worth noting that there is a vast amount of moral knowledge that is unavailable to children. While children test the limits of the permissible, so to speak, they do not engage in the full range of behaviors that would stimulate their caretakers to impart a good deal of moral information. Similarly, children are not exposed to examples of every possible situation in which a particular judgment is appropriate. And it is not until the adolescent is enrolled in Moral Philosophy 101 that she even ponders whether it is permissible for the authorities to hang an innocent man in order to avoid a bloody riot." (Dwyer, 1999, p. 172/173).

The first part of this quote means that caretakers cannot tell children the right action/judgement in many situations because children do not create/encounter all those situations. I take "[I]mpart [...] moral information" to mean explicit verbal feedback that may take either the form of moral principles or other feedback from the caretakers that only shows whether that very action is permissible, such as gestures or facial expressions of approval or disapproval. But what kind of "moral information" is she thinking of? Is she thinking of explicit utterances of principles that help judging a scope of situations or of judgements about the permissibility of single situations, and does Dwyer think that children have to encounter all token situations to hypothesize the right kinds of rules or that it suffices to encounter every unknown type of situation in order to learn them? Her example in the end is not very enlightening, either: Do children not come to decide about the very problem whether it is permissible to hang a man to avoid a bloody riot or the more general type of problem, such as whether it is permissible to sacrifice one in order to save many? As usual, I will adapt the position that is most favorable to the people who argue that moral principles are innate:

Children do not create or come across every single situation that would make their caretakers give them feedback about its permissibility. Children cannot judge in all situations because they have not been provided with feedback about all situations.

This means that to assess new situations, they have to extrapolate principles and apply them to unknown situations, hence generate judgements. This is only an issue if extrapolating the right kind of principle is impossible because children encounter neither negative evidence nor explicit instructions and get stuck with their over-generalizations (hence, the LPLA applies).

Fiona Cowie makes a similar point as Dwyer, concerning language, not morality:

"Worst of all, however, is the fact that there are infinitely many strings of (say) English words that the child will never encounter in the data. Some of them she will not hear because they are not, in fact, English sentences; others, however, are perfectly good sentences that are absent from the data for the simple reason that no-one has gotten around to uttering them. What this means is that the mere non-occurrence of a string in the data cannot by itself constitute negative evidence. So even if she were always corrected when she made a mistake; and even if her interlocutors invariably spoke impeccably, there will always remain an infinity of strings, some of them English and some of them gibberish, that the child has no information one way or the other about." (Cowie, 1997, p. 21/22)

Cowie's claim can be divided into two parts: The first part condenses what many proponents of the Logical Problem of Language Acquisition claim: That the mere non-occurrence of a particular sentence [178] cannot be negative evidence. This is because there are so many sentences that no-one ever uttered that are nevertheless correct. If all those sentences were taken as negative evidence, and hence either excluded from the number of sentences the learner uses or used to wrongly narrow down over-generalizations, the learner would end up with a very deprived version of their target language.

The first problem Cowie has touched upon becomes salient when we do not encounter enough explicit teaching:

Mikhail (Mikhail 2008, p. 355) and Harman (Harman, 2011, p. 18) have argued that if the learner has no explicit positive instructions such as utterances of the principles themselves, they will not be able to acquire moral principles without innate information.

[178] "Sentence" in this case is a simplification; for the sake of general understandability, I try not to use too many technical terms from linguistics where they are not absolutely necessary to understand the argument.

According to the LPLA, transferred to the realm of morality, they would be able to but have to extrapolate the principles themselves. However, if they additionally have no negative evidence about many cases, even if they manage to extrapolate some principles, they are at risk of over-generalization. And, according to Cowie (Cowie, 1997, p. 21/22) negative evidence cannot be inferred from the non-occurrence of sentences: There are just too many correct sentences that people never encounter, either, and those should not be regarded as negative evidence. What is left then is explicit negative evidence such as explicit corrections ("the sentence you uttered is wrong" or "this action is impermissible", if we follow my argumentation about the helpfulness of post-hoc evaluations above).

I will show that if people manage to extrapolate principles, it is no major problem that they may be over-generalized even if there is no negative evidence (as in the sense of disapproving looks or instructions that some action/judgement is wrong) - there may still be sufficient evidence to restrict those principles and safely land on the target principle: If particular actions or judgements that the learner has predicted do not occur, that is a reason to let go of the principles that generated the prediction. The same holds if actions and judgements that follow some principle never occur. I will elaborate on this on "But Cowie argues", in the paragraph after the next one.

The second part of Cowie's quote resembles Dwyer's argument as I decided to understand it: That there will always remain many language strings whose grammaticality the learner will never learn about because they never encounter them. Remember Dwyer's argument, that children are not exposed to every moral situation possible. We can draw the following analogy: To assess whether language strings are correct/moral situations are permissible, children have to hypothesize and apply principles. And, as Cowie added, this is impossible because they do not get any negative evidence and absence of positive evidence is not negative evidence. That a sentence/situation does not arise does not mean that it is not grammatical/permissible.

But Cowie argues - and I think we can transfer this to the realm of morality - that the learner could make predictions based on what they already know. If those predictions were wrong, that would be negative evidence against the hypothesized rule that has generated the prediction (Cowie, 1997, p. 39). So if you hypothesize that the past form of "going" is "goed", but people say "went" every time you are expecting them to say "goed", [179] that would disprove your hypothesis, or at least make it less probable: Of course, as Cowie admits, "[o]ther explanations for the failure of match are available. [...] perhaps both of [the verb forms] are grammatical in that context, etc. But, [...] the fact that a theory may always be saved in the face of recalcitrant experience is hardly news; and it's not news either that more than a single failure of prediction may be needed to overturn a cherished hypothesis." (Cowie, 1997, p. 39).

But does this apply to the realm of moral principles as well? Dwyer's example that

[179] Cowie uses a more complicated, but linguistically more realistic example.

people seldom think about whether it is appropriate to hang someone to avoid a bloody riot is not a typical (Trolley Type) example of moral decisions that are made following genetically constrained principles. Still, the principles underlying it may be the same as in other cases where many people decide systematically: As in the “Ned”/“Oscar” case, you could apply the Doctrine of Double-Effect here: Although a good consequence is intended (you would hang the man in order to avoid a bloody riot; Doctrine of Double-Effect rule number 1 is fulfilled), you would use hanging the man (the bad consequence of your action) as a means to avoiding the riot. So even if you can be sure that more than one person will get killed in the riot if you do not hang the man (hence, according to our definition before, the good effects outweigh the bad effects and DDE rule number 4 is fulfilled), you are not allowed to sacrifice the man if you follow the Doctrine of Double-Effect.

Let us now transfer the Cowie case to this case. The principle that is hypothesized by the learner (in the example I made above: ”The correct past form of all verbs is their stem with ’-ed’ added”) to underlie decision-making has something to do with the Doctrine of Double-Effect. As Dwyer’s/Cowie’s argument can be counted as part of the Logical Problem of Language Acquisition, let us assume that someone wrongly hypothesizes an over-generalized principle. The target principle in our case would be the Doctrine of Double-Effect. We could, for example, assume that someone hypothesizes something like a principle consisting of Rules 1 and 4 of the Doctrine of Double-Effect, but not 2 and 3: They might assume that it is permissible to do something with a good and a bad effect if

1. A good consequence is intended and

2. The good effects outweigh the bad effects.

They would, in everyday (or, hopefully, not so everyday) practice, for example expect that it is permissible to hit someone in order to keep them from hitting more people. But this will mostly not happen in adult circles that endorse the Doctrine of Double-Effect and if children do act this way, people will reprimand them. If, however, someone cannot stop the person hitting someone else without hurting that person or accidentally hitting people around them, most people endorsing the Doctrine of Double-Effect will find this permissible. This non-occurrence of stopping a person from hitting others by hitting the person (hence, hitting them as a means to save the other people) in circles that (explicitly or implicitly) act according to the Doctrine of Double-Effect then will be negative evidence for the over-generalized principle: The easiest way to stop the person would be to hit them and this would also be permissible according to our over-generalized partial Doctrine of Double-Effect principle (Rules 1 and 4 of our formulation of the Doctrine of Double-Effect). So, our learner who holds the over-generalized principle will probably expect people to do this in a situation where someone has to be kept from hitting others. But people will rather try to hold the person, even if this results in hurting the person or hurting people around them. [180] If there are several cases where expectations based on the over-generalized principle fail, the learner will probably start to change the hypothesis. And this is connected to a second argument Cowie has made: As she formulates it, "[Non-occurrence] can serve as evidence that whole classes of objects are not instances of the kind whose extension is being learned." (Cowie, 1997, p. 39).

So, if no situations with a certain structure (those situations that lack the kind of structure whose extension is being learned) occur, this non-occurrence serves as evidence that those situations are not instances of the kind whose extension is being learned. In our example: If no situations occur where people hurt others as means (and not merely as a foreseen side-effect), this non-occurrence serves as evidence that those situations are not instances of situations where the Doctrine of Double-Effect is fulfilled. The Doctrine of Double-Effect, in this case, is the kind of principle whose extension is being learned. The classes of objects that do not occur are instances of situations that have the following feature in common: Harming people (or some other kind of bad effect) is used as a means and not only as a foreseen side-effect to saving people (or some other kind of good effect).

The latter two kinds of negative evidence (and any negative evidence, that is) only count if we hold the picture of a learner who discards their hypothesized principles as soon as they encounter negative evidence (or as soon as they encounter a certain amount of negative evidence). Let us assume that someone comes up with rules that fit the regularities of the sentences they hear. But there might be two correct forms for the same thing, such as two correct past forms for the verb "go": Both "went" and "goed" could be correct. If the learner would just accept those new rules (one past form of "go" is "went") and additionally stick to their over-generalized rules so that both co-exist, they would not change their rules if their hypothesized form (the correct past form of "go" is "went") was not confirmed. Instead, they would just add a new rule like "for 'go', the past form can be 'went' or 'goed'". So the argument that non-confirmation of the over-generalized hypotheses that the learner holds only works if:

1. The rules permit only one correct form for the same mode of one verb AND
2. There is a probabilistic rule along the lines of "if the hypothesized form never (or only extremely seldom) occurs, but a different form does, then I should change my hypothesized principle so that it fits the form that does occur".

Transferred to the moral realm: If we, for some reason, expect someone to harm people as a means to saving more people, but they do not do that, and we allow for two principles

[180] I have assumed here that the expectations for the situation does not only depend on the moral principles the person hypothesizes, but also on how easy and how dangerous or potentially harmful to others an action is. I am also aware that we could explain people's preference to hold someone instead of hitting them by different moral principles, namely "avoid harm wherever you can". This fictional situation is only to illustrate how predictions could be made regarding moral situations.

to be correct that lead to different reactions in the same situation, the following would happen:

We would think that in this situation, you can either use people as a means (because that is your over-generalized principle) or you can refrain from doing so (this is what the Doctrine of Double-Effect would tell you) and both are permissible (analogue to: "goed" and "went" are both correct). If we, however, had a probabilistic rule saying that "if the hypothesized form never (or only extremely seldom) occurs, then I should change my hypothesized principle so that it fits the form that does occur", and we would encounter similar situations a few times, hence situations where no-one harms anyone as means although that would be easier, this would lead us to change our over-generalized hypothesis.

A learner without some probabilistic, holistic picture (a statistical overview of all moral actions and a prediction, based on this overview, of how likely they are to happen) would not admit the non-occurrence of structures that follow certain rules as evidence against those rules either: If you use non-occurrence as counter-evidence, you have to have some statistic in mind about how probable it is that a structure following a certain rule occurs. If we become aware that people do not harm others as means to save more people, that is because we either have a hypothesis about what is going to happen in this very situation (the case above) or we have a different kind of hypothesis: Some kind of measure about how often we expect cases with this kind of structure to occur over a certain time frame or relative to cases with different kinds of structures. If this hypothesis then gets rejected due to their non-occurrence in the predicted time frame or because other situations that the learner expected to happen significantly less often than the predicted situation have actually occurred more often, this is evidence against rules that would generate this kind of structure.

The picture of learners who learn like little scientists, however, can persist: They just have to either only allow for one right form per verbal mode/moral rule or allow for some more probabilistic hypothesis confirmation. And the latter comes closer to the picture of a learner that I have in mind.

I have recounted Fiona Cowie's view on negative evidence, i.e. that some kinds of negative evidence do not need to be explicit if we assume that people change their hypotheses based on probabilistic thinking, and shown, via analogy, that this theory is applicable to both language and morality. Cowie's theory may be contestable but has at least found to be serious enough to be published, and I do not see any way that this theory is more vulnerable when applied to morality than when we apply it to language. But, of course, this is only a suggestion as to how the problem raised by Dwyer 1999 might be solved and this would profit from being fleshed out in the future.

In the previous chapter, I have shown that behavior does not need to be universal to be innate and that different features we tested for above (fixed developmental order, being a solution to a Pleistocene problem etc.) can be sufficient to make the behavior a candidate for a Poverty of Stimulus argument (hence, an innateness argument). I have shown that if a behavior is genetically inheritable, it is not necessarily adapted, but it is necessary for behavior to be inheritable to be an Evolutionary Psychologist adaptation. I have expounded what generativity means in context of moral principles, hereby contributing to the Linguistic Analogy by making it more precise. This was necessary to show that and how the Logical Problem of Language Acquisition is applicable to moral principles. I have shown why proponents of the Linguistic Analogy present certain kinds of arguments such as the need for negative evidence or explicit instructions that are clearly connected to the Logical Problem of Language acquisition, but seldom spelled out clearly. I have shown that even for principles such as the Doctrine of Double-Effect, which most people who judge in accordance with it cannot identify as source for their judgements, explicit formulations exist and many people might or do come across them and hence it is not true that people cannot have learned the Doctrine of Double-Effect because they do not encounter explicit formulations of it. However, the (not very decisive) evidence showed that being familiar with the principle might not be correlated to judging according to it. This suggests that either we do have an innate moral faculty that makes it unnecessary for us to learn them as Dwyer and the others hypothesized, or that we over-generalize initially but encounter enough negative evidence to make it still possible to learn the right principle, or that people acquire moral principles in a different way than 'like little scientists'. I have also shown that the kind of generalized explicit instructions and post-hoc evaluations Dwyer conceded that caretakers provide could be very useful for learning a language and that negative evidence is not completely absent from normal learning environments.

Hence, all in all, I have shown that it would in principle be possible for children to acquire our three principles ‘like little scientists’ because caretakers (and religious or philosophical environments) do provide explicit instructions and negative evidence. It is an open question, however, whether they do indeed acquire the principles this way. In the following section, I will provide some further possibilities of how those principles could be acquired.

7 Alternatives to innateness: A Social Learning Account for Morality

Before I begin , a small caveat. The following (sub-)chapter is not meant to present a complete theory. I will explore a possibility of how children might learn moral principles and be motivated to judge according to them as a suggestion for further research.

I will first present a solution that Pinker proposed regarding the LPLA [181] and then adapt it to the realm of morality.

Next, I will explain how we can apply it to our three moral principles, and, lastly, show how children can not only learn those principles (hence, the systematicity or theory behind moral behavior) that way, but also at the same time adopt the validity that their caretakers assign to them, so they, for instance, not only know (or cognize) that it is bad to harm someone as a means to a good effect, but also find it bad (and might be even motivated to act according to that verdict). This will be a very brief account aiming to show that it is possible to learn both languages and moral principles by recognizing patterns and hypothesizing principles, even without innate, domain-specific information/constraints, hence, without genuinely moral adaptations. This account can solve the LPLA or similar problems. For a more extensive and comprehensive characterization of both the linguistic account and an empiricist account on how people become motivated to act morally, see (Pinker 1984) and (Prinz 2007b).

Pinker's way of learning principles has the learner hypothesizing one 'constraint' every time they hear a linguistic form in a context. As Cowie sums it up:

> "In constraint sampling, the learner uses her analyses of the primary linguistic data as a basis for restricting the application of an over-general rule. Given an input sentence, the learner randomly selects one feature of that sentence and applies it as a constraint on the rule. Each possible constraint has a non-zero probability of being hypothesized at any given time and every constraint adopted is retained until a sentence violating it is encountered in the input, at which point it is dropped from the constraint-set. [...] Eventually, the child attains the correct set of constraints on the rule, and her grammar no longer overgenerates in that respect." ((Cowie 1997, 41), referring to (Pinker 1986))

Pinker gives the following example:

[181] And ways to avoid the unlearnability of certain kinds languages with certain kinds of methods as proven in Gold's theorem; the latter brings up even more learning problems than the LPLA (see (Johnson 2004, 583)) but I have not elaborated them here because no Linguistic Analogy proponent has made arguments that clearly referred to his theorem, but some of their arguments are clearly at least inspired by the LPLA. Dwyer and consorts did not explicitly refer to the LPLA when arguing about morality, but I could not explain their references to negative evidence and explicit instructions differently and Dwyer cites a book about the Logical Problem in one of her texts.

> “For example, on one occasion a child might hypothesize that [“]wants[”] is sensitive to object animacy, and this constraint is retained until it is disconfirmed. On another occasion he or she might hypothesize that it is sensitive to subject number, and that hypothesis will be retained indefinitely, since no subsequent input will ever disconfirm it.” (Pinker 1986, 68)

Hence, when the child comes across the word “wants” in a sentence in a situation, it hypothesizes one constraint for the use of that word; e.g. that you can only want living things. The sentence it has encountered, in this case, must have been one where someone wanted something that was alive. The next time the child comes across one sentence in which someone expresses that they want a chair (or something different inanimate), or after it came across several sentences where the object of “wants” was inanimate, it will drop that constraint forever.

In the first version (’chair’), children only need a single sentence that does not conform with a hypothesis to disconfirm it; in the second (and more realistic) version (’a number of sentences’), children need more than one instance of language use that goes against their hypothesized constraint to disconfirm hypotheses. Our acquisition theory would theoretically work with both versions, but as children who once heard someone speaking incorrectly typically do not acquire a mistaken language, I will assume that the second version is correct.

If the child hypothesizes a constraint that is correct (in English), e.g. that only a single person “wants” something, it will hold this belief forever as it can only be disconfirmed by incorrect sentences which, in our first version, should never occur and in the second version, the one I adapted, only seldom. And it is very improbable that a child systematically and often hears sentences that disregard rules in the same way; hence, if we choose the second version, it is very improbable that children will drop any hypotheses that were correct when they learn languages the way Pinker proposed.

As this model is about constraints, it can solve the Logical Problem of Language Acquisition: When a child holds an over-generalized rule, according to the ’little scientist’ theory, it might utter wrong sentences (e.g., “they wants”), and without getting corrected (negative evidence), it will never realize that it over-generalized that rule. In Pinker’s case, however, the child is eventually going to hypothesize further constraints that are consistent with all (version one) or by far the most (version two) of the inputs it has hitherto had. In this case, it will sooner or later hypothesize that “wants” can only be used for single subjects (Pinker 1986, 69). [182]

[182] Laurence and Margolis, however, criticize that “a child might easily discard a correct principle because it generates the wrong results in interaction with various other principles the child endorsed.” (Laurence and Margolis 2001, 229). I do not know whether this holds for constraints as well and they do not elaborate on this argument, apart from writing that “linguists [in the dominant tradition] rely upon a smaller number of powerful general principles that only work in interaction with one another.” (Laurence and Margolis 2001, 229). Again, more research is needed.

In the view of Pinker, his theory has a nativist component: The constraints are

> "randomly assembled within the following boundary conditions: (a) all constraints must be admissible according to Universal Grammar; (b) all the features and government relations that potentially can be encoded in a constraint have a nonzero probability of being included in the constraint hypothesized on a given occasion; (c) every constraint must be consistent with the current input sentence-plus-context; (d) constraints that were hypothesized previously but disconfirmed by subsequent inputs are not hypothesized a second time (or at least have a successively lower probability of being hypothesized every time they are disconfirmed)." (Pinker 1986, 68)

(b) says that all possible linguistic constraints have some probability (in contrast to none) to be hypothesized at some point of time, while (a) says that this only holds for constraints that are within the realm of Universal Grammar which, according to most of Chomsky's accounts (he has changed them numerous times), is innate (see, for instance, (Laurence and Margolis 2001), who (with good reason) treat Chomsky as the midwife of linguistic nativism). (c) means that the constraints cannot collide with the sentence and its context; if, for instance, someone were to say "I want this", the constraint that "wanting" only goes with "animacy" is only possible if "this" is animate; otherwise, the constraint would not be consistent with the uttered sentence and its context and, therefore, a hypothesis that is disconfirmed by the sentence (and context) it is based on itself. (d) is similar to our versions one and two: In our and Pinker's version one, if a constraint is disconfirmed, it will not arise again, while in our version two, the child will drop the hypothesis once it has been disconfirmed several times and in Pinker's version two, the hypothesis will arise again once it has been disconfirmed, but with successively lower probabilities every time the learner faces a counter-example.

But Cowie thinks that this innateness might not be necessary:

> "I should note that although Pinker's constraint-sampling heuristic is quite general, he is a nativist, holding that the space of possible constraints for a given rule-type is specified innately by U[niversal]G[rammar]. It seems to me, though, that if we allow (as all models of language-learning in fact do) that a learner can use her current state of grammatical knowledge to perform some preliminary syntactic analysis of the input, possible constraints might be suggested by that analysis, rather than (as in Pinker's model) being selected from an innately-specified set." (footnote in (Cowie 1997, 49))

What Cowie likely means is that the space of possible constraints is diminished the more we know about the structure of the sentence: To be able to hypothesize that the object of "wants" needs to be "animate", you have to already know that "wants" usually needs an object. To me, this raises the question how the learner has acquired their "current state of grammatical knowledge". But there is a further argument that weakens Pinker's model as compared to a model without innate components:

The number of possible grammatical constraints that, according to Pinker, is limited by what the respective version of Universal Grammar allows may still be huge: Depending on which Universal Grammar hypothesis we choose, the limitations could to different degrees be helpful in hypothesizing the 'right' kinds of features that all hitherto (or currently) witnessed sentences have in common. This means that, depending on our Universal Grammar, the space of possible constraints may still be close to the space of possible constraints without Universal Grammar.

I assume that the less boundaries Universal Grammar imposes on possible constraint features, the more sentences will be necessary to encounter the right kind of language or a sufficiently similar language [183] because the space of possible constraints to choose from will be wider. And several other factors may play a role in how realistic it might be for a learner to encounter the right constraints in a reasonable amount of time (hence, in a 'normal' setting with 'normal' linguistic input during a few years): The more features the child has to pick and guess, the more possibilities it has when it hypothesizes constraints. It might, for instance, besides obvious grammatical features like verb endings, incorporate features such as modulation (which, in many languages, plays a role in assigning meaning and might hence be regarded as grammatically meaningful) into its constraint hypotheses. And if the child only hypothesizes one constraint per input (with input being, for instance, a sentence), it will need a longer time to encounter all the correct constraints.

As I have mentioned above, it is only necessary for the child to end up hypothesizing a *sufficiently similar* set of principles to learn its language because people speak idiolects. Even in one language community, everyone speaks a slightly different language and people in one community may even follow very similar, but not identical grammatical rules[184] without anyone realizing it. Even beyond idiolects, people might hold some flawed beliefs that never get corrected. They might, for instance, make mistakes only in forms that come up very rarely or use wrong cases that are pronounced very similarly to the right case, such as mixing up German articles "den" and "dem", and other people might just not notice (compare (Laurence and Margolis 2001, 265), citing (Cowie 1998): "Cowie also claims that nativists make the mistake of wrongly assuming that there are no differences in the end product among language learners (i.e., of the same natural language)"). We could define a sufficiently similar language in social terms: A language whose great majority of sentences is deemed grammatically correct by a great majority of the child's language community[185].

[183] I am going to elaborate on what I mean by "sufficiently similar" in the next paragraph

[184] And even very different principles can produce the same outputs, just as very different algorithms can execute the same function if we define "function" and "algorithm" like David Marr (Marr, 1982).

[185] The extension of the language community, hence, the amount of variance that is deemed correct (whether it, for instance, includes different dialects), depends on other social factors. But this is a topic

But how is this account applicable to moral principles? Think back to our Loop with -Heavy Object Dilemma situation: Let us assume that the learner witnesses someone telling a story about the following situation: A train is rushing towards five people. On the route between the train and the five people, a side-track diverts and loops back to the main track. If a train was to be diverged to the loop track, it would return to the main track and continue rushing towards the five people. But on the loop track, a person is standing with a heavy stone behind them. If the train gets redirected onto the loop track, the train will hit the one person and they will die, but the train will stop and not return to the main track where it would have hit the five people and killed them. Is it permissible to redirect the train?

And let us assume that the learner would witness everyone who is listening to the story nodding and confirming that it is permissible to redirect the train; here the learner would learn morality the way Pinker suggested people might learn languages. The leaner would work like a 'little scientist', look for regularities, in our case even witness explicit remarks about the permissibility of an action (but not about the rules that make it permissible), produce hypotheses about the rules that created the regularity, and thereby likely over-generalize some rules.

Maybe they had previously witnessed other instances where people find it permissible to sacrifice one person to save five. They might then make the over-generalization that it is always permissible to kill one to save five others. But then they witness people finding it impermissible to divert the train onto the same loop track, but without anything behind the person (the Loop case).

The learner might then make an assumption about the constraints of their rule.

They might assume that it is only impermissible to divert trains to kill one and save five if there is a stone behind the one person (although it is acceptable to kill one and save five in all other situations). However, the next situation they encounter is the same, but the heavy object is made of metal. Still, everyone finds it permissible to divert the train onto the loop track. So, if we assume version one (one counter-example suffices to make someone drop a constraint for good), the leaner drops the stone constraint forever, but comes up with a new constraint. This time, it might be that harming someone to save other people with a train is permissible only if there is a heavy object behind the person who gets sacrificed. When they come across other cases where someone is harmed by a train as a side-effect, but without heavy objects, they will drop that constraint as well until they eventually hypothesize the Doctrine of Double-Effect. This may take some time, depending on how many properties of situations people assume to be possibly moral. With growing moral competence, it might be more and more obvious that the properties of harm done are often related to moralities while the materials of which things consist are usually only relevant because of other properties such as their value or their capacity to

too wide to discuss here.

harm people (e.g. their rigidity). But with enough examples and accumulating evidence, this should be a way to learn just the right kind of moral principle. If, for instance, our learner had been more experienced and had come across other, non-train-related instances of impermissible sacrifices, they would have hypothesized the Doctrine of Double-Effect with fewer guesses.

And if, in the end, you maintain a constraint such as "you should never sacrifice anyone as a means or foreseen side-effect to save others" because you never came encountered a case (or enough cases) to make you drop the "foreseen side-effect" constraint, this would make you part of the 28% who judged the "Loop with Heavy Object" case impermissible in Cushman 2007 (Cushman et al. 2007, 8). As we have seen in the Universality chapter, people are far from concordant in their judgements even with similar cultural upbringing, especially when it comes to the less obvious and more complex principles like the Doctrine of Double-Effect. This might be a hint that they actually do learn their principles in a way similar to what Pinker suggested for language acquisition and I adapted for morality here: The part of populations that does not judge in accordance with our three moral principles might just hold under- or over-generalized rules. Depending on the rules, a possible under-generalization might look like one of our constraints above ("you are only allowed to sacrifice one to save five with a train if there is a metal object behind that person") and lead to the judgement that sacrificing one person as a side-effect is impermissible. Or it might look like a different constraint such as "it is only impermissible to sacrifice one to save five if the one is on a loop track" and lead to the judgement that sacrificing one as a means is permissible. In any case, how you judge a situation would depend on what you have experienced and inferred from your experience (if we do not assume Pinker's innate constraints). This would be consistent with Catholics judging differently than Atheists.

This hypothesizing could (and probably would) take place in a way that the learner has no conscious access to it and cannot make it explicit, because when acquiring both languages and morality we hardly ever ponder about hypotheses and disconfirmations and Cushman et al. 2007 have shown that most people are not able to make the Doctrine of Double-Effect explicit, even if their judgements were consistent with it.

I do not claim that the acquisition strategy I have sketched is the one that children actually employ, but it is one way in which children could possibly acquire those principles without employing congenital resources other than domain-general learning mechanisms. If we agree to that, Poverty of Stimulus arguments do not hold anymore:

According to Poverty of Stimulus arguments, we cannot have acquired morality only getting the inputs learners normally get if we did not make use of some innate, domain-specific information. But most of us have acquired morality with just those inputs. Hence, we must have some innate, domain-specific information that helps us acquire the moral system of our culture (or of all humans).

If the model I adapted from Cowie and Pinker works, then we can have learned morality empirically, even without innate, domain-specific information. Hence we can explain the acquisition process (and why everyone or almost everyone has acquired some moral system) without resorting to domain-specific, innate knowledge.

I have shown a possible way for children to acquire the tools to generate their own 'correct' judgements. If they learn morality like this, they 'know' or 'cognize' how to generate judgements according to our three principles; they have the ability to do so because they can recognize which criteria make an action morally right or wrong. But I have not yet explained why they should judge according to them; why they should actually find actions right or wrong. Just because we know how utilitarian judgements work and that many people in our surroundings use them, we do not automatically adopt a utilitarian moral system. And the same holds for moral principles: That our learner knows how to apply those principles in the same way as their environment applies them does not mean that they agree with the principles or endorse them.

To explain this part, I will resort to a theory by Jesse Prinz. His theory shows how children come to acquire the same validity for moral actions as their environments. Basically, it concerns "Emotional Conditioning": "If caregivers punish their children for misdeeds, by physical threat or withdrawal of love, children will feel badly about doing those things in the future. Herein lie the seeds of remorse and guilt." (Prinz 2007a, 404). If children do something that was not permissible according to our rules, their parents will judge it to be wrong. Even if they do not give explicit feedback like the reactions discussed above ("you should never lie" or something along those lines), they will let their children feel when they did something unwanted and make them feel uncomfortable. Prinz thinks that the negative emotions evoked by those negative reactions are the basis of some negative moral judgements (J. J. Prinz 2007b, 105). As he is a sentimentalist, he believes that moral terms like "right" and "wrong" express sentiments, sentiments being tendencies to have a certain emotion (Prinz 2006, 34). When we, therefore, judge that it is wrong (or impermissible) to kill someone as a means to saving someone else, we express that we have a disposition to feel some negative (moral) emotion when we witness or imagine someone doing that.

He also has a theory how children get taught morality: He thinks that caregivers condition children by "power assertion (physical punishment or threat of punishment), which elicits fear" or by "induction, which elicits distress by orienting a child to some harm she has caused to another person" or, thirdly, by "love withdrawal, which elicits sadness through social ostracism ('If you behave like that, I'm not going to play with you!')." (all three: (Prinz 2006, 32), based on (Hoffman 1983). Prinz thinks that this causes children to "experience negative emotions in conjunction with misdeeds", and later writes that "if a person did harbor a strong negative sentiment towards killing, we would say that she believes killing to be morally wrong, even if she did not have any explicit belief about whether killing diminished utility [...]." (both quotes: (Prinz 2006, 32)).

But 'conditioning' children by exerting influence on them is not the only way to make them feel negatively (or have a disposition to experience negative emotions) when they think about an action or situation: I assume that people who hear about situations where someone killed someone else as a means to saving five and morally disapprove of this action will be shocked and create an atmosphere that shows the learner that the person would have been blamed had they been present; hence, it creates fear of being in the agent's position. Even the atmosphere of shock might be sufficient to evoke negative emotions in the learner in the future when reminded of a situation like that. Negative emotions displayed by the judging person, anyway, tend to evoke negative emotions in the persons present: The name of this phenomenon is "Emotional Contagion" (for a classic on Emotional Contagion, see, for instance, (Hatfield, Cacioppo, and Rapson 1993).) If, as Prinz argues, negative moral judgements about a situation or action express negative sentiments toward that situation or action, and sentiments are a disposition to experience emotions, a person who judges tends to experience negative emotions when they make impermissibility judgements. In our example, the learner's environment will be contagious with negative emotions whenever killing one as a means to saving five comes up. If the learner adopts those emotions in this situation, they will in recurring situations acquire not only the structure (by induction as we have seen above), but also the valence of the moral judgements that their environment makes. Hence, the child starts having negative sentiments when it thinks about sacrificing someone as a means to saving five. If the child expresses those sentiments, it will make negative moral judgements about the situation.

I conclude that I have sketched a method how children could, in principle, acquire moral principles such as our three principles without innate, domain-specific information. Other linguists have made different proposals on how to learn languages than the one I have presented (See, e.g., (Pinker 1979) or (O Duda, E Hart, and G.Stork 2001), Chapters 8.7 and 8.8, pp. 429-434)). I do not claim that this is the only possible model to learn languages without innate, domain-specific information, nor do I claim that this is the only way to learn moral principles without said information. I have merely given a how-possible explanation. Now that I have shown a possible way to acquire moral principles without innate, domain-specific information (and, possibly, after others have created more detailed models of acquiring them this way), it is the Linguistic Analogy proponents' turn to show why they should not have been acquired like this. Our models are no longer dependent on innate information and we are free to develop a model of morality learning that accounts best for people's behaviors, which are far from universally homogeneous.

8 Conclusion

I have shown that Trolley Dilemma judgement mechanisms are not domain-specific and mechanistically modular. I have shown that Functional Modularity might not be the right paradigm to examine whether behaviors in social contexts are evolutionary adaptations. I have argued that children from 4 years on start making moral judgements along the lines of our three principles, and this might be because some domain-general components of the judgement mechanism such as understanding others' intentions and weighing sums develop at this time. To find out whether there is a critical stage of learning the principles we would need more empirical evidence. The same holds for the question of whether children at the age of 4 have not had enough moral input to empirically learn the principles without domain-specific, innate information, hence, whether the argument for Poverty of Stimulus applies. After this, I have shown that Trolley Dilemma judgements are not universal, neither inner- nor cross-culturally, and people with different upbringing show different judgement patterns. Hence, Linguistic Analogy proponents' argument that the cross-cultural universality of Trolley Dilemma judgements shows that such judgements may be innate should be dropped.

I have considered Poverty of Stimulus arguments, mainly those connected to the Logical Problem of Language Acquisition, and shown that under most conditions the learner has input that would keep them from over-generalization. Although, against the assumption of some Linguistic Analogists, some people are familiar with explicit versions of the Doctrine of Double-Effect, this does not seem to have an impact on their Trolley Dilemma judgements. I have argued that this might be because the learner acquires them implicitly by interaction with other people who judge according to the Doctrine of Double-Effect and sketched an account, based on Pinker and Cowie's account of learning languages and Jesse Prinz's moral sentimentalist framework, of how children might learn Trolley Dilemma principles empirically without domain-specific knowledge.

I conclude that in the light of all evidence discussed, it is very improbable that Trolley Dilemma judgement mechanisms are evolutionary adaptations or otherwise based on inheritable domain-specific information.

Recently, someone asked me whether I thought Experimental Philosophy had become better over the years. I responded, referring to Trolley Dilemma studies,[186] that I thought there had been a tremendous evolution in Experimental Philosophy: Trolley Dilemma research started with very general studies and hypotheses and has proceeded to very specialized studies, all the way accompanied by a dialogue between ethics and meta-ethics theorists and empirical researchers; and today many trained philosophers collaborate

[186] Notwithstanding that many of these studies had been made by psychologists or psychologists working with philosophers, lawyers and anthropologists, because I consider them as part of Experimental Philosophy in view of their reception and interpretations which have strongly influenced moral philosophy.

closely with psychologists and have acquired psychological methodology. I think that those collaborations are only the beginning and I hope, with this thesis, to make my own contribution to the realm where psychology and philosophy cross.

A Bibliography

Abarbanell, L., Hauser, M.D., 2010. Mayan morality: An exploration of permissible harms. Cognition 115, 207–224.

Abello-Contesse, C., 2009. “Age and the Critical Period Hypothesis.” ELT Journal 63 (2): 170–172.

Adriaans, Pieter, ”Information”, in: Edward N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Fall 2013 Edition), https://plato.stanford.edu/archives/fall2013/entries/information/.

Ahlenius, H., Tännsjö, T., 2012. Chinese and Westerners Respond Differently to the Trolley Dilemmas. Journal of Cognition and Culture 12, 195–201.

Anderson, M.L., 2010. Neural reuse: a fundamental organizational principle of the brain. Behav Brain Sci 33, 245–266; discussion 266–313.

Aquinas, S.T., 1265. Summa theologiae: Latin text and English translation, introductions, notes, appendices, and glossaries. Blackfriars; McGraw-Hill, New York; [Cambridge, England].

BAHFest, 2014. BAHFest 2013 - Zach Weinersmith: Weinersmith’s Infantapaulting Hypothesis. (Onlinevideo, https://www.youtube.com/watch?v=94_omZ2RnfI. Acc. 07/09/2015.)

Baird, J.A., Astington, J.W., 2004. The role of mental state understanding in the development of moral cognition and moral action. New Directions for Child and Adolescent Development 2004, 37–49.

Baldwin, C.Y., Clark, K.B., 2006. Modularity in the Design of Complex Engineering Systems, in: Braha, D., Minai, A.A., Bar-Yam, Y. (Eds.), Complex Engineered Systems, Understanding Complex Systems. Springer Berlin Heidelberg, pp. 175–205.

Banerjee, K., Huebner, B., Hauser, M., 2010. Intuitive Moral Judgments are Robust across Variation in Gender, Education, Politics and Religion: A Large-Scale Web-Based Study. Journal of Cognition and Culture 10, 253–281.

Baron, J., Granato, L., Spranca, M., Teubal, E., 1993. Decision making biases in children and early adolescents: Exploratory studies. Merrill Palmer Quarterly 23–47.

Baron, J., Ritov, I., 2004. Omission bias, individual differences, and normality. Organizational Behavior and Human Decision Processes 94, 74–85.

Barrett, H.C., 2012. A hierarchical model of the evolution of human brain specializations. Proceedings of the National Academy of Sciences 109, 10733–10740.

Barrett, H.C., 2009. Where there is an adaptation, there is a domain: The form-function fit in information processing, in: Foundations in Evolutionary Cognitive Neuroscience. Cambridge University Press.

Barrett, H.C., 2008. Evolved Cognitive Mechanisms and Human Behavior, in: Crawford, C., Krebs, D. (Eds.), Foundations of Evolutionary Psychology: Ideas, Issues, Applications and Findings. Erlbaum Associates, Mawah, NJ, pp. 173–191.

Barrett, H.C., 2005a. Enzymatic Computation and Cognitive Modularity. Mind & Language 20, 259–287.

Barrett, H.C., 2005b. Adaptations to Predators and Prey, in: Buss, D.M. (Ed.), The Handbook of Evolutionary Psychology. John Wiley & Sons, New York, NY, pp. 200–223.

Barrett, H.C., Kurzban, R., 2006. Modularity in cognition: framing the debate. Psychol Rev 113, 628–647.

Betzig, L., 1998. Not Whether to Count Babies, but Which, in: Crawford, C., Krebs, D.L. (Eds.), Handbook of Evolutionary Psychology: Ideas, Issues, and Applications. Lawrence Erlbaum Assoc Inc, Mahwah, NJ, pp. 265–273.

Böhning-Gaese, K., Oberrath, R., 1999. Phylogenetic effects on morphological, life-history, behavioural and ecological traits of birds. Evol Ecol Res 1, 347–364.

Bonvillian, J.D., Folven, R.J., 1993. Sign language acquisition: Developmental perspectives, in: Marschark, M., Clark, D. (Eds.), Psychological Perspectives on Deafness. Erlbaum, Hillsdale, NJ, pp. 229–265.

Bramble, D.M., Lieberman, D.E., 2004. Endurance running and the evolution of Homo. Nature 432, 345–352.

Broderick, J.P., 1975. Modern English Linguistics: A Structural and Transformational Grammar. Crowell.

Brown, J.E., Kahn, E.S., Hartman, T.J., 1997. Profet, profits, and proof: do nausea and vomiting of early pregnancy protect women from “harmful” vegetables? Am. J. Obstet. Gynecol. 176, 179–181.

Brown, R., Cazden, C., Bellugi, U., 1969. The Child’s Grammar from I to III, in: Minnesota Symposium on Child Psychology. University of Minnesota Press, Minneapolis.

Brown, R., Hanlon, C., 1970. Derivational complexity and order of acquisition in child speech, in: Hayes, J. (Ed.), Cognition and the Development of Language. John Wiley, New York.

Bruers, S., Braeckman, J., 2014. A Review and Systematization of the Trolley Problem. Philosophia 42, 251–269.

Brumbach, H.J., Jarvenpa, R., 2006. Gender dynamics in hunter-gatherer society: archaeological methods and perspectives, in: Nelson, S.M. (Ed.), Handbook of Gender in Archaeology. Altamira Press, Walnut Creek, CA, pp. 503–535.

Buller, D.J., 2006a. Adapting Minds: Evolutionary Psychology and the Persistent Quest for Human Nature. A Bradford Book, Cambridge, Mass.

Buller, D.J., 2006b. Evolutionary Psychology: A Critique, in: Sober, E. (Ed.), Conceptual Issues in Evolutionary Psychology. MIT Press, Cambridge, MA, pp. 197–214.

Buss, D.M., 2008. Evolutionary Psychology: The New Science of the Mind, 3rd edition. Pearson/Allyn and Bacon, Boston.

Buss, D.M., 1995. Evolutionary Psychology: A New Paradigm for Psychological Science. Psychological Inquiry 6, 1–30.

Buss, D.M., 1989. Sex differences in human mate preferences: Evolutionary hypotheses tested in 37 cultures. Behavioral and Brain Sciences 12, 1–14.

Buss, D.M., Abbott, M., 1990. International Preferences in Selecting Mates. Journal of crosscultural psychology, 21(1), 5-47

Buss, D.M., Shackelford, T.K., 2008. Attractive women want it all: Good genes, economic investment, parenting proclivities, and emotional commitment. Evolutionary Psychology 6, 134–146.

Calabretta, R., Nolfi, S., Parisi, D., Wagner, G.P., 2000. Duplication of modules facilitates the evolution of functional specialization. Artif. Life 6, 69–84.

Cameron, E.L., 2014. Pregnancy and olfaction: a review. Frontiers in psychology 5, 1–11.

Carruthers, P., 2006. The case for massively modular models of mind, in: Stainton, R.J. (Ed.), Contemporary Debates in Cognitive Science. Blackwell, pp. 3–21.

Carruthers, Peter. 2005. "Distinctively Human Thinking: Modular Precursors and Components." in: The Innate Mind, Peter Carruthers, Stephen Laurence, and

Stephen Stich (Eds.): Structure and Content, 69–88. New York: Oxford University Press.

Carruthers, P., Laurence, S., Stich, S., 2005. The Innate Mind: Structure and Contents. Oxford University Press, USA.

Cattell, N.R., 1972. The new English grammar: A descriptive introduction. MIT Press, Cambridge (Mass.).

Cavanaugh, T., 1997. Aquinas's Account of Double Effect. The Thomist, 61, 107–121.

Chib, V.S., Rangel, A., Shimojo, S., O'Doherty, J.P., 2009. Evidence for a Common Representation of Decision Values for Dissimilar Goods in Human Ventromedial Prefrontal Cortex. J. Neurosci. 29, 12315–12320.

Chomsky, N., 1991. Some Notes on Economy of Derivation and Representation, in: Freidin, R. (Ed.), Principles and Parameters in Comparative Grammar. MIT Press, Cambridge (Mass.): Department of Linguistics, 417–454.

Chomsky, N., 1980. Rules and Representations, 2005 ed. Columbia University Press.

Chomsky, N., 1965. Aspects of the Theory of Syntax, 1st Paperback Ed. The MIT Press.

Chomsky, N., 1957. Syntactic structures. Walter de Gruyter.

Chomsky, N., Hornstein, N., 1980. Rules and Representations, Second Edition,. ed. Columbia Univ Pr, New York.

Chomsky, N., Jakobovits, I.L.A., Miron, M.S., 1959. A Review of BF Skinner's Verbal Behavior. Language 35, 26–58.

Confer, J.C., Easton, J.A., Fleischman, D.S., Goetz, C.D., Lewis, D.M.G., Perilloux, C., Buss, D.M., 2010. Evolutionary psychology. Controversies, questions, prospects, and limitations. Am Psychol 65, 110–126.

Conroy-Beam, D., Buss, D.M., Pham, M.N., Shackelford, T.K., 2015. How Sexually Dimorphic Are Human Mate Preferences? Pers Soc Psychol Bull 41(8), 1082-1093.

Cosmides, L., Tooby, J., 2006. Evolutionary Psychology: Theoretical Foundations, in: Nadel, L. (Ed.), Encyclopedia of Cognitive Science. John Wiley & Sons, Ltd, pp. 54–64.

Cosmides, L., Tooby, J., 1997. The Modular Nature of Human Intelligence, in: Scheibel, A.B., Schopf, J.W. (Eds.), The Origin and Evolution of Intelligence. Jones & Bartlett Learning, pp. 71–99.

Cosmides, L., Tooby, J., 1994. Origins of domain specificity: The evolution of functional organization, in: Hirschfeld, L.A., Gelman, S.A. (Eds.), Mapping the Mind. Domain Specificity in Cognition and Culture. Cambridge University Press, New York, NY, US, Melbourne, Australia, pp. 85–117.

Cosmides, L., Tooby, J., 1994. Beyond intuition and instinct blindness: toward an evolutionarily rigorous cognitive science. Cognition 50, 41–77.

Couzens, D., Haynes, M., Cuskelly, M., 2012. Individual and environmental characteristics associated with cognitive development in Down syndrome: a longitudinal study. J Appl Res Intellect Disabil 25, 396–413.

Cowie, F., 2010. Innateness and Language, in: Edward N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Fall 2017 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/fall2017/entries/innateness-language/.

Cowie, F., 1998. What's Within?: Nativism Reconsidered, 1st edition. Oxford University Press, New York.

Cowie, F., 1997. The Logical Problem of Language Acquisition. Synthese 111, 17–51.

Craig, E. (Ed.), 1998. Routledge Encyclopedia of Philosophy. Routledge, London ; New York.

Cruz, H.D., Smedt, J.D., 2009. The Innateness Hypothesis and Mathematical Concepts. Topoi 29, 3–13.

Curran, C.E., 2008. Catholic Moral Theology in the United States: A History. Georgetown University Press, Washington, D.C.

Cushman, F., 2014. The Psychological Origins of the Doctrine of Double Effect. Criminal Law, Philosophy 1–14.

Cushman, F., Young, L., 2011. Patterns of Moral Judgment Derive From Nonmoral Psychological Representations. Cognitive Science 35, 1052–1075.

Cushman, F., Young, L., Greene, J.D., 2010. Our multi-system moral psychology: Towards a consensus view, in: The Oxford Handbook of Moral Psychology.

Cushman, F., Young, L., Hauser, M., 2006. The Role of Conscious Reasoning and Intuition in Moral Judgment: Testing Three Principles of Harm. Psychological Science 17, 1082–1089.

Cushman, F., Young, L., Kang-Xing Jin, R., Mikhail, J., Hauser, M., 2007. A Dissociation Between Moral Judgments and Justifications. Mind Language 22, 1–21.

Dahlgrün, M.H., 2015. The Broad Foundations of Adaptationist-Computational Evolutionary Psychology, in: Breyer, T. (Ed.), Epistemological Dimensions of Evolutionary Psychology. Springer New York, 19–68.

Damasio, A.R., 1996. The somatic marker hypothesis and the possible functions of the prefrontal cortex. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 351, 1413–1420.

de Graaf, G., van Hove, G., Haveman, M., 2013. More academics in regular schools? The effect of regular versus special school placement on academic skills in Dutch primary school students with Down syndrome. Journal of Intellectual Disability Research 57, 21–38.

Doris, J. M., Machery, E., & Stich, S. (2017, May 10). Can Psychologists Tell Us Anything About Morality? The Philosophers' Magazine. (Online magazine, http://www.philosophersmag.com/essays/153-can-psychologists-tell-us-anything-about-philosophy).

Dutton, D.G., Aron, A.P., 1974. Some evidence for heightened sexual attraction under conditions of high anxiety. Journal of Personality and Social Psychology 30, 510–517.

Dwyer, S., 2008. Dupoux and Jacob's moral instincts: throwing out the baby, the bathwater and the bathtub. Trends in Cognitive Sciences 12, 1–2.

Dwyer, S., 2006. How Good is the Linguistic Analogy?, in: Carruthers, P., Laurence, S., Stich, S.P. (Eds.), The Innate Mind, Vol. 2: Culture and Cognition. Oxford University Press, 237-256.

Dwyer, S., 1999. Moral Competence, in: Murasugi, K., Stainton, R. (Eds.), Philosophy and Linguistics. Boulder: Westview Press, 169–190.

Dwyer, S., Huebner, B., Hauser, M.D., 2009. The Linguistic Analogy: Motivations, Results, and Speculations. Topics in Cognitive Science 2, 486–510.

Epstein, R.A., 2004. Cases and Materials on Torts, 8th ed. Aspen Publishers, Aspen.

Ermer, E., Cosmides, L., Tooby, J., 2007. Functional Specialization and the Adaptationist Program, in: Gangestad, S.W., Simpson, J.A. (Eds.), The Evolution of Mind: Fundamental Questions and Controversies. Guilford Press, New York, NY, US, 153–160.

Finley, S., 2012. The Role of Negative and Positive Evidence in Adult Phonological Learning, in: Penn Working Papers in Linguistics, 18(1), 61-68.

Fischer, J.M., Ravizza, M., 1992. Ethics: Problems and Principles. Harcourt Brace Jovanovich College Publishers, Fort Worth, TX.

Fischer, J. M., Ravizza, M., & Copp, D. (1993). Quinn on Double Effect: The Problem of "Closeness." Ethics, 103(4), 707–725.

Fodor, J.A., 1985. Précis of The Modularity of Mind. Behavioral and Brain Sciences 8, 1–5.

Fodor, J.A., 1983. The Modularity of Mind. MIT Press, Cambridge, MA, US.

Fodor, J.A., 1981. The Present Status of the Innateness Controversy, in: Fodor, J.A. (Ed.), Representations. MIT Press, Cambridge, MA, US, pp. 257–316.

Fodor, J.A., Pylyshyn, Z.W., 1988. Connectionism and cognitive architecture: a critical analysis. Cognition 28, 3–71.

Foot, P. (1967). The Problem of Abortion and the Doctrine of Double Effect. Oxford Review, 5, 5–15.

Fraser, B., Hauser, M., 2010. The Argument from Disagreement and the Role of Cross-Cultural Empirical Data. Mind & Language 25, 541–560.

Gamez, D., 2009. The Potential for Consciousness of Artificial Systems. International Journal of Machine Consciousness 1, 213–223.

Garcia, C. L. (2007). Cognitive modularity, biological modularity, and evolvability. Biological Theory, 2(1), 62–73.

Geary, D.C., Huffman, K.J., 2002. Brain and cognitive evolution: forms of modularity and functions of mind. Psychol Bull 128, 667–698.

Gell-Mann, M., 1988. Simplicity and Complexity in the Description of Nature. Engineering and Science 51, 2–9.

Gerhart, J. C., & Kirschner, M. W. (2003). Evolvability. In B. K. Hall & W. M. Olson (Eds.), Keywords and Concepts in Evolutionary Developmental Biology, 133–137. Cambridge, MA: Harvard University Press.

Gittleman, J.L., Anderson, C.G., Kot, M., Luh, H.-K., 1996. Phylogenetic lability and rates of evolution: a comparison of behavioral, morphological and life history traits, in: Martins, E.P. (Ed.), The Comparative Method in Animal Behavior. Oxford University Press, Oxford, 166–205.

Gold, N., Colman, A., Pulford, B., 2014. Cultural Differences in Responses to Real-Life and Hypothetical Trolley Problems. Judgment and Decision Making 9, 65–76.

Gold, N., Pulford, B.D., Colman, A.M., 2015. Do as I Say, Don't Do as I Do: Differences in moral judgments do not translate into differences in decisions in real-life trolley problems. Journal of Economic Psychology 47, 50–61.

Gould, S.J., 1978. Sociobiology: The art of storytelling. New Scientist 80, 530–533.

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

Greene, J.D., 2008. The Secret Joke of Kant's Soul, in: W. Sinnott-Armstrong (Ed.), Moral Psychology. (Vol. 3. The neuroscience of morality: Emotion, brain disorders, and development), 35–80. Cambridge (Mass.): MIT Press.

Greene, J.D., 2004. Cognitive Neuroscience and the Structure of the Moral Mind, in: Carruthers, P., Laurence, S., Stich, S. (Eds.), The Innate Mind: Structure and Contents. Oxford University Press, New York, Oxford, 338–352.

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

Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M., Cohen, J.D., 2001. An fMRI Investigation of Emotional Engagement in Moral Judgment. Science 293(5537), 2105–2108.

Greene, J., Haidt, J., 2002. How (and where) does moral judgment work? Trends Cogn. Sci. (Regul. Ed.) 6, 517–523.

Grossi, G., 2014. A module is a module is a module: evolution of modularity in Evolutionary Psychology. Dialect Anthropol 38, 333–351.

Gustafson, J.M., 1998. A Protestant Ethical Approach, in: Lammers, S.E., Verhey, A. (Eds.), On Moral Medicine: Theological Perspectives in Medical Ethics. Wm. B. Eerdmans Publishing, 600–612.

Haidt, J., Baron, J., 1996. Social roles and the moral judgement of acts and omissions. Eur. J. Soc. Psychol. 26, 201–218.

Haidt, J., Joseph, C., others, 2007. The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules, in: The Innate Mind. Oxford University Press New York, 367–392.

Haidt, J., Koller, S.H., Dias, M.G., 1993. Affect, culture, and morality, or is it wrong to eat your dog? J Pers Soc Psychol 65, 613–628.

Haith, M.M., 1998. Who put the cog in infant cognition? Is rich interpretation too costly? Infant Behavior & Development 21, 167–179.

Hamilton, W.D., 1964. The genetical evolution of social behaviour. II. J Theor Biol 7, 17–52.

Harman, G., 2008. Using a Linguistic Analogy to Study Morality, in: Sinnott-Armstrong, W. (Ed.), Moral Psychology, Vol 1: The Evolution of Morality: Adaptations and Innateness. MIT Press, Cambridge, MA US, 353–359.

Harman, G., 2000a. Explaining Value and Other Essays in Moral Philosophy. Oxford University Press.

Harman, G., 2000b. Moral Philosophy and Linguistics, in: Explaining Value and Other Essays in Moral Philosophy. Clarendon Press, Oxford, 217–226.

Harman, G., 1999. Moral Philosophy and Linguistics, in: The Proceedings of the Twentieth World Congress of Philosophy 1, 107–115.

Hatfield, E., Cacioppo, J.T., Rapson, R.L., 1993. Emotional Contagion. Current Directions in Psychological Science 2, 96–100.

Haugeland, J., 1995. Mind Embodiend and Embedded, in: Houng, Y., Ho, J. (Eds.), Mind and Cognition. Academia Sinicy, Taipei, 207–237.

Hauser, M., 2006. Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong. Ecco.

Hauser, M., Young, L., Cushman, F., 2008. Reviving Rawls' linguistic analogy. Moral psychology 2, 107–144.

Hoffman, M.L., 1983. Affective and cognitive processes in moral internalization: An information processing approach, in: Higgins, E.T., Ruble, D.N., Hartup, W.W. (Eds.), Social Cognition and Social Development: A Sociecultural Perspective. John Wiley, New York, 236–274.

Hudson, N.J., 2011. Musical beauty and information compression: Complex to the ear but simple to the mind? BMC Research Notes 4,9.

Jacob, P., Dupoux, E., 2007. Universal moral grammar: a critical appraisal. Trends in Cognitive Sciences 11, 373–378.

Johnson, J.S., Newport, E.L., 1989. Critical period effects in second language learning: the influence of maturational state on the acquisition of English as a second language. Cognitive psychology 21, 60–99.

Johnson, K., 2004. Gold's theorem and cognitive science. Philosophy of Science 71, 571–592.

Johnson, M., 2011. There is no moral faculty. Philosophical Psychology 25, 409–432.

Joyce, R., 2013. The Many Moral Nativisms, in: Sterelny, K., Calcott, B., Fraser, B. (Eds.), Cooperation and Its Evolution. MIT Press, Cambridge, MA, 549–572.

Kahane, G., 2015. Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. Social Neuroscience, 10(5), 551–560.

Kahane, G., 2012. On the Wrong Track: Process and Content in Moral Psychology. Mind & Language 27, 519–545.

Kahane, G., Everett, J. A. C., Earp, B. D., Farias, M., & Savulescu, J., 2015. 'Utilitarian' judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. Cognition, 134, 193–209.

Kahane, G., Shackel, N., 2010. Methodological Issues in the Neuroscience of Moral Judgement. Mind & Language 25, 561–582.

Kahneman, D., Tversky, A., 1984. Choices, values, and frames. American Psychologist 39, 341–350.

Kamm, F. M., 1989. Harming Some to Save Others. Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 57(3), 227–260.

Kant, I., 1785. Groundwork of the Metaphysics of Morals. (M. Gregor, Trans.). Cambridge: Cambridge University Press.

Kaplan, H., Hill, K., Lancaster, J., Hurtado, A.M., 2000. A theory of human life history evolution: Diet, intelligence, and longevity. Evol. Anthropol. 9, 156–185.

Kaufmann, D., 2015. Philosophical commitments in the Neuroscience of Morality (PhD Thesis). Graduate School of Systemic Neurosciences, Munich.

Kauppinen, A., 2015. Moral Intuition in Philosophy and Psychology, in: J. Clausen & N. Levy (Eds.), Handbook of Neuroethics, 169–183. Springer, Dordrecht.

Kelly, D., Stich, S., Haley, K.J., Eng, S.J., Fessler, D.M.T., 2007. Harm, Affect, and the Moral/Conventional Distinction. Mind & Language 22, 117–131.

Kelly, R.L., 1995. The foraging spectrum: Diversity in hunter-gatherer lifeways. Smithsonian Institution Press, Washington.

Kirkby, D., Hinzen, W., Mikhail, J., 2013. Your theory of the evolution of morality depends upon your theory of morality. Behavioral and Brain Sciences 36, 94–95.

Kuhl, P.K., Conboy, B.T., Padden, D., Nelson, T., Pruitt, J., 2005. Early Speech Perception and Later Language Development: Implications for the "Critical Period." Language Learning and Development 1, 237–264.

Kurzban, R., 2012. Why Everyone (Else) Is a Hypocrite: Evolution and the Modular Mind. Princeton University Press.

Laurence, S., Margolis, E., 2001. The Poverty of the Stimulus Argument. Br J Philos Sci 52, 217– 276.

Lehrman, D.S., 1953. A critique of Konrad Lorenz's theory of instinctive behavior. Q Rev Biol 28, 337–363.

Lenneberg, E.H., 1967. Biological Foundations of Language. Wiley, New York.

Levy, N., 2004. Evolutionary Psychology, Human Universals, and the Standard Social Science Model. Biology and Philosophy 19, 459–472.

Lillo-Martin, D., 1999. Modality effects and modularity in language acquisition: The acquisition of American Sign Language, in: Ritchie, W.C., Bhatia, T.K. (Eds.), Handbook of Language Acquisition. Academic Press, San Diego, CA, 531–568.

Lucas, M.M., Wagner, L., Chow, C., 2008. Fair game: The intuitive economics of resource exchange in four-year olds. Journal of Social, Evolutionary, and Cultural Psychology 2, 74–88.

Mach, E., 1894. Popular Scientific Lectures, 1943rd ed. Open Court, Illinois.

Machery, E., 2008. Massive Modularity and the Flexibility of Human Cognition. Mind & Language 23, 263–272.

Machery, E., 2011. Discovery and Confirmation in Evolutionary Psychology, in: Prinz, J.J. (Ed.), The Oxford Handbook of Philosophy of Psychology. Oxford University Press, Oxford, England.

Machery, reviewed by E., Barrett, H.C., 2006. David J. Buller: Adapting Minds: Evolutionary Psychology and the Persistent Quest for Human Nature,. Philosophy of Science 73, 232–246.

MacWhinney, B., 2004. A multiple process solution to the logical problem of language acquisition. J Child Lang 31, 883–914.

Mangan, J., 1949. An Historical Analysis of the Principle of Double Effect. Theological Studies 10, 41–61.

Marcus, G.F., 2006. Cognitive architecture and descent with modification. Cognition 101, 443–465.

Marks, I.M., 1987. Fears, phobias, and rituals: panic, anxiety, and their disorders. Oxford University Press, New York.

Marr, D., 1982. Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information. Henry Holt and Company.

McClamrock, R., 2006. Modularity, in: Encyclopedia of Cognitive Science. John Wiley & Sons, Ltd.

McIntyre, A., 2014. Doctrine of Double Effect. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Winter 2014). https://plato.stanford.edu/archives/win2014/entries/double-effect/.

McNeill, D., 1966. The creation of language by children, in: Wales, R. (Ed.), Psycholinguistics Papers. University of Edinburgh Press, Edinburgh.

Mercier, H., Sperber, D., 2010. Why Do Humans Reason? Arguments for an Argumentative Theory, Behavioral and Brain Sciences 34(2), 57-74

Mikhail, J., 2011. Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgement. Cambridge University Press.

Mikhail, J., 2008. The poverty of the moral stimulus, in: Sinnott-Armstrong, W. (Ed.), Moral Psychology, Vol 1: The Evolution of Morality: Adaptations and Innateness. MIT Press, Cambridge, MA US, 353–359.

Mikhail, J., 2007. Universal moral grammar: theory, evidence and the future. Trends in Cognitive Sciences 11, 143–152.

Mikhail, J., 2007b. Moral Cognition and Computational Theory, in: Walter Sinnott-Armstrong (ed.), Moral Psychology, vol. 3: The Neuroscience of Morality: Emotion, Brain Disorders, and Development, Cambridge, MA: MIT Press, 81-92.

Mikhail, J., 2006. Moral Heuristics or Moral Competence? Reflections on Sunstein, 28 Behav. & Brain Sci. 557-558

Mikhail, J., 2002. Aspects of the Theory of Moral Cognition: Investigating Intuitive Knowledge of the Prohibition of Intentional Battery and the Principle of Double Effect (Georgetown Public Law Research Paper 762385). Washington, DC: Georgetown University Law Center. Available at http://ssrn.com/abstract=762385.

Mikhail, J., 2000. Rawls' Linguistic Analogy: A Study of the "Generative Grammar" Model of Moral Theory Described by John Rawls in "A Theory of Justice." (Phd Dissertation, Cornell University, 2000).

Mineka, S., Cook, M., 1988. Social learning and the acquisition of snake fear in monkeys, in: Zentall, T.R., Galef, B.G. (Eds.), Social Learning: Psychological and Biological Perspectives. Erlbaum, Hillsdale, NJ, 51–73.

Mineka, S., Davidson, M., Cook, M., Keir, R., 1984. Observational conditioning of snake fear in rhesus monkeys. J Abnorm Psychol 93, 355–372.

Moon, C., Lagercrantz, H., Kuhl, P.K., 2013. Language experienced in utero affects vowel perception after birth: a two-country study. Acta Paediatr 102, 156–160.

Morishita, H., Hensch, T.K., 2008. Critical period revisited: impact on vision. Current Opinion in Neurobiology, Development 18, 101–107.

Müller-Lyer, F.K., 1889. Optische Urteilstäuschungen, in: Archiv Für Anatomie Und Physiologie, Physiologische Abteilung. 263–270.

Murray, H.A., 1943. Thematic apperception test. Harvard University Press, Cambridge, MA, US.

Neander, K., 1996. Dretske's Innate Modesty. Australasian Journal of Philosophy 74, 258–74.

Need, A., Graaf, N.D. de, 2005. Conversion and switching between religious denominations in the Netherlands. Mens en maatschappij 80, 288–304.

Nelkin, D. K., & Rickless, S. C., 2015. So Close, Yet So Far: Why Solutions to the Closeness Problem for the Doctrine of Double Effect Fall Short. Noûs, 49(2), 376–409.

Nelson, K., 1973. Structure and Strategy in Learning to Talk. Monographs of the Society for Research in Child Development 38, 1–135.

Nelson, S.A., 1980. Factors influencing young children's use of motives and outcomes as moral criteria. Child Development 51, 823–829.

Ney, J.W., 1992. On Generativity: The History of a Notion That Never Was. Historiographia Linguistica 20, 341–454.

Nichols, S., 2005. Innateness and Moral Psychology,in: P. Carruthers, S. Laurence, and S. Stich (Eds.), The Innate Mind: Structure and Contents. New York: Oxford University Press New York.

Duda, Richard O., Peter E. Hart, David G. Stork, 2001. Pattern Classification, 2nd edition, Wiley Interscience.

Oxford, M., Spieker, S., 2006. Preschool language development among children of adolescent mothers. J Appl Dev Psychol 27, 165–182.

Pablo Pineda Ferrer 2014. (Online Document, Wikipedia, http://de.wikipedia.org/w/index.php?title=Pablo_Pineda_Ferrer&oldid=126148048. Acc. 30/04/14.)

Peck, J.R., 1994. A ruby in the rubbish: beneficial mutations, deleterious mutations and the evolution of sex. Genetics 137, 597–606.

Petitto, L.A., Marentette, P.F., 1991. Babbling in the manual mode: evidence for the ontogeny of language. Science 251, 1493–1496.

Piaget, J., 1948. The moral judgment of the child. Free Press, New York, NY, US.

Piattelli-Palmarini, M. (Ed.), 1980. Language and Learning: The Debate Between Jean Piaget and Noam Chomsky, Reprint. ed. Harvard University Press, Cambridge, Mass.

Pinker, 2004. Clarifying the logical problem of language acquisition. Journal of Child Language 31, 949–953.

Pinker, S., 1997. How the Mind Works. Penguin UK.

Pinker, S., 1986. Productivity and Conservatism in Language Acquisition, in: Demopoulos, W., Marras, A. (Eds.), Language Learning and Concept Acquisition: Foundational Issues. Ablex Publishing Corporation, Norwood, NJ, 54–79.

Pinker, S., 1984. Language Learnability and Language Development. Harvard University Press, Cambridge, MA.

Pinker, S., 1979. Formal models of language learning. Cognition 7, 217–283.

Pinker, S., Bloom, P., 1990. Natural language and natural selection. Behavioral and Brain Sciences 13, 707–784.

Popper, K.R., 1959. The logic of scientific discovery. Basic Books, Oxford, England.

Powell, D., Cheng, P., & Waldmann, M. R., 2014. How should autonomous vehicles behave in moral dilemmas? Human judgments reflect abstract moral principles, in: A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), Proceedings of the 38th Annual Conference of the Cognitive Science Society, 307–312. Austin, TX: Cognitive Science Society.

Powell, N.L., Derbyshire, S.W.G., Guttentag, R.E., 2012. Biases in children's and adults' moral judgments. J Exp Child Psychol 113, 186–193.

Prinz, J., 2006. The emotional basis of moral judgments. Philosophical Explorations: An International Journal for the Philosophy of Mind and Action 9(1), 29-43.

Prinz, J.J., 2007a. Is Morality Innate?, in: The Evolution of Morality: Adaptations and Innateness, Moral Psychology. MIT Press, Cambridge Mass., London, 367–407.

Prinz, J.J., 2007b. The emotional construction of morals. Oxford University Press.

Profet, M., 1992. Pregnancy sickness as adaptation: A deterrent to maternal ingestion of teratogens, in: Barkow, J.H., Cosmides, L., Tooby, J. (Eds.), The Adapted Mind: Evolutionary Psychology and the Generation of Culture. Oxford University Press, New York, NY, US, 327–365.

Pullum, G.K., Scholz, B.C., 2002. Empirical assessment of stimulus poverty arguments. The Linguistic Review 19, 9–50.

Pust, J., 2017. Intuition. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Summer 2017). https://plato.stanford.edu/archives/sum2017/entries/intuition/.

Rai, T.S., Holyoak, K.J., 2010. Moral principles or consumer preferences? Alternative framings of the trolley problem. Cognitive Science: A Multidisciplinary Journal 34, 311–321.

Ramscar, M., Yarlett, D., 2007. Linguistic Selfcorrection in the Absence of Feedback: A New Approach to the Logical Problem of Language Acquisition. Cognitive Science 31, 927–960.

Rawls, J., 1971. A theory of justice. Belknap Press of Harvard Univ. Press, Cambridge, Mass.

Ritov, I., Baron, J., 1994. Judgements of compensation for misfortune: The role of expectation. Eur. J. Soc. Psychol. 24, 525–539.

Ruti, M., 2015. The Age of Scientific Sexism: How Evolutionary Psychology Promotes Gender Profiling and Fans the Battle of the Sexes, 1st ed. Bloomsbury Press, New York.

Sag, I.A., Wasow, T., 2011. Performance-Compatible Competence Grammar, in: Borsley, R.D., Börjars, K. (Eds.), Non-Transformational Syntax. Wiley-Blackwell, 359–377.

Samuels, R., 2008. Is Innateness a Confused Concept?, in: Carruthers, P., Laurence, S., Stich, S. (Eds.), The Innate Mind, Vol. III, Foundations and the Future. Oxford University Press, New York, NY, US, pp. 17–36.

Sarva, A.R., 2003. The Concept of Modularity in Cognitive Science. Dissertation, Leland Stanford Junior University.

Scanlon, T., 2009. Moral Dimensions. Harvard University Press.

Scanlon, T.M., 2010. Moral Dimensions: Permissibility, Meaning, Blame, Reprint. ed. Harvard Univ Pr, Cambridge, Mass.

Schmidhuber, J., 2009. Driven by Compression Progress: A Simple Principle Explains Essential Aspects of Subjective Beauty, Novelty, Surprise, Interestingness, Attention, Curiosity, Creativity, Art, Science, Music, Jokes, in: Pezzulo, G., Butz, M.V., Sigaud, O., Baldassarre, G. (Eds.), Anticipatory Behavior in Adaptive Learning Systems, Lecture Notes in Computer Science. Springer Berlin Heidelberg, 48–76.

Schmitt, D.P., Pilcher, J.J., 2004. Evaluating evidence of psychological adaptation: How do we know one when we see one? Psychol Sci 15, 643–649.

Schwitzgebel, E., January 17, 2015. Moral principles and DDE judgments.[personal communication]

Schwitzgebel, E., 2013. The Splintered Mind: How Subtly Do Philosophers Analyze Moral Dilemmas? (Blog Entry, http://schwitzsplinters.blogspot.de/2013/12/how-subtly-do-philosophers-analyze.html. Acc. 30/04/2018.]

Schwitzgebel, E., Cushman, F., 2012. Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-Philosophers. Mind & Language 27, 135–153.

Schwitzgebel, E., Cushman, F.A., 2013. (Supplementary Online Material, http://www.faculty.ucr.edu/~eschwitz/SchwitzPapers/OrderEffects-SOM-110113.pdf. Acc. 25/03/2015.)

Scott-Phillips, T.C., 2010. Evolutionary psychology and the origins of language: (Editorial for the special issue of Journal of Evolutionary Psychology; on the evolution of language). Journal of Evolutionary Psychology 8, 289–307.

Seok, B., 2006. Diversity and Unity of Modularity. Cognitive Science 30, 347–380.

Shapo, M., 2003. Principles of Tort Law, 2 ed. West Academic Publishing, St. Paul, Minn.

Shenhav, A., Greene, J.D., 2010. Moral Judgments Recruit Domain-General Valuation Mechanisms to Integrate Representations of Probability and Magnitude. Neuron 67, 667–677.

Siegler, R.S., 1989. Mechanisms of Cognitive Development. Annual review of psychology 40, 353–79.

Simon, H.A., 1962. The architecture of complexity, in: Proceedings of the American Philosophical Society, 467–482.

Sosa, E., 2007. Intuitions: Their Nature and Epistemic Efficacy. Grazer Philosophische Studien, 74(1), 51–67.

Sousa, P., Holbrook, C., Piazza, J., 2009. The morality of harm. Cognition 113, 80–92.

Sperber, D., 2002. In defense of massive modularity, in: Dupoux, E. (ed.), Language, brain, and cognitive development: Essays in honor of Jacques Mehler, 47–57. Cambridge, MA: The MIT Press

Sperber, D., 1994. The modularity of thought and the epidemiology of representations, in: L. A. Hirschfeld & S. A. Gelman (Eds.), Mapping the mind: Domain specificity in cognition and culture (pp. 39-67). New York, NY, US: Cambridge University Press.

Sripada, C.S., 2008. Reply to Harman and Mikhail, in: Sinnott-Armstrong, W. (Ed.), Moral Psychology, Vol 1: The Evolution of Morality: Adaptations and Innateness. MIT Press, Cambridge, MA US, 361–366.

Sripada, C.S., Stich, S.P., 2005. A Framework for the Psychology of Norms, in: Carruthers, P. et al. (Eds.): The Innate Mind: Culture and Cognition, 280–301. Oxford University Press

Stich, S., & Doris, J. M., 2016, March 17. Moral Disagreement, Moral Realism and Moral Grammar. Draft to Circulate.

Stratton-Lake, P., & Zalta, E. N., 2016. Intuitionism in Ethics, in: The Stanford Encyclopedia of Philosophy (Winter 2016). https://plato.stanford.edu/archives/win2016/entries/intuitionism-ethics/

Street et al., 2009, Faith in Flux: Changes in Religious Affiliation in the U.S. Pew Research Center's Religion & Public Life Project. (Website with survey, possibly funded by non-scientific sources. urlhttp://www.pewforum.org/2009/04/27/faith-in-flux/)

Swallow, B.L., Lindow, S.W., Aye, M., Masson, E.A., Alasalvar, C., Quantick, P., Hanna, J., 2005. Smell perception during early pregnancy: no evidence of an adaptive mechanism. BJOG 112, 57–62.

Symons, D., 1979. The Evolution of Human Sexuality. OUP USA.

Thagard, P., 2012. Cognitive Science, in: Zalta, E.N. (ed.), The Stanford Encyclopedia of Philosophy (Fall 2015 Edition),https://plato.stanford.edu/archives/fall2014/entries/cognitive-science/

Thaler, R.H., 1999. Mental Accounting Matters. Journal of Behavioral Decision Making - J BEHAV DECIS MAKING 12, 183–206.

Thaler, R.H., 1988. Anomalies: The Ultimatum Game. Journal of Economic Perspectives 2, 195–206.

Thomson, J. J., 1985. The Trolley Problem. The Yale Law Journal, 94(6), 1395–1415.

Thomson, J.J., 1999. Physician Assisted Suicide: Two Moral Arguments. Ethics 109, 497–518.

Tomasello, M., 2009. Why We Cooperate. MIT Press.

Tooby, J., Cosmides, L., 1992. The psychological foundations of culture, in: Barkow, J.H. (Ed.), The Adapted Mind. Oxford University Press, New York, 19–136.

Treves, A., Naughton-Treves, L., 1999. Risk and opportunity for humans coexisting with large carnivores. J. Hum. Evol. 36, 275–282.

Trivers, R.L., 1972. Parental Investment and Sexual Selection, in: Campbell, B. (Ed.), Sexual Selection and the Descent of Man: The Darwinian Pivot. Transaction Publishers, New Brunswick (U.S.A.), 137–181.

Trivers, R.L., 1971. The Evolution of Reciprocal Altruism. Quarterly Review of Biology 46, 35–57.

Turiel, E., 1983. The Development of Social Knowledge: Morality and Convention. Cambridge University Press.

van Inwagen, P., 1997. Materialism and the Psychological-Continuity Account of Personal Identity. Philosophical Perspectives, 11, 305–319.

van Wouwe, J.P., van Gameren-Oosterom, H.B.M., Verkerk, P.H., van Dommelen, P., Fekkes, M., 2014. Mainstream and Special School Attendance among a Dutch Cohort of Children with Down Syndrome. PLoS One 2014; 9(3): e91737.

Waal, F.B.M. de, 2012. The Antiquity of Empathy. Science 336, 874–876.

Wagner, G. P., 1995. Adaptation and the modular design of organisms. In F. Morán, A. Morán, J. J. Merelo, & P. Chacón (Eds.), Advances in artificial life, 317–328. Berlin: Springer Verlag.

Waldmann, M.R., Nagel, J., Wiegmann, A., 2012. Moral judgment, in: Holyoak, K.J., Morrison, R.G. (Eds.), The Oxford Handbook of Thinking and Reasoning. Oxford University Press, New York, 364–389.

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, 364–389. New York: Oxford University Press.

Waldmann, M.R., Wiegmann, A., 2010. A Double Causal Contrast Theory of Moral Intuitions in Trolley Dilemmas, in: Ohlsson, S., Catrambone, R. (Eds.), Proceedings of the Thirty-Second Annual Conference of the Cognitive Science Society. Cognitive Science Society, Austin, TX, 2589–2594

Watt, H., 2017. Double Effect Reasoning: Why We Need It. Ethics & Medicine: An International Journal of Bioethics, 33(1), 13–19.

Wheeler, B., 2014. Laws, Natural Properties and Algorithmic Compression. (Online Document], http://philsci-archive.pitt.edu/10778/. Acc. 12/07/2014.)

Williamson, T., 2004. Philosophical “Intuitions” and Scepticism about Judgement. Dialectica, 58(1), 109–153.

Wimmer, H., Perner, J., 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children’s understanding of deception. Cognition 13, 103–128.

Yuill, N., 1984. Young children’s coordination of motive and outcome in judgments of satisfaction and morality. British Journal of Developmental Psychology 2, 73–81.

B Appendix

List of Evolutionary Psychologists' statements advocating cognitive modularity:

(Leda Cosmides and Tooby, 1994, p. 85):

"By establishing that domain-specific machinery is necessary to explain human cognitive performance, psychologists who advocate modular or domain-specific approaches have found themselves in an unanticipated situation. [. . .] From this emerging integrated perspective, the domain-specific mechanisms or modules cognitive psychologists have been studying can be readily recognized for what they are - evolved adaptations, produced by the evolutionary process acting on our hunter-gatherer ancestors." (referring to Cosmides & Tooby, 1987)

(Leda Cosmides and Tooby, 1994, p. 105):

"The evolvability considerations discussed earlier suggest that our species-typical architecture can be expected to contain not only a large number of domain-specific mechanisms that generate knowledge, but also a large number of domain-specific mechanisms that otherwise function to regulate and generate behavior [...]."

(Cosmides and Tooby 1997,"The Modular Nature of Human Intelligence", p. 81):

"[...]Such dedicated minicomputers are sometimes called modules. There is, then, a sense in which one can view the brain as a collection of dedicated minicomputers - a collection of modules. [...] So, more precisely, one can view the brain a s a collection of dedicated minicomputers whose operations are functionally integrated to produce behavior."

(Pinker, 1997, p. 27):

"The mind, I claim, is not a single organ but a system of organs, which we can think of as psychological faculties or mental modules."

(Pinker, 1997, p. 21):

"The mind is organized into modules or mental organs, each with a specialized design that makes it an expert in one arena of interaction with the world. The modules' basic logic is specified by our genetic program."

(Sperber, 2002, p. 1):

"I was arguing that domain-specific abilities were subserved by genuine modules, that modules came in all format and sizes, including micro-modules the size of a concept, and that the mind was modular through and through." (referring to Sperber 1994)

(Sperber, 2002, p. 3):

"Fodor is understandably reluctant to characterize a module merely as a "functionally individuated cognitive mechanism", since "anything that would have a proprietary box in a psychologist's information flow diagram" would thereby "count as a module" (Fodor 2000:56). If, together with being a distinct mechanism, being plausibly a distinct adaptation with its own evolutionary history were used as a criterion, then modularity would not be so trivial."

(Carruthers, 2006, p. 18):

"What emerges, then, is that there is a strong case for saying that the mind is very likely to consist of a great many different processing systems, which exist and operate to some degree independently of one another. [...] And all of these systems will need to be frugal (come to some result) in their operations, hence being encapsulated in either the narrow-scope or the wide-scope sense." "In its narrow-scope form, an encapsulated system would be this: concerning most of the information held in the mind, the system in question can't be affected by that information in the course of its processing. [...] In its wide-scope form, on the other hand, an encapsulated system would be this: the system is such that it can't be affected by most of the information held in the mind in the course of its processing." (most systems not processing most other systems' data). "Moreover, the processing that takes place within each of these systems will generally be inaccessible elsewhere." "It is certainly a form of massive modularity in the everyday sense that we distinguished at the outset."

Clark Barrett speaks up for massive modularity first (Barrett and Kurzban, 2006, p. 629):

"Because there are extensive and exhaustive reviews elsewhere (Carruthers, 2005; Coltheart, 1999; Pinker, 1997; Samuels, 1998, 2000; Segal, 1996), we review here only briefly our view of modularity. Proponents of massive modularity have offered several reasons for expecting mental processes to consist of multiple specialized systems, rather than a single general purpose one."

(Barrett and Kurzban, 2006, p. 630):

"Our position, then, is that functionally specialized mechanisms with formally definable informational inputs are characteristic of human (and nonhuman) cognition and that these features should be identified as the signal properties of "modularity.". But in his later work ((Barrett, 2009), Barret does not use the word "module" anymore; he instead uses the notion of "Enzymatic Computation" to refer to a new style of computation and later: "adaptations" for adapted mechanisms that work enzymatically:

(Barrett, 2009, p. 98):

"But, importantly, all aspects of information processing that exhibit design (that process information in a systematic, functional, goal-directed manner) are either adaptations, parts of adaptations, comprised of adaptations, or the result of adaptations."

(Barrett, 2005a, p. 285):

"What is likely to be incorrect is the premise that all systems composed of specialized devices must be Fodorean modular systems.";

(Barrett, 2005a, p. 1):

"I've argued to this point that the brain consists of a large number of specialized systems, or modules, with various functions associated with solving our ancestors' adaptive problems. I've argued that some of these systems feed information to one another and some don't. Some instances in which information doesn't flow are evident in [...] moral dumbfounding."

(Kurzban, 2012, p. 62):

"Of the many possible non-Fodorean architectures, one is explored here that offers possible solutions to computational problems faced by conventional modular systems: an 'enzymatic' architecture."). This is possibly a critique concerning the use of the notion "modular" (Barrett, 2005a, abstract).

Universality Chapter/Reviewed Empirical Papers

Binary Choice Dilemmas/Dilemmas with three reply options

Publication: Cushman et al. 2007, subjects: 5000 subjects covering 120 countries, but effect of nationality only tested for Australia, Brazil, Canada, India, US, UK

Question Type	Reply "Yes" in %	Reply "No" in %	Features of the harm
Switch Type action morally permissible?	89	n. a.	Far, no direct contact, side-effect
Push Type action morally permissible?	11	n. a.	Close, direct contact, means
Loop with Heavy Object action morally permissible?	72	n. a.	Far, no direct contact, side-effect
Loop action morally permissible?	56	n. a.	Far, no direct contact, means

Results/interpretation: The subjects judged actions in line with Doctrine of Double-Effect and with the Close Contact Harm Principle permissible more often than actions that are not.

Publication: Not published, but mentioned in Hauser et al. 2008, pp. 298/299. According to a preliminary online version of the paper [https://pdfs.semanticscholar.org/6a26/55ecdc14cb54abaccac234eeba544b279e22.pdf (1/8/17), p. 298]: Subjects: 15 Hadza from Tanzania.

Question Type	Reply "Yes" in subjects	Reply "No" in subjects	Features of the harm
Push Type action, question n. a., assumption: Permissible?		At least 12 (assumption)[187]	Close, direct contact, means
Switch Type action, question n. a., assumption: Permissible?	At least 12 (assumption)		Far, no direct contact, side-effect

Results/interpretation: A small sample of Hadza from Tanzania judged Switch Type actions more permissible than Push Type actions.

[187] As Hauser et al. wrote that 12 out of 15 Hadza judged "these cases as do web-savvy westerners",(Hauser et al. 2008, p. 136), I assume that 12 judged the Switch Type Action permissible AND the Push Type Action impermissible in a within-subjects design. It is, however, possible that 12 subjects judged the Switch Type Action permissible OR the Push Type Action impermissible in a between-subjects design. Considering the small (and uneven-numbered) subject group, I think a within-subjects design would have made more sense and is therefore more likely.

Publication: Mikhail 2002/2011, subjects: 40 adult volunteers whose majority were “United States citizens or members of other Western nations” (Mikhail 2002, p. 39) and 39 adult volunteers who had recently immigrated from China.

Question Type	**Reply “Yes” in %**	**Reply “No” in %**	**Features of the harm**
Condition 2 with 39 adults from China			
Switch Type action morally permissible?	79		Far, no direct contact, side-effect
Push Type action morally permissible?	14		Close, direct contact, means
Condition 1 with 65 adults, mostly from US and “other Western nations”			
Switch Type action morally permissible?	76		Far, no direct contact, side-effect
Push Type action morally permissible?	8		Close, direct contact, means

Results/interpretation: The subjects from China judged Push Type and Switch Type cases very similarly to subjects from “Western nations”, suggesting that they similarly judge according to the Close Contact Harm Principle and/or the Doctrine of Double-Effect.

Publication: Ahlenius and Tännsjö 2012, subjects: 1000 from each China, Russia and USA; reply options: "Yes", "No", "I don't know"

Question Type	**Reply "Yes" in subjects**	**Reply "No" in subjects**	**Features of the harm**
US inhabitants			
Should you do Switch Action?	81	13	Far, no direct contact, side-effect
Should you do Loop Action?	60	32	Far, no direct contact, means
Should you do Push Action?	39	56	Close, direct contact, means
Inhabitants of Russia			
Should you do Switch Action?	63	20	Far, no direct contact, side-effect
Should you do Loop Action?	54	23	Far, no direct contact, means
Should you do Push Action?	36	45	Close, direct contact, means
Inhabitants of China			
Would it be morally permissible to do Switch Action?	52	36	Far, no direct contact, side-effect
Would it be morally permissible to do Loop Action?	34	52	Far, no direct contact, means
Would it be morally permissible to do Push Action?	22	68	Close, direct contact, means

Results/interpretation: The preference order for the actions was the same for subjects from China, Russia and the USA; less subjects from Russia and even less from China generally judged harming one person permissible.

Publication: Gold et al. 2014, subjects: 55 "British" and 45 "Chinese" subjects; "[a]ll the British were native English speakers, none of the Chinese was."(Gold et al. 2014, p. 69)

"British" subjects			
Would you do Switch Type Action?	76.36	n. a.	Far, no direct contact, side-effect
Is it morally wrong for you to do Switch Type Action?	45.45	n. a.	Far, no direct contact, side-effect
"Chinese" subjects			
Would you pull the lever in Switch Type Action?	64.44	n. a.	Far, no direct contact, side-effect
Is it morally wrong for you to do Switch Type Action?	35.56	n. a.	Far, no direct contact, side-effect

Results/Interpretation: "British" and "Chinese" subjects judge this Switch Type Action more similarly than in the previous survey with the results for the "British" participants diverging from results of "Westerners" or British citizens in other surveys.

Within-subject designs

Publication: Cushman et al. 2006b, subjects: 332, 88%: English as primary language, most subjects from US, Canada and UK

Question Type	Scale labels	Mean value of judgments	Features of the harm
Action/Omission permissible?	From 1 (forbidden) over 4 ("permissible") to 7 (obligatory)	n. a.	various

Results/interpretation: The subjects judged actions in line with Doctrine of Double-Effect, the Close Contact Harm Principle and the Action/Omission Principle more permissible than those that are not.

Publication: Abarbanell & Hauser 2010, subjects: 30/30/31/29/30 per condition, with the first four groups subjects from a "rural" Mayan population and the last 30 subjects from a younger, more educated and "more urban" Mayan population.

Condition 1, subjects from "rural" population sample			
What do you think about Switch Type Action? (truck)	From 1 (very impermissible) to 5 (very good)	4.07	Far, no direct contact, side-effect
What do you think about Push Type Action? (truck)	"	1.63	Close, direct contact, means
What do you think about Loop Type Action? (truck)	"	1.90	Far, no direct contact, means
What do you think about Loop Type Omission? (truck)	"	1.90	Far, no direct contact, means, omission[188]
Condition 2, subjects from "rural" population sample			
What do you think about Switch Type Action? (boat)	From 1 (very impermissible) to 5 (very good)	4.03	Far, no direct contact, side-effect, action
What do you think about Switch Type Omission? (boat)	"	4.10	Far, no direct contact, side-effect, omission
What do you think about Loop Type Action? (boulder)	"	3.30	Far, no direct contact, means, action
What do you think about Loop Type Omission? (boulder)	"	3.60	Far, no direct contact, means, omission

[188] A person dying as a "means" is tricky when it comes to omissions, but "means" are a tricky concept in all cases. Here, I want to indicate that the sacrificed person's presence in front of the truck was necessary for stopping the truck in both the "means, action" and the "means, omission" case.

Condition 3, subjects from "rural" population sample			
What do you think about unintentionally Drop from arms Action? (peccaries)	3.44	Close, direct contact, side-effect	
What do you think about intentionally Drop from arms Action? (peccaries)	2.67	Close, direct contact, means	
Condition 5, dilemma texts identical with condition 1/Switch Type and Loop Type action, and (the last) two dilemmas from condition 2, subjects from "urban" population sample			
What do you think about Switch Type Action? (truck)	From 1 (very impermissible) to 5 (very good)	2.83	Far, no direct contact, side-effect
What do you think about Loop Type Action? (truck)	"	2.30	Far, no direct contact, means
What do you think about Loop Type Action? (boulder)	"	1.97	Far, no direct contact, means, action
What do you think about Loop Type Omission? (boulder)	"	2.87	Far, no direct contact, means, omission

Results/interpretation: A "rural" Mayan population judged actions in line with Doctrine of Double-Effect more permissible than those that are not, but did not prefer actions in line with the Close Contact Harm Principle and the Action/Omission Principle over those that are not. An "urban" Mayan population that was not tested for the Close Contact Harm Principle preferred both actions in line with the Doctrine of Double-Effect and, unlike the "rural" Mayan subjects, harmful omissions over harmful actions.

Publication: Schwitzgebel and Cushman 2012, subjects: 264 philosophers, 1801 non-philosophers for Switch vs. Loop Type Case[189] , and "324 'philosophers' (completed MA or PhD in philosophy), 753 'academic non-philosophers' (completed Master's or PhD not in philosophy), and 1389 'non-academics' (no Master's or PhD in any field)"(Schwitzgebel and Cushman 2012, p. 7), recruited by e-mails to universities and by academic blogs.

Question	**Equivalent ratings in %**	**Ratio who favours first scenario action in %**	**Ratio who favours second scenario action in %**	**Principle tested**
Switch Type Action/Push Type Action is 1 (Extremely Morally Good) to 4 (Neither Good Nor Bad) to 7 (Extremely Morally Bad)	Switch and Push Type Action: 50-78	Favour Switch Type Action: 22-50 (assumption)[190]	Favour Push Type Action:[191] 5	Doctrine of Double-Effect, Close Contact Harm Principle
Switch Type Action/Loop Type Action is 1 (Extremely Morally Good) to 4 (Neither Good Nor Bad) to 7 (Extremely Morally Bad)	Switch and Loop Action: 80-83	Favour Switch Type Action: 10-12	Favour Loop Type Action: 1 (Ethics PhDs, N=73) to 11 (non-Ethics PhDs, N=1992)	Doctrine of Double-Effect

Results/interpretation: 50% to 78% (half to a majority) of subjects judge actions that are in accordance with the Doctrine of Double-Effect AND the Close Contact Harm Principle equivalently as actions that go against both. 80-83% (a large majority) of subjects judge actions that are in accordance with the Doctrine of Double-Effect equivalently as actions that go against it.

[189] According to Eric Schwitzgebel's mail (E. Schwitzgebel, personal communication, January 17, 2015), where he refers to his 2012 article with Fiery Cushman. The article mentions "324 'philosophers' (completed MA or PhD in philosophy), 753 'academic non-philosophers' (completed Master's or PhD not in philosophy), and 1389 'non-academics' (no Master's or PhD in any field)"(Schwitzgebel and Cushman 2012, p. 7); the higher subject count in the mail might be due to participants who completed the survey after the paper draft was written or because replies that shifted in the unpredicted direction (favoured the harm used as a means over the foreseen harm) were excluded from the equivalence analysis in the paper.

[190] As the persons who favoured the Push Type action, in my reading of the paper, were excluded before calculating the equivalence ratings (see footnote 5), the people who did not vote equivalently favoured the Switch Type action.

[191] Excluded from equivalence analysis, ambiguous wording (possibly referring to scenarios 14-17), see Cushman and Schwitzgebel 2012, p. 11.

Publication: Cushman et al. 2007, subjects: 2612[192] out of 5000 visitors of Moral Sense website (see table p. 11) for within-session data and 207 for between-session data with on average 20 weeks between presentation of Loop and Loop with Heavy Object scenario; those 207 were a sample of the 2612 subjects who had only judged one of the two scenarios and either judged the Loop with Heavy Object scenario permissible or the Loop scenario impermissible.

Question	**Equivalent ratings in %**	**Ratio who favours any of the two actions in %**	**Principle tested**
Loop/Loop with heavy Object action permissible? Yes/No (within-session)	Loop and Loop with heavy Object Action: 94.2 (assumption, 100% minus ratio who favours either action)	5.8	Doctrine of Double-Effect
Loop/Loop with heavy Object action permissible? Yes/No (between-session)	Loop and Loop with Heavy Object Action: 67% (assumption, 100% minus ratio who favours either action)	33	Doctrine of Double-Effect

Results/interpretation: Only 5.8% (when tested in one session) and 33% (when tested in two sessions with 20 weeks in between on average) of subjects judged two scenarios differently of which one was in accordance with the Doctrine of Double-Effect and one was not. Hence, a large majority did not judge along the lines of the Doctrine of Double-Effect.

Publication: Schwitzgebel and Cushman 2015 (formerly: In draft), subjects: "497 respondents reporting graduate degrees in philosophy ("philosophers")[. . .]; and 921 respondents reporting graduate degrees, but not in philosophy ("non-philosophers")", recruited via e-mails collected from university department sites. (Schwitzgebel and Cushman 2015, p. 129)

Question	**Equivalent ratings in %**	**Scale labels**	**Mean value of judgments**
Flipping the switch [in Drop/Switch condition] is: (Supplementary Material of Schwitzgebel and Cushman 2015)	Drop/Switch: 46-70 (depending on presentation order)	From 1 (Extremely Morally Good) to 4 (Neither Good Nor Bad) to 7 (Extremely Morally Bad)	3.7 (Switch), 4.5 (Drop)

Results/interpretation: Roughly half of the subjects or more than half of the subjects (depending on which dilemma was presented first) in Schwitzgebel and Cushman 2015 did not judge the Switch- and the Drop Case differently, hence, depending on the presentation order, half of the subjects or only a minority judged according to the Doctrine of Double-Effect.

[192] Assumption, as both "Exposure to Moral Philosophy/Yes/No-replies and Gender/Male/Female add to 2612 and, considering the standards in this field of science at 2007, I am assuming a gender-binary design in this study. 19 scenarios were presented to the 5000 subjects, but only 4 evaluated for this paper, of which the Loop- and Loop with Heavy Object-scenarios were two (Cushman et al. 2007, p.5)

Acknowledgments

I want to cordially thank my supervisor Prof. Dr. Stephan Sellmaier and my thesis advisors Prof. Dr. Dr. Hannes Leitgeb, Dr. Michael Öllinger and Dr. Michael von Grundherr for all the constructive reviews and professional advice they gave me and to the Graduate School of Systemic Neurosciences for funding me.

And thank you so much to my family, friends and colleagues for your extraordinary support and for making this possible, particularly to Rahim, Rosi, Sarah, Zita, Felix, Fred, Johanna, Sebastian, Timmo, Tobi, David, Janett, Joachim, Feli, Steffen, James, Julia, Nico, Tobi, Jesse, Joerg, Frederic, Lena, Mauricio, Max, Sylwia, Taib, Giles, Daniel, David and Paula.

Eidesstattliche Versicherung/Affidavit

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation "Are Trolley Dilemma Judgement Mechanisms Evolutionary Adaptations?" selbstständig angefertigt habe, mich außer der angegebenen keiner weiteren Hilfsmittel bedient 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 "Are Trolley Dilemma Judgement Mechanisms Evolutionary Adaptations?" is the result of my own work and that I have only used sources or materials listed and specified in the dissertation.

2.5.2018

München, den/Munich, date

Lara Pourabdolrahim Seresht Ardebili