The University Of Sheffield.

"Ought Implies Can" as a Principle of the Moral Faculty

By:

Miklós Kürthy

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

The University of Sheffield

Faculty of Arts and Humanities

Department of Philosophy

January 2019

Abstract

This thesis is a contribution to moral psychology, the systematic study of the cognitive processes underlying moral judgment. It has two main aims. First, it attempts to show that the so-called Linguistic Analogy (LA) is the most productive framework for the study of moral cognition. As its name suggests, LA has it that moral psychology can be fruitfully modelled on linguistics, in particular on the Chomskyan project of detailing the architecture of the Language Faculty (FL)—a domain-specific cognitive system dedicated to language. This means, amongst other things, that the fundamental task of moral psychology is to discover and detail the representations, principles, and computational operations of the Moral Faculty (FM)—a domain-specific cognitive system that underpins the human capacity for moral judgment. Second, the thesis argues that the "Ought Implies Can" principle (OIC)—according to which if an agent *ought* to perform an action, then she *can* perform that action—is one of the central principles of FM, and proposes a novel account of how OIC is implemented in FM. To achieve this second aim, the thesis presents novel empirical evidence on intuitive moral judgments of ordinary people and argues that the best explanation of these data is to consider OIC as a processing constraint on the operations of FM.

Acknowledgements

Above all, I owe an enormous debt of gratitude to my amazing supervisors, Luca Barlassina and Stephen Laurence. I am lucky enough to have been Luca's very first PhD student. In the past 4 years or so, we would repeatedly have 3-4 hour discussions on some of the topics discussed in the thesis (as well as many others). Given Luca's wide-ranging knowledge and creativity, these discussions have always been the most intellectually inspiring moments of my years as a doctoral student. I only hope that the thesis reflects these discussions. When Luca and I meandered too far off the path, Steve was always there to nudge us back. His advice has always been extremely sensible and pragmatic. I must also thank both Luca and Steve for being available whenever needed, and, most importantly, for their patience.

I am very grateful to the Department of Philosophy for letting me be part of this excellent department and for generously funding my PhD. In particular, I am much indebted to Jenny Saul for her support with various administrative issues.

I want to thank Paulo Sousa and Holly-Lawford Smith for the collaboration, without which Chapter 3 would not have seen the light of day. Paulo invited me to Belfast no less than three times. During my visits, we had some extremely intellectually rewarding discussions, from which much of my thesis has benefitted. When I first arrived in Sheffield, Holly was one of the first friend/colleague who made me feel welcome. Our friendship has led to two collaborative papers and many interesting discussions.

I also have to thank some of my friends who in one way or another have contributed fruitfully to my work. They are (in no particular order): Dávid Metzger, who is responsible for some of the most exciting bits in the thesis: the figures in the first two chapters. I can't wait to return the favour Fecókam. Alex Duval, who has contributed to my thesis by lending me a laptop on which I ended up typing up the final two chapters, and more importantly by discussing some of the ideas in the thesis, and many others besides. I owe my greatest gratitude to José Avalos Díaz for drinking all my coffee and spoiling everything in Game of Thrones—and of course for being the best flatmate ever. Ryo Yokoe provided me with another laptop as well as endless conversations about topics of various degrees of seriousness. I thank Francesco Antilici for the peanuts, the gym tips, and all the heated arguments about conceptual content. I'm also grateful to Beth Flood for giving me inspiration for the examples in Chapter 1, Section 6.

The Postgraduate student community of the Philosophy Department has also been a most welcome influence on my PhD years. In particular, I should mention (in random order) Lewis, Anton, Damiano, Emma, James, Barend, Phil, Gonzalo, Neri, Tony, Armin and Ahmad. Other Sheffield friends have also

shaped my experience in many unforgettable ways: Josie, Christina, Elle, Ingrid, Alexandra, John, Luisa. Thanks for all the memories. Many thanks also to my Budapest friends for always being there for me during my all too infrequent visits: Soma & Kata, Dávid & Anna, Koffer & Andi, Ancsa & Zsombor, Kata és a Kubasch család.

Last, but by no means least, a massive thanks to my lovely family in Budapest: Mama, Apa, András, Ádám, Kristóf és Pau, Balázs és család. I miss you all!

Table of Contents

List of figures

List of abbreviations

AMG	Argument for Moral Grammar
DLPFC	Dorsolateral prefrontal cortex
DP	The Dual Process framework for the study of the mind
FL	Language Faculty
FM	Moral Faculty
FLB	Broad Language Faculty (or Language Faculty in the Broad sense)
FLN	Narrow Language Faculty (or Language Faculty in the Narrow sense)
LA	Linguistic Analogy
MM	Myopic module (see Chapter 2, Section 3.2)
NP, CP, VP	Noun Phrase, Complementiser Phrase, Verb Phrase
OIC	The *Ought Implies Can* principle
PC, PF, SP	Physical contact, personal force, spatial proximity (see Chapter 2, Section 3.2)
PH	Pragmatic Hypothesis (for OIC's descriptive success)
PLD	Primary Linguistic Data
PMD	Primary Moral Data
RT	Reaction Time
SH	Semantic Hypothesis (for OIC's descriptive success)
UG	Universal Grammar
UMG	Universal Moral Grammar
VMPFC	Ventromedial prefrontal cortex

Introduction

Moral judgment is a ubiquitous aspect of our everyday lives. We naturally judge our own actions and (perhaps even more so) those of others in terms of their perceived moral status, such as 'right', 'wrong', 'obligatory' or 'forbidden'. What are the cognitive processes that render us capable of making such judgments? This is the fundamental question of moral psychology, an interdisciplinary research project spanning philosophy, psychology, computer science, neuroscience, anthropology, and evolutionary biology. The present thesis is a contribution to this overarching research project. In it, I defend two main claims.

The first one concerns what the best way of doing moral psychology is. I defend the so-called Linguistic Analogy (LA), that is, I argue that moral psychology should be modelled on linguistics—in particular, on generative, Chomskyan linguistics. Accordingly, the main task of moral psychology should be conceived of as follows: to discover and detail the representations, principles, and computational operations of the Moral Faculty (FM), a specific-purpose cognitive system that underpins the capacity for moral judgment. Here is another way to articulate my first claim:

> (i) the human mind/brain contains a cognitive mechanism, FM, specialised for generating moral judgments. FM takes representations of actions as input (e.g., "John was bored, so pulled Mary's hair to have some fun"), and, on the basis of a body of rules or principles, it outputs a moral judgment ("John did something wrong");
>
> (ii) the main aim of moral psychology is to find out what the principles of FM are, how they are computationally implemented, and how they are acquired.

My second claim is not methodological, but substantive: I propose that "Ought Implies Can"—according to which if an agent ought to perform an action, then she can perform that action—is one of the fundamental principles of FM, and I put forward a novel account of how OIC is implemented in FM. To argue for this second claim, I adopt the following two-pronged strategy. First, I present the results of empirical studies I conducted in collaboration with Holly Lawford-Smith (Melbourne University) and Paulo Sousa (Queen's University Belfast), which indicate that the intuitive moral judgments of ordinary people conform to OIC. Second, I argue that the best explanation is to posit OIC as a processing constraint on the operations of FM.

The thesis is divided into two main parts. In the first one (Chapters 1 and 2), I argue in favour of LA as the best framework for the study of moral cognition. I begin by introducing and motivating LA (Chapter 1).

Then I show the superiority of LA over a very influential framework for moral psychology—namely, the framework behind Joshua Greene's dual-process theory—and I defend LA against popular objections in the literature (Chapter 2). The second part of the thesis (Chapters 3 and 4) turns to OIC. In Chapter 3, I present experimental evidence for the claim that OIC is descriptively adequate—that is, for the claim that it captures patterns of ordinary moral judgment. Finally, in Chapter 4, I argue that the best explanation of these results is that OIC is a processing principle of FM. Below, I provide a slightly more detailed summary of each chapter.

Chapter 1: The Linguistic Analogy as a Framework for the Study of Moral Cognition

The phenomenon of moral judgment raises many questions worth investigating. In Chapter 1, I first take account of some of these and zero in on the one that forms the topic of this thesis, namely, the question of what cognitive mechanisms and processes underlie moral judgment. Then, I introduce the framework, the Linguistic Analogy (LA), that I believe is best suited for the study of moral cognition. LA provides substantial hypotheses about the nature of moral cognition, the terms in which this phenomenon is to be accounted for, and the ways in which theories about it are to be evaluated. I review these hypotheses one by one, introducing the relevant arguments, concepts and distinctions along the way. Of particular importance is the Argument for Moral Grammar, according to which the productivity of moral judgment suggests that its explanation will have to appeal to abstract principles and representations over which those principles are defined. To a large extent, this argument sets much of the explanatory agenda of the present thesis. I review this argument in detail and argue for a modified version of it, which preserves the crucial part of its conclusion, namely, the theoretical necessity of appealing to principles. The discussion of what the form of these principles is constitutes the final part of this chapter.

Chapter 2: The Linguistic Analogy—Comparisons, Objections and Replies

In Chapter 2, I introduce a rival explanatory framework, the Dual Process framework (DP), and a specific theory that has been proposed from the perspective of this framework, namely, Joshua Greene's dual-process theory of moral judgment. I assess Greene's theory critically and point out some of its shortcomings. However, my central point here is not so much that Greene's DP theory *per se* is problematic, but that—in contrast to LA—the DP approach is an inadequate theoretical framework, in that it does not provide the tools, theoretical concepts, and distinctions needed for the study of moral cognition. In the final part of the chapter, I respond to similar concerns raised against my favoured framework. Since its inception, LA has been

subjected to serious criticism in the literature. I discuss and evaluate these from the point of view of the version of LA that I endorse in Chapter 1. The conclusion will be that none of the objections present a serious obstacle to LA.

Chapter 3: The "Ought Implies Can" Principle and Descriptive Adequacy

One of the important consequences of adopting LA as an explanatory framework is that explanation in moral psychology is expected to proceed by identifying the principles of FM. I propose that a viable strategy to individuate such principles is to consider ethical principles on which there is considerable philosophical agreement. A principle that until quite recently enjoyed a great deal of consensus in moral philosophy is the "Ought Implies Can" principle (or "OIC" for short), according to which if an agent *ought* to perform an action, then the agent in question *can* perform it. In this chapter, I propose that OIC is descriptively adequate, that is, it correctly describes the competence of an idealised individual in terms of his or her moral judgments, and thus it is plausible to take it as one of the principles of FM. To establish this, I discuss recent purported evidence that ordinary people do not reason in line with OIC. However, I show that much of this evidence is methodologically problematic and/or inconclusive. I also present the results of my own empirical studies, conducted in collaboration with Holly Lawford-Smith and Paulo Sousa, which provide evidence in favour of the descriptive adequacy of OIC.

Chapter 4: OIC Meets Cognitive Science—Hypotheses, Old and New

In this final chapter I provide a novel explanation of the descriptive adequacy of OIC. While standard approaches have attempted to explain OIC in terms of some aspects of language—namely, either its semantics or its pragmatics—, I argue that such accounts are on the wrong track. Instead, I propose that OIC is grounded in the nature of the computational operations peculiar to FM. I propose two types of ways in which this claim can be cashed out based on the approach defended in the first half of the thesis. According to the first, OIC is a synchronic processing principle, that is, it reflects the online operations of FM. According to the second, OIC is a diachronic acquisition principle constraining the types of moral rules (or "faculty principles") that are acquirable by FM. Given the currently available evidence, it is difficult to judge how these proposals will fare in the future. Nevertheless, they provide viable alternatives to the accounts available in the literature.

Chapter 1: The Linguistic Analogy as a Framework for the Study of Moral Cognition

Overview

The present thesis addresses aspects of moral-deontic thinking from the point of view of the Linguistic Analogy, a particular way of pursuing the inquiry into the nature of moral cognition, famously suggested by John Rawls (1971) and most comprehensively articulated by the philosopher and legal scholar, John Mikhail (Mikhail, 2000, 2009, 2011). According to the Linguistic Analogy, the study of moral cognition can be fruitfully modelled on the study of language as undertaken in the generative tradition in linguistics, inaugurated by Noam Chomsky in the 1950s. This chapter is devoted to illustrating the main aspects of the Linguistic Analogy as well as defending its viability as an appropriate and worthwhile approach to the study of moral psychology.

The chapter is structured as follows. First, in Section 1, I introduce the variety of questions that may come to mind when terms such as "morality" or "moral judgment" are mentioned and single out the problems that will be central in this thesis—namely, questions concerning moral cognition—whilst mentioning a few that will not be addressed at all—namely, questions concerning normative ethics and metaethics. I defend the position, which may be referred to as the "independence thesis", according to which the kind of inquiry I am interested in pursuing can be done so without confronting and taking sides in normative and metaethical controversies. Second, in Section 2, I introduce the Linguistic Analogy as a substantive framework for the study of moral cognition, focusing on its assumptions concerning the character of the explanandum, the nature of explanation, and the empirical standards of evaluation of theories. As for the first, in Section 3, I argue that the proper explanandum for moral psychology is moral competence (or I-morality), intended as a computational system that underlies the productivity of moral cognition. In Section 4, I propose that the natural explanation for such an explanandum is computational in nature. In Section 5, having taken the core arguments of the Linguistic Analogy seriously, I address the question of how to deal with the core implication of the productivity argument as outlined earlier in Section 3, namely, the idea that the notion of moral principle should be at the heart of theories of moral psychology. In the final section, I discuss what the Linguistic Analogy has to say concerning the ways in which theories in moral cognition (and computational theories more generally) ought to be evaluated.

1. The multiplicity of questions about morality

Picture Roy, who is having a stroll in the forest on his own. Soon, he happens to stumble by a charming little pond. To his dismay, he notices a small girl who appears to be drowning. There is no one else around, and the closest village is miles away. Roy doesn't know who the girl is; he has never seen her in his life. Even in the absence of considering any further details of the case, one thing seems crystal clear: Roy *must* do something. He must help the little girl. It is his moral obligation.

There are many deep questions raised by this little story. For example, it seems to be more than simply a matter of opinion that Roy has an obligation to help the little girl: it appears *true*. But what makes this the case? If I say "Paris is the capital of France", there is a simple way of verifying or falsifying my claim. Is there anything analogous with deontic or moral claims, such as "Roy must help the little girl"? And, relatedly, how do we have access to truths of this kind, assuming they exist? More generally, how can it be a matter of truth rather than opinion that Roy *should* do certain things and refrain from doing others in a way that is independent of what he (or anyone else for that matter) wants to achieve? Shifting to a more practical point of view, if Roy has this obligation, he presumably also has others. And so do we. But what are they? And, perhaps more urgently, how do we find out about them? Can we deduce them from some first principles? Or should we just follow our emotions or intuitions? There is no shortage of attempts to tackle these and other related problems of course. Such attempts are subsumed under the scope of metaethics, which investigates questions about the semantics, epistemology and metaphysics of morality, or normative ethics, which deals with questions about what we ought to do.

The thesis, however, will focus on a third type of question. Namely, what is it about our *minds* that makes it the case that we can think in such terms—namely, in terms of what is *right* and what is *obligatory*—to begin with? To be more precise, I will be concerned with our capacity to assign moral-deontic value to actions.[1] (If I use terms such as "moral judgment" along the way, this is what I mean.) There is good reason to believe that this capacity is underlain by a dedicated cognitive system of sorts due to some properties of moral judgment, such as its early emergence in ontogeny (Baillargeon et al. 2015; Hamlin, 2015), its universal occurrence (Brown, 1991), the automaticity with which such judgments are made (van Lier et al., 2013), aspects of its neural organisation (Zinchenko & Arsalidou, 2018), and aspects of its content (e.g. Turiel, 1983; Mikhail,

[1] By "deontic" here I mean the standard set of deontic concepts as interdefined by the deontic hexagon (see e.g. McNamara, 2018; Joerden, 2012). That is, OBLIGATORY, FORBIDDEN, and DISCRETIONARY, plus the negation of each. By "moral", I mean simply that the relevant obligations and prohibitions are intuitively seen as of the "moral" type, whatever exactly that comes to (see more on which in the next chapter).

2011; Haidt,2013). The present inquiry concerns the description of some features of this putative system, which I will sometimes refer to as the moral faculty (or FM) from a cognitive science perspective, or, more specifically, from the perspective of the Linguistic Analogy (or "LA" for short). More generally, this thesis is a work in moral psychology—it is a study of some aspects the psychological capacity for morality.

The presumption that cognition about moral issues can be legitimately studied from a psychological perspective is not controversial. In fact, this presumption is by no means a recent invention: it is deeply rooted in the history of the study of morality. To take but one example, some philosophers of the Enlightenment, collectively referred to as "the moral sense philosophers", such as Adam Smith, Francis Hutcheson, or the 3rd Earl of Shaftesbury, speculated amply about the psychology of moral judgment as well as its origins. In fact, some of the conclusions of contemporary empirical studies were anticipated by these thinkers, such as the automaticity or involuntariness of moral evaluation: we cannot but see the world in moral terms, and our judgments are often unaffected by our explicit ideas about morality. It is worth mentioning that based on such considerations, these philosophers concluded that humans are endowed with a moral sense analogous to our linguistic and visual cognitive capacities.

What may be slightly more controversial, however, is whether we can do moral psychology *without saying anything substantive about metaethics or normative ethics* at all. It would be concerning if taking a position in respect of the problems raised by these disciplines were a necessary preliminary step in the study of moral psychology, partly because there is nowhere near a general agreement about some of the most basic ones of these, including the questions raised above. However, I believe there is no reason to worry. Take the psychology of vision, which is a success story in cognitive science if there ever was one. Notice that philosophers still disagree about the epistemology and metaphysics of colours. This is not a problem *for psychologists*, however. They can detail the mechanisms responsible for colour perception even though they are unsure as to what colours really *are*. This analogy—along with the assumption, put forward above, that moral thinking is a natural phenomenon—suggests that not taking sides in metaethical or normative issues need not prevent one from tackling questions in moral psychology: the explanandum for moral psychology seems independent from substantive views in metaethics and normative ethics (as also noted by the moral sense theorists). Consequently, the success of the former enterprise is not predicated on success of the latter ones. For better or worse, I shall proceed under this additional—somewhat stronger—presumption, which we may refer to as "the independence thesis".

2. Two approaches to the Linguistic Analogy

At its simplest and most straightforward, the Linguistic Analogy is an approach to studying moral cognition with tools borrowed from the systematic study of language as pursued in the framework of generative linguistics launched and developed by Chomsky and his colleagues in the decades following the 1950s.

Chomsky's approach was novel on account of being explicitly mentalistic in contrast to the then prevalent behaviouristic approach to language: Chomsky contended that it is facts about the mind that explain our ability to speak and understand language. Mentalism about language has two important consequences. First, since the mind is by all accounts finite, and since we nevertheless have the capacity to understand, evaluate and produce any number of novel sentences, knowledge of language has to consist of more than a list of words and sentences: it has to contain combinatorial principles that are capable of generating all that variety. I shall refer to this as the *Argument for Mental (Linguistic) Grammar*. Second, children have no trouble acquiring these principles despite not being explicitly taught. But since children succeed in selecting the actual rules and principles of their language despite the fact that many different generalisations would be equally consistent with the linguistic data that is available to them (i.e. the so-called *primary linguistic data* or PLD), children must have an innate component that forms the basis of the process of language acquisition. Chomsky referred to this innate component as Universal Grammar (or UG). This, in a nutshell, is the *Argument for Universal Grammar*.

One approach to the Linguistic Analogy is thus to construct analogous arguments in the case of moral cognition. First, it does not seem to be the case that the number of actions and situations we can evaluate in moral terms has an upper limit. Second, the acquisition of moral rules, principles, and distinctions often seem to proceed without instruction or conscious awareness, on the basis of the arguably fragmentary data available to the child (i.e. what we may refer to as the "primary moral data" or PMD). On this approach to the Linguistic Analogy, the success of LA is predicated upon the extent to which these arguments are successful.

A second way to understand the Linguistic Analogy is to see it as a heuristic tool for the study of moral cognition. The strength of this approach can be understood as varying along a continuum, the extremes of which correspond to a weaker and a stronger version of this view. The weaker version merely assumes that some of the technical concepts and distinctions made by generative linguists might help make sense of the

data in moral psychology.[2] The stronger version has it that the parallel will be robust; on this view, moral cognition is expected to be much like language relative to some relevant dimensions of similarity. Relatedly, all or most distinctions made by Chomsky and his followers will have a counterpart in the moral domain.[3]

To summarise, the Linguistic Analogy can be conceived of as an endorsement of the applicability in the case of moral cognition of (a) one or both of the two central arguments of Chomsky's programme or (b) some or most of the distinctions made by Chomsky in respect of language. In fact, a combination of (a) and (b) is possible as well, and this is the approach to moral psychology that I favour. In particular, I will endorse the following: (i) the mentalistic (*qua* computationalist and representationalist) approach to morality, (ii) a version of the Argument for Moral Grammar, (iii) the relevance of the competence-performance distinction as well as (iv) the relevance of the perception-production distinction. I introduce these elements in sections 3-4 below. I will argue that the joint endorsement of these distinctions and arguments warrants an approach to the study of moral cognition that involves the search for moral principles and the representational repertoire they are specifiable in terms of. Towards the second half of the chapter (Section 5), I will go into some detail as to how "principle"—one of the core terms of LA—ought to be understood in the context of moral psychology. Finally, in Section 6, I illustrate how a theory of moral cognition (and more loosely, hypotheses about particular moral principles) ought to be evaluated as suggested by the Linguistic Analogy.

3. The character of the explanandum

My reading of the Linguistic Analogy has it that (i) the study of moral cognition needs to subscribe to mentalism (in the form of computationalism and representationalism), (ii) the capacity for moral judgment is in some sense combinatorial, (iii) moral competence should be distinguished from moral performance, and

[2] Here is a sample of expressions of this weaker version of LA from the literature: "On the weak version, the Linguistic Analogy is merely a heuristic for posing the right sorts of questions about the nature of our moral competence. On this version, it matters little whether morality works like language. What matters is that we ask about the principles that guide mature competence, work out how such knowledge is acquired, understand whether and how competence interacts with both mind internal and external factors to create variation in performance, and assess how such knowledge evolved and whether it has been specially designed for the moral sphere" (Hauser et al., 2008a, *p.* 139). "At the very least, analogies [such as LA] suggest questions and shape the direction of research" (Roedder & Harman, 2010, *p.* 273). "At least in terms of its main questions, fundamental conceptual distinctions, key methodological commitments, and overarching theoretical goals, the theory of moral cognition would benefit from drawing on the basic terminology and theoretical apparatus of Universal Grammar" (Mikhail, 2011, *p.* 308).

[3] "On the strong version, language and morality work in much the same way: dedicated and encapsulated machinery, innate principles [guiding] acquisition, distinctions between competence and performance, inaccessible and unconscious operative principles, selective breakdown due to [selective brain] damage, and constraints on the evolvable and learnable languages and moralities" (Hauser et al., 2008a, *p.* 139).

(iv) the problems of moral perception and moral production should be kept distinct. Here, I argue for each claim in turn.

3.1. Mentalism

3.1.1. From behaviourism to cognitive science

It is a characteristic assumption in cognitive science that "intelligent" behaviour can only be properly explained by attributing internal mental states and processes to the organism whose behaviour we are interested in making sense of. For historical reasons, this view is often contrasted with behaviourism, the proponents of which either deny the reality of internal mental states altogether or else they argue that mental states cannot be the proper object of scientific investigation, despite the indubitable success of mentalistic explanation both in the domain of scientific psychology and in ordinary reasoning about behaviour. Unlike behaviourism, mentalism assumes that the most straightforward, and indeed the best, account of the success of explaining and predicting intelligent behaviour by reference to internal mental states is that such mental states do in fact exist and play a genuine causal role in the generation of behaviour.

To take a toy example, consider a very basic aspect of our mathematical competence: keeping track of the numerosity of a small number of objects. For instance, when we see two identical objects successively disappear behind a screen, we can form the expectation that there are (at least) two objects behind the screen. How do we go about explaining this apparently simple inference, which, as a matter of fact, even 5-month-old infants are capable of (*cf.* e.g. Wynn, 1992)? Mentalism (qua computationalism and representationalism) suggests that we entertain mental representations of the relevant objects and manipulate them according to some laws or regularities, such as the one according to which "1 + 1 = 2".

The computational premise in the Chomskyan type of mentalism assumes that thinking is computation, where computation is understood as a form of symbol manipulation, the symbols being the mental representations over which cognitive processes (such as additive inferences) are defined. This traditional picture entails that the explication of cognition proceeds along the following lines: identify the computational principles and processes that the relevant kind of thinking (such as addition) consists in and identify the representations which those computations presuppose.

This much is little more than the recapitulation of an orthodoxy in the philosophy of cognitive science. The reason why I highlight mentalism is simple: even though practically everybody agrees that moral cognition consists in some psychological processes, it is also the case that researchers sometimes overlook that psychological processes are manipulations of representations. But by overlooking this, they fail to focus on

the proper task of moral psychology: studying what representations are involved in moral thinking and how, that is, according to what principles, these representations are manipulated. In contrast, the Linguistic Analogy makes these important assumptions explicit—unlike other popular frameworks, such as the dual-process theory of moral judgment (discussed in the next chapter).

3.1.2. E vs. I

The Chomskyan shift from behaviourism to mentalism about language can be illustrated in terms of Chomsky's distinction between E-language and I-language (Chomsky, 1986). The former had been the traditional object of study of pre-Chomskyan linguistics, which took language, albeit mostly implicitly, as an external entity independent of the workings of individual minds. Thus, language would be viewed either as some sort of an abstract entity or to exist (in some unspecified form) in the community. Chomsky instead argued that the primary object of inquiry for linguistics is the cognitive system rooted in individuals mind/brain that makes them capable of linguistic communication. In Chomsky's view, the former conception of language, to the extent it is a coherent conception at all, should be regarded as parasitic on the latter one. The former he referred to as *E-language*, or "externalised language", the latter *I-language*, or "internal language".

In fact the "I" in *I-language* stands for three features: internalised, individualised, and intensional (Mikhail, 2011; Chomsky, 2018). I-language is *internalised* because it is taken to be internal to the mind (rather than being an emergent social or cultural phenomenon). I-language is *individualised* on account of the assumption that the explanation of an individual's capacity to process (speak and understand) his or her language is logically prior to that of a larger group of individuals (such as a dialect group). Finally, I-language is *intensional* in the sense that properly characterising the computations performed by the mind when processing language (or some narrower task, such as mapping a set of phonemes to syntactic structures), we are interested in those computations as they are performed rather than the way in which they could be performed given functional/extensional equivalence, that is, identical input-output mapping.

The analogy with morality is tempting: we can distinguish between, on the one hand, the cognitive structures that enable an individual to cognise in moral terms and, on the other hand, morality in the sense of culturally shared codes of conduct. That is, we may distinguish I-morality and E-morality, respectively (Mikhail, 2011, *pp*. 24-26). In contrast to an I-morality, that is, a person's cognitive structures and mechanisms that subserve his or her moral judgments, an E-morality may be understood as the property of a community made up of individuals with more or less identical I-moralities, such as libertarian Americans, for

example (to the extent that they can be said to be a relatively "morally homogeneous" group).[4] Since it may easily be the case that it is more or less the latter sense that many of us have in mind when we talk about morality, the distinction between I- and E-moralities is an important one worth bearing in mind. In keeping with the mentalist approach of Chomsky as well as modern cognitive science, the proper explanandum for moral psychology must be I-morality.

There is one point that is worth clearing up before we go on. One may assume that if the primary goal of moral psychology is to describe and explain an I-morality, then there is no need to investigate populations, just one individual will do. This argument is based on a mistaken conception of the utility of idealisation in science. As Chomsky has pointed out, from a scientific perspective, an individual's language, along with all of its idiosyncrasies, is utterly uninteresting *per se*. The need to describe an I-language first and foremost derives from the mentalist premise. According to Chomsky, the ultimate goal of linguistics is to account for how humans' mind/brain makes it possible for individuals to acquire a language (in the sense of I-language) based on the available data (i.e. the PLD) they are exposed to in the course of ontogeny. But, clearly, this can only be done once we have an accurate description of the system that is acquired. Since that system is a mental system, we need to understand it in mentalistic terms. Studying what patterns of judgments populations of individuals make is a necessary step towards specifying the robust elements of I-languages: only phenomena exhibiting sufficient stability and universality (at least in terms of a chosen population) are worth investigating seriously.

Of course, there is a significant assumption lurking in the background, namely that there will be stability in terms of I-moralities/languages in the population. This assumption is not seriously questioned in modern linguistics. An analogy may help understand the above point. Consider our model of the human skeleton. It is a structure of immense complexity, involving hundreds of bones (206 in the adult human body), and three different types of joints and cartilages. When we look at a model skeleton, what is embodied in the model is a generalised understanding of an idealised human (in terms of its skeletal properties). There is no assumption to the effect that the model perfectly represents all humans, nor in fact any one human in particular. What is important, though, is that it captures the universal aspects of the structure of the human skeletal system, that is, those that are widely shared across individuals. In this respect, a model is a collection of statistical averages

[4] This is just one reading of what an E-morality may be. In Chomsky's original formulation, the concept of E-language subsumes formal conceptions of language as well, such as Lewis's (1975), whereby language is seen as an abstract mapping between sentences and meanings. The crucial property of an E-language, however, is that it is "understood independently of the properties of the mind/brain" (Chomsky, 1986, p. 20). Similar considerations apply in the case of E-moralities. In any case, here the focus in on I-morality, not on E-morality, so—in keeping with the spirit of the independence thesis—we need not pronounce definitively on this issue.

across a population. For this reason, it is important that no idiosyncratic features are built into the model, such as supernumerary digits, for instance.

3.2. Productivity in moral cognition: The Argument for Moral Grammar

3.2.1. The Argument for Mental (Linguistic) Grammar

Consider the following sentence:

> The reason why I am writing this chapter is that my best friend, the alcoholic aardvark, who was calculating the derivative of the natural logarithm on my brother's mantelpiece yesterday, told me to do so

I am fairly sure that neither the reader nor anyone else has ever heard, read or uttered this sentence before. Nevertheless, and this is one of the fascinating facts about language and thought, understanding the sentence will not have posed any substantial difficulties. This sentence gives testimony to the unbounded capacity of language to express an incalculable—strictly speaking, infinite—number of thoughts.[5] This unbounded capacity, referred to as the *productivity* of language, is a crucial explanandum for theories of (I-)language.

Let us consider a short illustration of the productivity of (the English) language. Take the sentence:

(1) [[the aardvark] [that [solved [my homework]]]] [ate [my breakfast]][6]

Traditional grammatical analysis interprets this sentence as consisting of a subject ("the aardvark that solved my homework") and a predicate ("ate my breakfast").[7] At the syntactic level of analysis, the subject is taken up by what linguists refer to as a noun phrase (NP). Crucially, the subject NP of (1) contains a relative clause, "that solved my homework" (technically a complementiser phrase or CP) which itself contains an NP: "my homework". This is an example of recursive embedding, because a phrase of a given type (NP in our case) is embedded in a phrase of the same type. Infinity follows from properties of I-languages such as recursive embedding. (See Jackendoff, 2002, *pp*. 38-67 for other examples of linguistic productivity.)

One significant technical consequence of the productivity of language is that, contra some theories of language and the mind (such as linguistic structuralism, behaviourism or connectionism), language competence (I-language) cannot consist in a mere enumeration of grammatical sentences: to account for the

[5] This formulation is somewhat loose, because it could be the case that the number of possible *sentences* is infinite, while the number of *thoughts* is not. But it is very probably not the case, as I will suggest below.

[6] The square brackets indicate phrase boundaries.

[7] In modern syntax, the predicate is the verb, the subject and object(s) being its arguments. We can ignore this complication for present purposes.

ability to understand not to mention produce a potentially infinite number of novel sentences, the theorist needs to posit (a) a lexicon, that is, a store of a large but finite number of basic lexical elements out of which sentences can be built and, crucially, (b) a closed set of *rules* or *principles of combination* that specify how these elements can be assembled to form *grammatical* sentences.[8] Additionally, notice that it is not sufficient for the lexicon to consist of a mere list of *words*: it is also necessary to specify these words in terms of syntactic (as well as other linguistic-functional) categories, because the rules of syntax are defined over these representations (and *mutatis mutandis* for morphological and other types of linguistic rules), rather than the words themselves (see *fn*. 8); this is what is referred to as the *structure dependency* of linguistic principles, that is, the property of such principles to make reference to structural constituents—rather than lexical items or ordinal positions—in a sentence. Clearly, if the rules were defined over the words themselves, then there would be as many rules as there are words—not a desirable or tenable theoretical position.

The structure of the foregoing argument has been something along the following lines: a fundamental property of language, namely productivity, that is, the ability to parse and produce a potentially infinite number of novel expressions, places certain constraints on theories of language, such that they are required to account for this property on pain of being *prima facie* disfavoured. The best explanation of productivity—and in fact the only one we are aware of—is that there is a *mental-linguistic grammar* specifying a set of representations (such as NPs and CPs) and the syntactic rules that are defined over them (*cf.* Jackendoff, 1994, Chapter 2), which jointly have the power of accounting for linguistic productivity, at least in principle.[9]

3.2.2. The Argument for Moral Grammar

As suggested by the Linguistic Analogy, there is an analogous case to be made in the domain of moral cognition. The guiding thought is that moral judgment appears potentially infinite too, suggesting that some sort of productivity might also characterise the moral faculty. If there are productive processes in moral cognition, this would indicate the existence of a "moral" *grammar*, that is, a set of generative principles from

[8] Examples of such a rule are the so-called *rewrite rules* (or phrase structure rules) specifying the legitimate (i.e. grammatical) ways in which phrases can be formed out of the syntactic categories of the constituents. So, for example, to capture our example in the text (1), an NP rewrite rule is: "NP → NP CP", that is, an NP may consist of the concatenation of an NP and a CP.

[9] Jackendoff points out, quite rightly in my view, that the word *representation* is somewhat unhelpful here (as well as other related intentional terms, such as *knowledge* or *information*). For example, an NP is not a representation *of* anything: it is simply a mental symbol that the theorist needs to posit to explain certain mental-linguistic phenomena (e.g. Jackendoff, 2007, *pp*. 5-7). For this reason, Jackendoff has suggested alternative terms, such as "mental structure". Though I am sympathetic to Jackendoff's terminological worries, I shall stick to the more traditional nomenclature (*cf.* also Chomsky, 2000, Chapter 6; Collins, 2004, *pp*. 512-3).

which particular moral judgments can be derived in more or less the same way as whether particular sentences are grammatical can be derived from principles of syntax.

The argument for a principles-based moral cognition is by no means novel. For example, Hume made the following remarks in his *Treatise of Human Nature*:

> "It may now be ask'd *in general*, concerning this pain or pleasure, that distinguishes moral good and evil, *From what principles is it deriv'd*, and whence does it arise in the human mind? To this I reply, *first*, that 'tis absurd to imagine, that in every particular instance, these sentiments are produc'd by an *original* quality and *primary* constitution. For as the number of our duties is, in a manner, infinite, 'tis impossible that our original instincts should extend to each of them, and from our very first infancy impress on the human mind all that multitude of precepts, which are contain'd in the completest system of ethics. Such a method of proceeding is not conformable to the usual maxims, by which nature is conducted, where a few principles produce all that variety we observe in the universe, and every thing is carry'd on in the easiest and most simple manner. 'Tis necessary, therefore, to abridge these primary impulses, and find some more general principles, upon which all our notions of morals are founded." (Hume, 1739, *p*. 473)

Hume's argument is an ostensibly empiricist one: rather that innately possessing numerous (or an infinity of) principles that make us capable of morally appraising novel actions or situations, it is more plausible to posit only a few of such principles to account for this unbounded capacity. It is also possible to turn the argument on its head, thereby producing a rationalist version of it: it is wildly implausible to assume that we have been exposed to a "replica" of each situation we are nevertheless readily capable of evaluating in moral terms. Indeed, completely novel situations present themselves on a daily basis and are judged about as effortlessly as in the case of the perception of novel sentences such as the one about the alcoholic aardvark—or indeed all the others in this thesis. Irrespective of which version of the argument one is in favour of, the pertinent point here is that Hume recognised the productivity of moral judgment as an important explanandum for a naturalistic theory of moral cognition and he also made a defeasible inference to the existence of (a limited number of) principles guiding moral judgment. On this view, just as in the case of language, what is required is a set of principles from which the judgments can be derived; grammaticality judgments in the case of language, permissibility judgments in the case of the moral sense.

It is probably not a coincidence that the notion that moral judgment may be in some sense productive became the subject of considerable discussion in the wake of the inception of the burgeoning literature on the "trolley" dilemmas, in which the question is whether it is permissible to interfere in various ways to stop

a trolley from killing (usually) five people thereby killing one (see the next chapter, Section 1 for more). Two notable observations with respect to the trolley dilemmas in this context are the variety of novel situations the literature on them engendered (see e.g. Greene et al. 2009; Mikhail, 2011) and the obvious novelty (and artificiality) of each of the cases from the point of view of ordinary experience. Nevertheless, evaluating the dilemmas—unlike explaining the principles on which such evaluations are based—rarely produces any notable challenge for experimental subjects, and there is remarkable consistency and systematicity in the results.

To return to the main point, what is referred to by Mikhail as the *Argument for Moral Grammar* (AMG), which is more or less explicit in the above paragraphs, holds that, parallel to the case of I-language, I-morality contains (a) a finite and presumably small set of principles as well as (b) a finite although quite possibly large set of representations over which the principles are defined. It is worth quoting Mikhail for his formulation of AMG:

> "The novelty and unboundedness of moral judgment, together with its frequent predictability, stability, and other systematic properties, implies the existence of a moral grammar, because to explain how an individual is able to project her finite experience to entirely new cases, we must assume that she is guided, implicitly, by a system of principles or rules. Without this assumption, her ability to make these novel judgments – and our ability to predict them – would be inexplicable" (Mikhail, 2011, *p*. 72).[10]

Thus far, I believe that Mikhail's reasoning is unassailable. However, elsewhere, he writes that "the properties of moral judgment 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" (Mikhail, 2007, *p*. 144; emphasis added). It is apparent from this latter quote that Mikhail takes the Argument for Moral Grammar to be more or less exactly parallel to the Argument for Mental (Linguistic) Grammar (as clearly articulated by Jackendoff, 1994, *pp*. 8-20, for instance). That is, Mikhail believes that the bottom line of the argument is that moral competence (I-morality)

[10] Jackendoff also proposes a similar but more general argument: "The basic observation is that humans manage to participate in and understand an unlimited number of social interactions, most of which they have never encountered before in exactly the same form. The ability to interact socially must therefore involve a combinatorial system of principles in each individual's mind/brain, which make it possible to build up understanding of particular situations from some finite stock of stored elements" (2007, *p*. 149).

consists in a *generative* capacity, and the principles of I-morality are analogous to the principles of syntax.[11] I doubt, however, that this strong conclusion is warranted by the above argument. In the next section, I hope to make clear in what sense I think moral cognition *not* to be generative.[12]

Issues with the Argument for Moral Grammar

With an eye on amending the above argument, let us begin by considering some relevant disanalogies between language and moral cognition. First, although grammaticality judgments are in one sense analogous to permissibility judgments, the parallel is misleading.[13] To begin with, in language, the fact that a sentence is perceived as ungrammatical is an epiphenomenon that is due to whether the arrangement of functional elements in the sentence happens to conform to the principles of syntactic combination. In other words, there is no analysis of grammaticality *per se*, only an analysis of a sentence that is either successful or it isn't. In the latter case, that is, if the analysis fails because the constituents are not appropriately arranged relative to the rules of syntax, then the sentence is perceived as ungrammatical.[14] In other words, as far as we know there is no dedicated mechanism that takes as input a sequence of words (phonemes, sounds, syntactic categories), and outputs a binary valued variable representing the sentence in terms of its (un)grammaticality. In contrast, in the moral case, presumably, the whole point of the analysis is to appraise acts in terms of their deontic status, so a judgment of impermissibility is not due to a "failure" of the analysis, quite the contrary.[15]

The second crucial difference is between the respective sources of productivity. In the case of language, the sources of productivity are the principles along with the representations they are applied to, which formally guarantee that the sentences of a language can be generated.[16] In other words, once the principles are properly formulated, productivity follows. *Prima facie* at least, the same is *not* true in the case of moral cognition: it appears that an exhaustive description of moral principles and the relevant representations will

[11] *Cf*. also Hauser: "To attain its limitless range of expressive power, the principles of our moral faculty must take a finite set of elements and recombine them into new, meaningful expressions or principles" (2006, *p*. 47).

[12] To be fair to Mikhail, elsewhere, he acknowledges that the analogy may not be perfect: "language is an infinite combinatorial system in a way that morality (as distinct from the cognitive systems by which actions are mentally represented) may not be" (Mikhail, 2017, *p*. 241).If this is the case, the two arguments cannot be completely isomorphic.

[13] *Cf*. also Dwyer (2008) who expresses some doubt concerning the equivalence between the respective roles of these judgments in linguistic vs. moral theory.

[14] Technically, "unacceptable", but in this case, the unacceptability is due to ungrammaticalness (see Chomsky, 1965, *p.* 10ff).

[15] Mikhail is of course aware that the analogy between grammaticality and permissibility is not perfect, but he only makes the less theoretically relevant point that "recognizing whether a given structure is grammatical is *less central* to everyday behavior than the ability to recognize whether particular conduct is morally permissible" (2017, *p*. 240; emphasis added).

[16] Although as Chomsky often points out, how they are actually generated is something of a mystery to do with the interface between language and thought.

not guarantee productivity, only that, in principle at least, moral cognition can "harness" the productivity of the representations providing input to the moral faculty. In other words, we don't generate representations of situations or actions based on moral principles. Of course we do generate moral *evaluations*, but these, in contrast to the representations of actions and events that are the subject of evaluation, are neither potentially infinite nor in fact do they appear to be open ended.[17]

A friendly amendment

Mikhail might try to defend his version of the Argument for Moral Grammar as follows. It seems reasonable to accept that human *thought* is capable of producing a potentially infinite number of representations of actions (and presumably other things too, like states of affair, events, and so on). An informal way of illustrating this productivity is by considering works of fiction: although there is some repetition of themes and motifs in the history of literature, we never seem to run out of things to write, read, and thus more generally to think about, notably so in the domain of human action. Just as a whole book can be written in a single sentence (a recent example is Mike McCormack's *Solar Bones*), so can a whole book be written about a single action (Odysseus's journey home comes to mind).

A slightly more formal illustration of this point is provided by Jackendoff's analysis of action (see Jackendoff, 2007, especially Chapter 4). To cut a rather long story (very) short, action representations can be best understood as having complex hierarchical structure involving recursive embedding at different levels of the "action tree". For example, preparing coffee may involve taking the coffee out of the fridge (that is where my flatmate keeps it). This latter representation (which is itself an action!) is *part* of the former action, embedded in a complex structure consisting of a Preparation (preparing the ground for the main action), a Head (the part that constitutes the goal of the action), and sometimes an optional Coda (restoring the *status quo*). On Jackendoff's account, these are the proprietary representations of the action representation system (analogous to NPs and CPs in the case of syntax; see above). Irrespective of whether one wants to buy into the details of Jackendoff's account, the point of recursive embedding stands: I can think of the action of taking the coffee out of the fridge as a complete action and an end in itself, for example because I don't think it makes sense to keep it there. I can also entertain the thought of the action of preparing coffee for some visitors, in

[17] Trivially, one could say that "Jim's killing the Indian is impermissible" is a novel representation compared to "Jim's killing the Indian", by reason of which moral cognition may be said to be a source of productivity. But intuitively, this is not the interesting kind of productivity observed in the case of language and thought. Still, how novel situations and actions can be evaluated effortlessly and automatically is a genuine problem for a theory of moral cognition (as pointed out in the text).

which case the representation of the whole action as considered thus far (making coffee) would be embedded in a larger structure of the same type, for example, welcoming the visitors. And so on, in both directions.

To return to the original point, in spite of this productivity (and despite the potential novelty of action representations), at no point do we stop being able to evaluate actions in moral-deontic terms. In other words, moral cognition deals with the productivity of the "grammar of action", even though as argued above it doesn't account for it.[18] Nevertheless, moral judgment clearly needs some resources to tackle the productivity of the representational system the products of which its operations take as input. Just as Hume suggested, principles of *some* description appear to be the best candidates for this purpose—as long as there exists a representational format (analogous to NPs and CPs) that abstracts away from the vagaries of "fully explicit" action representations—otherwise positing principles would only delay rather than solve the productivity problem; much like positing syntactic rules defined over particular words would in the case of language. These representations do need to be "generated" by what Mikhail refers to as *conversion rules*, that is, rules that dictate how sensory and perceptual representations are turned into a more abstract representational format (involving representations of causes and intentions) that principles evaluate actions in terms of. But it would be odd to suppose that the product of conversion rules is potentially infinite: the very point of conversion rules is to reduce variability after all, not to create it (nor as a matter of fact are conversion rules likely to be specific to FM, on any plausible individuation of FM).

As usual, an analogy with language helps here: we parse sound waves into a finite set of phonemes. The conversion rules that achieve this feat reduce the variability of and eliminate noise from the input by means of ignoring some features of it and magnifying others (this process is called "categorical perception"). The details won't have to detain us here, the point is merely that the output is a finite set of representations (the set of phonemes in the given I-language), and there is no need for *combinatorial* rules at this stage. (To be sure, this analogy is also imperfect, since the variability of the sound signal is of a different kind from that in the input to moral evaluation. Also, some aspects of phonology—referred to as phonotactics—do contain combinatorial rules defined over phonemes, but, again, there is no analogy for *that* in moral cognition as far as I can tell.) Also relevant is the fact that for language processing to take place, the input activating the early

[18] Note that syntactic productivity doesn't account for the productivity of *thought*, either. Nevertheless, the former clearly has its own source of productivity (e.g. rewrite rules), which is perhaps part of what makes it capable of dealing with the productivity of thought. But note that the two kinds of productivity are not entirely overlapping: many well-formed sentences can be generated which nevertheless have no obvious semantic interpretation ("colourless green ideas sleep furiously" is one such popular example). And vice versa, I assume there are many thoughts that are impossible to express in language. In any case, the important point here is merely that unlike I-language, it is far from obvious that is I-morality has its own source of productivity.

auditory system has to be transformed into this format. Similarly, if I represent Frank hitting Roy *merely* as a physical event, I won't be evaluating it morally.

Thus, the difference between principles of I-*language* (particularly, principles of syntax) and principles of I-*morality* is that, *prima facie* at least, only the former are (or need be) principles of *combination*. I shall have more to say regarding the question of just what types of principles a theory of moral cognition should be expected to posit in Section 5.

3.3. Competence vs. performance

3.3.1. Language

As discussed above, Chomsky noted that there are rules of grammar that entail that the set of possible grammatical sentences is infinite. Yet, obviously, no human can ever produce more than a finite set of utterances, and there are very many (in fact an infinite number of) *grammatical* sentences that speakers of a given language would find very hard to understand[19] or would never understand[20] in spite of all the considerations in favour of their "knowing"[21] the principles from which those sentences can be generated (as in *fn*. 19). However, it is clear that some of these limitations derive not from our knowledge of language (i.e. our I-language) *per se*, but from other sources.

In part to capture such constraints imposed by language-external considerations, in the 1960s Chomsky introduced the distinction between *competence*, that is, those aspects of the mind responsible for the generative principles, or the "knowledge of language" (which later came to be referred as *I-language*) and *performance*, that is, the overall behaviour that results from other extraneous factors partially constraining the expression of that knowledge, including limitations of memory, shifts of attention and interest, as well as aspects of language processing (Chomsky, 1964, 1965; Miller & Chomsky, 1963).[22] This distinction was especially relevant in the historical context of the birth of generative grammar, because the then prominent

[19] Classical examples are constructions involving repeated centre embedding, such as in the sentence: "the person who the girl who my colleagues mentioned fell in love with walked past the cafe". In Chomsky's terminology, such sentences are "unacceptable" but clearly grammatical (in the sense of being entailed by the rules of I-language). A serious difficulty with these constructions is that while one has to process them, one has to keep multiple subjects in mind as well as connect them subsequently with the correct predicate (Miller & Chomsky, 1963; for another theory, see Jackendoff, 2002, *p*. 32).

[20] Infinite sentences would be prime examples of this latter category.

[21] Jackendoff refers to this kind of "knowing" as *f-knowledge* (*f* for "functional"). The point is that there is no *propositional* knowledge implied by "knowledge" in Chomsky's sense (see also Chomsky, 2018).

[22] Note that although the concepts of I-language and competence are closely related (see the text), the same is not ture in the case of the concepts of E-language and performance (for instance, only the former is supposed to be mind independent).

language theorists, namely behaviourists and structuralists, failed to draw such a distinction, which resulted in their inability (often explicitly endorsed) to distinguish between aspects of the data requiring explanation in terms of linguistic theory proper from aspects of the data for which theoretical explanation would more appropriately be expected from other domains of inquiry (a "theory of performance"). In other words, for behaviourist linguistic theory, there was no principled difference between data and explanandum.[23]

One intuitive way to illustrate the ineluctability of the distinction is by considering the case of a person with acquired deafness and muteness. Although in connection with such a person, there is no overt language use to speak of, neither on the perceptive nor on the productive side, one would naturally assume that the cognitive structures used to subserve the person's capacity to speak and understand language would not cease to exist subsequent to the onset of his or her disabilities. Intuitively, one would assume that the relevant cognitive structures still in place are more central to explaining the person's language capacity even prior to the onset of the disabilities than whatever caused them not to be able to exercise their language skill. The former are aspects of language competence, the latter are aspects of language performance. The empirical details of how to outline just what needs to be included in competence are murky and controversial, but that a line ought to be drawn somewhere is not so contentious (indeed, in the case of language, a great deal of the disagreement over competence vs. performance is *where* to draw the line, not *whether* it should be drawn; see e.g. Culicover, 2013).

3.3.2. Moral cognition

With respect to the study of moral cognition, the distinction between competence and performance can be understood in two ways. The first is to see it as the desired shift away from paradigms openly or covertly influenced by the behaviourist paradigm whereby, as in the linguistic case, performance on a task is mistaken for the explanandum of the theory of moral cognition (Mikhail, 2011). This interpretation is in line with the spirit of this section. The other way of seeing the relevance of the distinction is from the perspective of distinguishing between the contribution of moral competence from that of other mental faculties and processes that are often involved in moral reasoning or tasks designed to assess moral cognition. Both of these views are legitimate and they are in no way mutually exclusive.

Of course, as in the case of language, identifying and isolating the functions performed by (dedicated faculties of) the mind is by no means a straightforward procedure. To take an example from another cognitive

[23] To be sure, data have to be explained by any theory worth its salt, irrespective of the particular empirical enterprise, but it is rarely if ever the case that all aspects of the data are—or should be—treated with equal attention. In particular, it is a virtue to be able to separate noise from relevant aspects of data.

domain, carrying out even the simplest of behavioural tasks, such as multiplying a pair of two-digit numbers, involves the operation of both "horizontal" or relatively domain general faculties (such as memory) and "vertical" or relatively domain specific faculties (such as reading). Implicated in the transition from input to output are myriads of complex processing steps from the transduction of light into electrical signals through the transformation of the ensuing visual information into abstract numerical representations over which the mathematical operations are performed, all the way to the motor response (jotting down the result or saying it out loud). Intuitively, some of these will be crucial from the perspective of the study of mathematical cognition, but others will be only partially of interest or entirely irrelevant. In this example, the former category constitutes the competence for multiplication, the latter are aspects of the performance of the task that partially draw on that knowledge—as well as many other perceptual and cognitive abilities and processes besides.

Similarly, it is not controversial that certain elements of the complete set of processing steps of arriving at a moral judgment will be of limited interest from the point of view of the study of moral cognition. In an experimental situation, for example, the subject is typically exposed to a stimulus, usually in the form of a description of a story he or she is expected to read. This involves the visual system and the system dedicated to reading, *inter alia*. Neither of these systems seems to be essential for moral cognition: for example, blind and illiterate people can think in moral terms. (Imagine assessing the moral capacities of an illiterate individual by means of a test involving reading a story, and, based on their failure to provide the appropriate answers concluding that they are deficient in terms of their moral faculty.) Some aspects of judgment clearly belong to performance, but the inverse question, again, is a difficult ultimately empirical one. Even if we narrow down moral cognition to the ability to assign a moral-deontic value to an action, it is far from straightforward what cognitive structures and operations exactly moral competence should be understood to encompass. To take but one example, it is famously unclear whether emotions play an essential role in the generation of moral judgment (*cf.* Huebner et al., 2009). One way to phrase this question is to ask whether emotions are part of moral competence or performance.

A complicating factor is that some cognitive components may be necessary but not unique to the normal exercise of the cognitive capacity being investigated—as indeed in the case of emotion. With regard to language, there is an ongoing debate over what cognitive mechanisms contributing to language perception and production are *specific* to language competence—referred to as FLN, that is, the *Narrow Language Faculty*—and which ones are shared across other domains—referred to as FLB, that is, *Broad Language Faculty* (Hauser, Chomsky, & Fitch, 2002; Pinker & Jackendoff, 2005, Jackendoff & Pinker, 2005; Fitch,

Hauser, & Chomsky, 2005).[24] The same question can be raised with respect to morality. For example, it forms the basis of practically universal agreement that aspects of social cognition contribute to moral cognition. Thus, while the representation of intentions is often claimed to be central from the point of view of moral cognition (e.g. Wellman & Miller, 2008, but see Barrett et al. 2016, for example), we clearly think about intentions in contexts not involving moral considerations. A persistent quest in moral psychology has been to identify the "FMN" (Narrow Moral Faculty), although the debate is rarely phrased in such terms. I shall return to this issue further below.

In the moral domain, at the very least, the distinction is useful as a heuristic device that draws attention to the fact that moral judgment is likely the result of an interaction between numerous cognitive processes and systems, only some of which constitute the appropriate object of inquiry as far as the study of moral cognition is concerned. I believe this point is generally often overlooked in the literature, which is connected to the phenomenon that psychologists often look for interesting effects rather than investigating and developing an understanding of mental competences (see also Cummins, 2000). Nevertheless, the distinction between competence and performance (as well as the related—though conceptually distinct—separation of I-morality and E-morality) has some serious repercussions with respect to what we take morality to consists in. In particular, as discussed in the next chapter, it makes no sense to draw either of the relevant distinctions if moral cognition is *not* a natural kind. Irrespective of whether that assumption is correct, however, the distinction will be useful in making sense of proposed explanations of certain patterns of moral-deontic reasoning (in Chapter 4).

3.4. Production and perception

An aspect that is part of performance in the Chomskyan tradition is language processing.[25] Consider a sentence, such as "octopi are erudite". There are two different ways of arriving at the linguistic (phonological, syntactic, etc.) structures that enable us to understand this sentence: having the thought first and "translating" it into language (this was the way I came up with it) or being exposed to it in a written or auditory format (this is how you did it). Although these routes involve rather different cognitive processes, the linguistic structures we arrive at eventually are supposed to be identical (which is part of what makes communication possible). The fact that both (constructive and perceptual) processes need access to the relevant

[24] A relevant additional complication is the question of the extent to which aspects of FL are shared across species (e.g. vocal imitation). I'll ignore that topic here (but see e.g. de Waal, 1996).

[25] "When we say that a sentence has a certain derivation with respect to a particular generative grammar, we say nothing about how the speaker or hearer might proceed, in some practical or efficient way, to construct such a derivation. These questions belong to the theory of language use—the theory of performance" (Chomsky, 1965, *p.* 9).

representations (albeit from different directions) motivates the relegation of processing to the theory of performance. Consequently, a theory of performance on this view can be constructed insofar as we have a good theory of competence (otherwise, it is unclear what the representations are the construction of which processing theories are expected to explain). However, the relevant point I wish to emphasise here is that perception and production bifurcate: despite the shared representations they make use of, they require different kinds of explanations, as they involve different kinds of processes.

The distinction between perception and production has a direct analogy in the case of moral cognition, and it similarly bisects the inquiry into two related but importantly different explanatory endeavours in a way that has often been overlooked even in modern moral psychology and also in evolutionary approaches to moral cognition, which often equate a tendency towards cooperation or altruism with morality (but see e.g. Joyce, 2006). The main problem of perception in the case of moral cognition is the problem of how we evaluate other agents' actions in moral terms. This question has its own proprietary explanandum (such as the appraisal problem—see the next chapter), which is characteristically different from production problems, a typical example of which is the task of integrating the outputs of the moral faculty into an overall decision making process.[26] Thus, there are two points suggested by LA with respect to the relationship between production and perception. First, both processes will access at least some of the same representational repository. Second, production and perception, although overlapping problems, will require importantly different explanations.[27]

To connect two significant points, ideally, principles might be some of those cognitive structures characterising competence (or I-morality) that are accessed by both productive and perceptual processes. Of course there is no guarantee that on the discovery of a principle which correctly generates moral judgments of a particular type, those judgments are best explained by reference to competence or performance factors (see e.g. Nichols, 2005—more on this in the following chapter). Nevertheless, the ideal of searching for principles characterising moral competence (i.e. competence for moral judgment) is an ideal that is worth

[26] Jesse Prinz, a passionate critic of the Linguistic Analogy, acknowledges the utility of this distinction: "Laudable behavior can exist without the capacity to praise it as such. One of the exciting features of [LA] is that [it directly investigates] moral judgments, rather than morally praiseworthy behavior" (2008, *p*. 165). Ironically, the clear division between the problems of production and perception is one of the aspects of LA that would not be suggested by a "vision analogy", which is recommended by Prinz as an equally valid analogy for the study of morality (see the next chapter), although a "motor analogy" (also mentioned by Prinz) would perhaps fare better in this respect (though not so in many others).

[27] This is commensurable with some evolutionary arguments concerning the primacy of perception in this domain (*cf*. e.g. DeScioli & Kurzban, 2013). Thus, this might be one example of the many potential disanalogies between language and morality (which is noticed only if the analogy is acknowledged first).

aspiring to, I believe. This brings us to the question of just *what* a principle in the context of the current framework. This question is answered in Section 5 below, after considering the question of what shape explanation is to take in the theory of moral cognition if we take LA seriously.

4. The nature of explanation

A theory of moral cognition needs to do more than analyse the nature of the phenomenon it attempts to explain (namely, moral thinking): its success will be predicated on those of specific explanatory proposals for particular aspects of moral thinking. But how are such theories to be formulated? So far, I have been concerned only with outlining the substantive assumptions of the Linguistic Analogy in terms of the nature of the explanandum: the phenomenon we are trying to understand and explain. In this section, I go on to unpack some of the implications of this model with respect to the shape of the broader explanatory project.

4.1. Computationalism and the search for moral principles

A substantive theory of what thinking—and *ipso facto* moral thinking—is (i.e. mentalism *qua* computationalism and representationalism) rather directly sets the agenda for the explanans: the terms in which the phenomenon is to be explained. To reiterate, the (rather) general consensus in cognitive science is that thinking is a form of symbol manipulation. Once this picture is taken seriously, it follows that explanation of cognitive processes such as thinking—and *ipso facto* moral thinking—will have to be computational. Fortunately for us, there exist detailed accounts dedicated to fleshing out what form computational explanation is to take, in particular with respect to a theory of mental structure.

Although we could illustrate the same points by reference to explanation in linguistics, for historical reasons, I will use David Marr's ground-breaking analysis of explanation in the cognitive science of vision as outlined in his book, *Vision*. This analysis applies equally well in other areas of cognitive science, such as linguistics or problem solving. Marr conceived of explanation in cognitive science as involving three fundamental levels: the functional level, the computational-algorithmic level, and the level of implementation (Marr, 1982).[28] At the functional level, the main task is to characterise the function carried out by a mental mechanism in terms of an input-output relation. That is, the focus is on *what* it is that the mechanism does.

[28] I slightly depart from Marr's terminology, who describes the first level as the level of computational theory and the second level as the algorithmic level. My only reason for doing so is that, for better or worse, I find referring to the first level as "computational" slightly misleading. Marr's use of the term is somewhat unusual, referring to computational tasks rather than computational processes in the sense of symbol manipulation (see Miłkowski, 2013, *p.* 114; also see Sterelny, 1990, who refers to the first level as "ecological").

At the computational-algorithmic level, the analysis involves the particular representations involved as well as the processing steps that are used to carry out the transformation of the input into the output. That is, the focus is on *how* the mechanism does what it does. Finally, at the level of implementation, the aim is to explain how the mechanism as well as the representations involved are physically realised or implemented in the physical structure that carries out the computation (the brain in our case).

As an analogy, consider how we might go about investigating a mathematical function of a calculator in a case in which our initial epistemic access is limited to the input and output of the function. First, we try to find out what the function does [try $f(2) = 4$; $f(3) = 9$; etc. $\rightarrow f(a) = a2$] and what its domain and range are [try $f(-2.5) = 6.25 \rightarrow a$; $f(a) \in \mathbb{Q}$]. Then, the next step is to decide what algorithm realises the generation of the output [$f(a) = a \times a$; or $f(a) = (a-1) \times (a+1) + 1$, which are extensionally equivalent]. Finally, the last step towards a complete understanding of the function is the investigation of how the algorithm is physically realised (e.g. how a calculator's transistor controls the movement of electrons, and so on)

This is analogous to the investigation of the human mind in general or aspects (i.e. mechanisms) of the human mind. On the most general interpretation of this framework, this work may proceed in the following way: we provide the human a proximal stimulus in the lab, and record what it does with it. But instead of stopping at describing the input and the (behavioural) output, we speculate what processing steps are involved, and what operational principles these are constrained by. Finally, once we are fairly certain that our analysis is correct, we have a look at how the brain carries out the identified processing steps. Of course, this picture is far too oversimplified, because it treats the human as the function, yet we know that any carrying out a behavioural task will involve multiple processing steps. Nevertheless, in an ideal state of cognitive science, each of these steps would be described in functional and computational-algorithmic terms based on data gained from sophisticated experimental procedures.[29]

[29] An illustrative example is the theory of early (low- and intermediate level) visual processing, which has achieved a fairly sophisticated understanding in terms of all of Marr's levels (see e.g. Kandel et al. 2012, *pp.* 577-620). Many argue of course that this type of explanation gets more and more unachievable as we move from "peripheral" perceptual processes to more "central" ones. The paradigm case is Fodor (1983; 2000). But many disagree (see e.g. Barrett, 2015; Carruthers, 2006; see also Chomsky, 2018). This is one of those (very many) fascinating topics that will not be settled in this thesis. Suffice it to say that Fodor's worries about central cognition did not stop researchers from making remarkable progress in these domains. Furthermore, regarding our topic here, it is unclear as to what extent moral cognition is properly regarded as central, however one might proceed in terms of establishing the boundary.

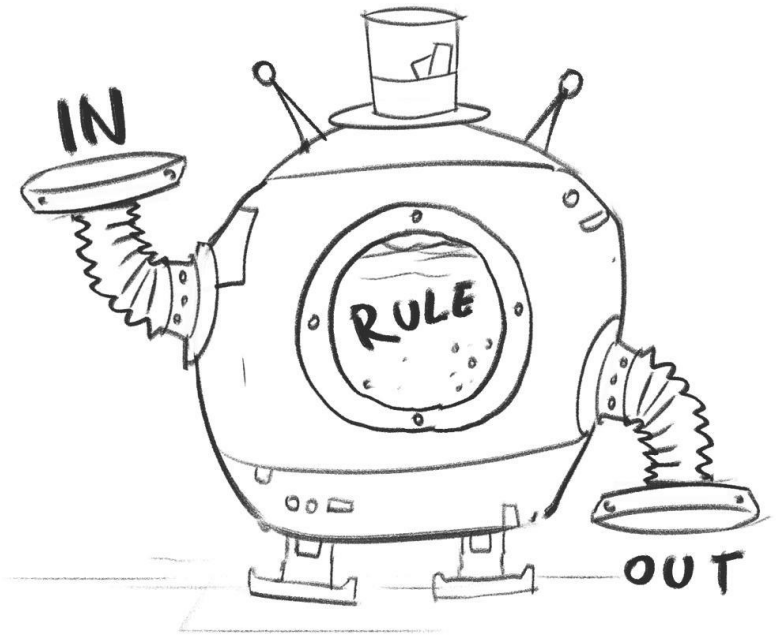

Figure 1.1: The function machine

In the broadest sense, the study of moral cognition is the study of the functions that are implemented by the mind/brain in the course of moral thinking and reasoning, involving the characterisation of input-output relations as well as an algorithmic description of what computations those relations are performed by. As mentioned earlier, and as suggested by even the modified, more modest version of the Argument for Moral Grammar endorsed above, the preliminary search for representations and functions is based on the assumption that moral judgment is driven by rules and principles, accounting for the productivity and unboundedness it is designed to harness. The hope is that by identifying candidate principles that have a running hope of achieving descriptive adequacy, and by subjecting them to systematic empirical and theoretical inquiry, eventually we can explain those principles in rather more explicit computational terms.

5. What is a principle?

Thus far, I have not said much about what form principles are expected to take, nor what principles *are* to begin with, beyond pointing out that they are part of the solution to the productivity problem and that, contrary to what AMG is generally taken to show, they need not be combinatorial. Although in general I don't think it is a very good idea to start explicating a technical term by reference to a dictionary definition, I cannot resist that temptation here, because the following dictionary entry happens to prove rather

instructive.[30] According to the Oxford Dictionary a principle is "a natural law forming the basis for the construction or working of a machine" (sense 2.1). I find this definition a reasonable start as it draws attention to the two central ways in which a principle can feature in the description and explanation of a mechanism or a system, namely, as guiding its construction or operation (see below). But of course for our purposes it is far from being sufficiently nuanced.

Chomsky characteristically uses the term with a "systematic ambiguity", one not unlike that involved in his use of the terms *grammar* or *theory of language* (1965, *p*. 25).[31] "Grammar" may either refer to an individual's *mental* or *internal* grammar (i.e. the "knowledge" characterising linguistic competence—see above), or to the linguist's model of it (a generative grammar capable of generating all and only the grammatical sentences of a given language). Analogously, "theory of language" may refer either "to the child's innate predisposition to learn a language of a certain type and to the linguist's account of this" (ibid.). Similarly, *principle* may denote either something that is built in the cognitive machinery (an internally represented "principle"), or it may refer to something that is part of the cognitive scientist's model of it; a "model principle", something that helps establish the appropriate functional (input-output) relations whether or not it describes an algorithm that corresponds to cognitive structures and processes operative in the modelled speaker's mind. Those concerns correspond to Marr's first and second levels, respectively (see Section 4). In a successful case, that is, when an aspect of (mental) grammar is appropriately characterised or explained in the form of an explanation pitched at the computational-algorithmic level, the two kinds of principles will be strictly isomorphic. At the outset of the inquiry, though, there is no assumption to the effect that the linguist's principles are explicitly (or otherwise) represented in individuals' minds. That is, as per the normal course of an explanatory endeavour, the explanation begins at the functional level (see more on this topic in the next chapter).

A second distinction is related to the equivocation in the above quoted definition, namely, whether the principle concerns the "construction" or "working" of the machine—or the mind in our case. As noted in Section 2 above, a central problem in language (as well as in moral psychology) is to explain how we get to the mature state of linguistic (or moral) competence. To reiterate, Chomsky's answer to this question involves

[30] Still, it should be born in mind that this section is most definitely not an exercise in conceptual or linguistic analysis. The present aim is to develop an understanding of the ways in which it makes sense to talk about principles in the context of the study of moral cognition as suggested by LA (by contrast, in the context of law, other constructs may be more appropriate, see e.g. Robinson 2016, *pp.* 38-42). The intention is that by drawing these distinctions, it will be easier to make sense of some claims we will encounter in later chapters.

[31] In fact, it was due to the persistent misunderstandings and misinterpretations on account of this ambiguity that Chomsky eventually introduced the term *I-language* (in Chomsky, 1986).

positing an innate system of the mind/brain (the initial state of the language faculty or FL) that, in its interaction with the data the child is exposed to in the course of his/her ontogenetic development (or the "primary linguistic data" or PLD), is responsible for the acquisition of the I-language the child ends up with (the adult state of FL). Thus, FL (in its initial state) is what the mind contributes specifically to the task of language acquisition.[32] The task of the linguist is to provide an abstract model of the initial state of FL (hence, there is an ambiguity with respect to FL that is akin to that with respect to either a grammar or a principle—see also Collins, 2004). This model involves principles that capture or guide the diachronic development of FL (for example by specifying what data it requires to arrive at the construction of particular mental grammars). Analogously, proponents of LA have hypothesised that the initial state of FM is non-zero, that is, the mind is innately prepared to develop what eventually ends up being the adult state of FM: I-morality.

An instructive classification scheme is advanced by Dwyer, who distinguishes between *norms*, *rules*, and *principles* in a way that is not customarily done in the moral psychology literature (Dwyer, 2008, *pp*. 411-414). According to Dwyer's proposal, norms should be regarded as emergent, group-level phenomena; those standards of conduct that are publicly accessible and (to varying degrees) collectively enforced by means of shared public attitudes towards violators and in some cases even legal sanctions. In contrast, rules are understood as individual level phenomena: they are represented in individuals' minds and thus guide their behaviour (production) and/or the evaluation of other agents' conduct (perception).[33] Consequently, although norms and rules may share content, it is technically possible for them to be doubly dissociated: a member of a moral community may privately (explicitly or otherwise) disagree with a prevailing norm while "having" idiosyncratic rules that are not widely shared. It may be apparent that the distinction between systems of norms and systems of rules is very similar to the distinction between E-languages and idiolects, that is, I-languages. On this conception, therefore, the study of E-moralities has norms as its proper subject, while the study of I-morality consists to a large extent of an inquiry into the nature and content of rules. That leaves us with principles, which, for Dwyer, characterise the moral faculty. More specifically, principles characterise the initial state of FM such that they explain the kinds of (I-)moralities we end up with—thus, they define the

[32] As mentioned in the previous section (Section 4.2.2 to be exact), there emerged a subsequent distinction between a FLN and FLW in the literature. To reiterate, the first is the language faculty in the narrow sense, i.e. system(s) that are entirely domain specific in the sense of being dedicated to language acquisition or knowledge and nothing else. FLW is not dedicated to language *per se* but it comprises the systems that contribute crucially important functions to language, whether in terms of acquisition or knowledge. On some extreme empiricist models (e.g. Christiansen and Chater, 2008), there is no need for positing FLN (or UG), because what the mind contributes to language is never exclusive to language itself, namely, general purpose or domain general learning principles and processing capacities. There is a hotly debated analogous question with respect to morality (see the next chapter).

[33] Thus, somewhat confusingly, *rule* more or less corresponds with what Sripada and Stich (2006) refer to as *norms* (although at times they are somewhat ambiguous in the use of the term).

space of possible I-moralities (*cf*. Moro, 2008 in the domain of language). That is, they guide and constrain the ontogenetic development of moral competence.

Dwyer rightly criticises moral anti-nativists (i.e. those that doubt whether there is an FM in the narrow sense) such as Prinz (see e.g. Prinz 2008, 2009) for not adequately identifying the explanatory level at which they believe innateness claims ought to be made or contested. As Dwyer points out, it makes little sense either to entertain or reject the possibility that E- and/or I-moralities, and, *a fortiori*, norms and rules might be innate. The only legitimate target of nativist claims and their rebuttals is if they are addressed at the faculty level (both in the case of language and morality), that is, if they concern *principles* in the way in which Dwyer defines them.[34]

Given how little we know concerning the constraints that regulate the development and acquisition of I-moralities, it is not surprising that Dwyer is uncertain as to how much at present there is to be said in connection with them. It is worth quoting Dwyer on this at some length:

> "To be frank, the form and content of the principles that I claim characterize the moral faculty remain a mystery. But what this approach predicts is that the articulation of such principles is unlikely to involve the use of terms with which moral philosophers are currently familiar. [...] As to the form of principles, I think we need to think more creatively about the nature of constraints in general. Not all constraints take the form of imperatives like "No flip-flops allowed in the front bar!" In cognitive science, we can think of constraints as ways of blocking a cognitive movement; a sort of "you-can't-get-there-from-here" admonition. Consider the moral judgment that it is good to torture small babies for fun. That has the feel of something no "normal" moral creature could generate. At a really fundamental level, then, the idea is that the principles of the moral faculty are what explain why such judgments cannot be generated" (Dwyer, 2008, *p*. 414).[35]

Although, as I have suggested, Dwyer's tripartite distinction is useful (for example on account of rendering the moral nativism versus anti-nativism debate more tractable), it misses an important point that reaches beyond the mere exegetical fact that other proponents of LA, such as Rawls and Mikhail (and indeed Chomsky himself), customarily make use the term *principle* in a rather different fashion, namely as being

[34] Claiming that a rule (such as the "I-prohibition" against killing) is innate might be seen as roughly equivalent to claiming that an English speaker's past tense formation rule is innate (and *mutatis mutandis* for norms). (Although notice that even such rules could be properly described as innate for instance under the "innateness as canalisation" view; e.g. Ariew, 1999).

[35] Or, to put it in terms of the current discussion, why I-moralities consistently lack the rule "torturing babies for fun is morally good".

more or less synonymous with Dwyer's *rule*, that is, a denizen of our mental lives that serves the purpose of guiding or generating moral judgment—or as formal tool of a model of moral judgment.[36] Rather, the problem is that exclusive emphasis on the explication of rules and principles (in Dwyer's sense) as the primary aim of moral psychology overlooks the fact that an exhaustive characterisation of rules (again, in Dwyer's sense, that is, *qua* mentally represented "injunctions") is unlikely to be sufficient to solve the problem of the productivity of moral cognition (see Section 3.2), for example since reasoning about and passing judgment on novel morally salient actions does not reduce to categorising them as instances of morally significant act-types, such as murder or theft.[37] For instance, principles such as the Action Principle (according to which harm caused by an action is morally worse than that caused by an omission) or the Principle of Double Effect or PDE (which determines overall permissibility given a *prima facie* wrong and a morally desirable outcome that is achieved through it) are not specified over particular action types at all. Inputs to these principles apparently involve deontic statuses of acts (*prohibited* or *impermissible*), and they arbitrate over the permissibility of such acts given a set of circumstances taken into account by the principle in one way or another.[38]

Thus, with an eye on a more nuanced (though not necessary exhaustive) topography of the possibility space, I distinguish between three broad types of principles. First, as alluded to before (and as will be discussed more extensively later in the thesis), there are constraints on what types of representations are potentially subject to moral-deontic evaluation. For example, my suspicion is that representations of events not involving agents—though their consequences may be "morally" undesirable (think of wildfires or earthquakes)—fall outside the moral domain (as understood in this thesis). Principles regulating the mental properties of the input to moral-deontic evaluation will be referred to as *constraint principles*. Constraint principles determine the kinds of things are subject to moral evaluation. Second, there are regularities concerning *what* deontic values particular acts are understood as having. These are supposed to be captured by *faculty* principles. Faculty principles—such as the prohibition of intentional homicide—are perhaps the closest to rules in Dwyer's sense. Finally, I shall refer to principles determining overall permissibility given situational constraints (such as the Action Principle or PDE—see the above paragraph) as *conflict principles*, because they

[36] Take Rawls, for example: "what is required is a formulation of a set of principles which, when conjoined to our beliefs and knowledge of the circumstances, would lead us to make these judgments with their supporting reasons were we to apply these principles conscientiously and intelligently" (1971, *p.* 47).

[37] Although it is also possible that rules go well beyond that simplistic characterisation. The representational format of norms/rules is far from being a settled issue in cognitive science (*cf.* e.g. Sripada & Stich, 2007).

[38] Mikhail (2002, 2011); Donagan (1977); Rawls (1971); see further below.

resolve a conflict between two (or more) apparently incongruous deontic rationales, that is, *prima facie* obligations or prohibitions.[39]

I should note in passing that there are other related but distinct classifications in the legal or moral philosophy literature regarding norms (i.e. Dwyer's rules and our faculty principles), but I find them in this context not especially helpful. For example, in their attempt to classify the "Ought Implies Can" principle (see chapters 3 and 4), Fox and Feis (2018) consider distinctions such as that between *primary* and *secondary* norms. The first difficulty is that different authors use these terms in somewhat diverging ways. For instance, in Kelsen's terminology, primary norms are descriptive in the sense that they specify the sanctions warranted by different types of wrongdoing, such as 'the hand of a thief is cut off'. In contrast, secondary norms specify what one should do in order to avoid the relevant sanctions, such as 'thou shalt not steal' (Kelsen, 1967; Navarro, 2013). Hart, on the other hand, regards norms of the latter type as primary norms; for him, secondary norms are those that bestow the power of the authorship of primary norms on a lawgiver. (Fox and Feis also discuss related concepts, such as those of "iterated norm" and "higher order norm"—see von Wright 1983). More importantly, although these distinctions are instructive, and they clearly bear some relevance for our purposes here, they are not particularly well suited for the study of moral cognition. For this reason, I shall rely on my own terminology, because it is designed to deal with moral judgment in the context of cognitive architecture and as such it is more helpful from the perspective of the issues discussed in this thesis.

Notice that in distinguishing between three different types of principles (constraint, faculty and conflict), we did not leave room for principles in Dwyer's proprietary sense, that is, *qua* principles characterising (the initial state of) FM. To reiterate, these principles (the form and content of which remain subject to speculation given the current state of the inquiry—see the quote from Dwyer above) are supposed to regulate the acquisition of the moral sense; or as a *bona fide* Chomskyan would put it, the growth of the moral faculty. To capture this sense of the term, I distinguish between *acquisition principles*, that is, principles in Dwyer's sense, and *processing principles*, that is, principles in more or less the senses we have been considering in the previous paragraph (and also in Rawls's and Mikhail's sense, for instance).[40] Thus, this latter category includes constraint, output and conflict principles. A standard assumption would be that processing principles develop as a function of acquisition principles plus the external input that shapes the development of FM (in which sense the two types of principles are systematically related). As we will see further below (Section 6),

[39] A related term is *ordering principle* (Mikhail, 2002, *p.* 25; Donagan, 1977, *pp.* 157-164).

[40] By referring to these principles as *processing* principles, I am deliberately ignoring an intricate issue that we touched upon above (Section 3.4), namely, that theories of processing may be understood as theories of performance rather than competence, as it involves the *use* of information stored or supplied by FL/FM. I merely raise this issue to put it to one side: including this distinction here would further complicate an already rather complex classification scheme.

these two types of principles, acquisition principles and processing principles, form parts of different, albeit related, explanatory endeavours, namely, those dedicated to solving the problems of explanatory and descriptive adequacy, that is, the problem of explaining synchronic and diachronic aspects of moral cognition, respectively. The distinction will also be important from the point of view of chapters 3 and 4.

Despite Dwyer's pessimism regarding our current understanding of the nature of acquisition principles, we might speculate a bit further. For example, there is what we might refer to as *principles of derivation* that purport to drive ontogenetic development, and perhaps even beyond. A principle of this kind, although never explicitly articulated as such, may be that which is held to distinguish the moral from the non-moral domain according to some theories. On a typical view of this kind, such as social domain theory, for example (e.g. Turiel, 1983; Nucci, 2001), moral violations share the property of involving unprovoked harm, right violation or injustice against an innocent agent, where "moral violations" are operationally defined as those instances of wrongdoing which are seen as general in scope (in terms of spatial and temporal applicability), authority independent, and more serious than other types of non-moral/conventional norm violations. Actions of such description are pan-culturally regarded as moral violations by children as young as 3 years old. The point here is not so much to defend or attack this view (for that, see e.g. Fessler et al. 2015 vs. Sousa & Piazza, 2014; Piazza et al. 2018) as to introduce one way of understanding its central claim within the present framework. The idea is that entertaining an abstract description of an action on top of a more "concrete" one, such as [A causes injustice to P], engages the FM in such a way that the action will be stored as morally forbidden.[41][42] On more ambitious proposals of this kind, *all and only* rules (faculty principles) of I-morality will share this abstract structural description (we will consider a recent theory in this vicinity in the next chapter). It may be an open question whether principles of derivation (if they exist) have a critical or sensitive period, or they remain operative in adults. This latter possibility (invariably assumed in the debates mentioned) is why derivation principles may overlap with processing principles: (a) they might be seen as constraint principles of a special kind, and (b) they might be taken to obviate the need for faculty or even

[41] Or it will enter a Sripada & Stich (2007) style norm database automatically (see Chapter 4, Section 4.2.2), that is, without the presence of the usual cues indicating the presence of a norm, such as different kinds of norm implicating behaviour. Again, such a story is speculative and provisional, and there is no assumption as to the mechanism that makes it the case that the relevant actions will be seen as morally significant.

[42] Notice also that there are two potential issues not strictly separated here: how a "more concrete" act description (such as *X kills Y*—which is still terribly abstract by the way) can be ontogenetically assigned a deontic status, on the one hand, and how an action is judged "on-line" in terms of this deontic status, on the other.

conflict principles altogether (the latter only if there is a *single* derivation principle, which is not the case in pluralistic theories, such as Haidt's, for instance).[43]

How do we know what the principles guiding moral cognition are? Perhaps we only need to ask. In fact, this reasonable looking assumption dominated some of the earliest attempts at a systematic scientific inquiry into moral cognition, such as Piaget's (1932) or Kohlberg's (1969, 1984), whose developmental theories were constructed on the basis of interviews in which children were asked to reason through moral scenarios in an attempt to discover the principles driving their moral judgment, such as Heinz's dilemma:

> A woman was on her deathbed. There was one drug that the doctors thought might save her. It was a form of radium that a druggist in the same town had recently discovered. The drug was expensive to make, but the druggist was charging ten times what the drug cost him to produce. He paid $200 for the radium and charged $2,000 for a small dose of the drug. The sick woman's husband, Heinz, went to everyone he knew to borrow the money, but he could only get together about $1,000 which is half of what it cost. He told the druggist that his wife was dying and asked him to sell it cheaper or let him pay later. But the druggist said: "No, I discovered the drug and I'm going to make money from it." So Heinz got desperate and broke into the man's laboratory to steal the drug for his wife. Should Heinz have broken into the laboratory to steal the drug for his wife? Why or why not?

Based on patterns in children's responses to such dilemmas, Kohlberg theorised that moral development consists of three levels (each of which he further subdivided into two "stages"). The classification went along the following lines. If the child appealed to Heinz's self-interest, then he or she would be said to be at the "pre-conventional" level (i.e. the lowest one). Second, appeal to what is accepted and required by the community would place the child at the "conventional" level. Finally, reasoning in terms of a social contract or universal ethical principles would be taken to indicate that the child is at the "post-conventional" level. Strikingly, *what* the child's actual answer was did not matter for Kohlberg's purposes, only the way the judgment (whatever it was) was justified.

This approach contrasts rather starkly with that assumed in this thesis. For a start, in linguistics, the principles driving grammaticality judgments are rarely if ever assumed to be available to conscious reflection, which is why surveys in grammaticality studies rarely even ask participants *why* they think a certain

[43] If they were descriptive theories, Kantian and utilitarian ethics could be understood as proposing their respective derivation principles; one very abstract principle explaining why particular acts are (seen as) morally right or wrong (*cf.* Section 6).

construction is (un-)grammatical.[44] Indeed, more recently, the Kohlbergian assumption has been quite radically challenged in empirical moral psychology due to the recognition that the justifications subjects provide for their judgments are often thoroughly insufficient to account for the patterns in their judgments (see e.g. Haidt, 2001; Hauser et al., 2007). (Haidt famously termed the related phenomenon whereby a person is surprised to find out that he or she cannot provide a satisfactory explanation of their judgment *moral dumbfounding*.)

To capture the difference between principles we appeal to in justifying our judgments and those that are causally responsible for the generation of the latter, Mikhail distinguishes between *express* and *operative* principles, respectively (Mikhail, 2011, *pp*. 19-21; see also Hauser, 2006, *pp*. 37-8 and Hauser et al., 2008a, *p*. 109, who talk about "expressed" and operative *knowledge*). It should be rather evident that the primary goal of a science of moral cognition is aimed at discovering the *operative* principles of moral judgment. As with many of the other distinctions introduced by LA, this observation need not depend on a *close* analogy between language and moral cognition. Indeed, the distinction is customarily observed in many areas of cognitive science, such as the study of probabilistic thinking, visual perception, or mathematical cognition. For instance, purported (operative) principles of probabilistic reasoning, such as the representativeness heuristic[45], are not discovered by asking experimental subjects to identify the inferential steps via which they arrive at their particular probability judgments, rather, it requires the systematic study of the patterns in their judgments. Nevertheless, as with other distinctions introduced in this chapter, it will be prudent to avoid being dogmatic about the divide. In particular, there are no obvious conclusive theoretical or empirical reasons in favour of the assumption that the two types of principles *cannot* collapse in particular cases—that is, it remains a live possibility that in some instances, operative principles may also be express—see for example Cushman, Young, & Hauser (2006), where some principles that accurately explained judgment also proved accessible to conscious reasoning. We shall return to this problem in Chapter 3.[46]

---

[44] Compare Chomsky regarding this issue: "Obviously, every speaker of a language has mastered and internalized a generative grammar that expresses his knowledge of his language. This is not to say that he is aware of the rules of the grammar or even that he can become aware of them, or that his statements about his intuitive knowledge of the language are necessarily accurate. Any interesting generative grammar will be dealing, for the most part, with mental processes that are far beyond the level of actual or even potential consciousness; furthermore, it is quite apparent that a speaker's reports and viewpoints about his behavior and his competence may be in error. Thus a generative grammar attempts to specify what the speaker actually knows, not what he may report about his knowledge" (1965, p. 8).

[45] According to the representativeness heuristic, "the subjective probability of an event, or a sample, is determined by the degree to which it: (i) is similar in essential characteristics to its parent population; and (ii) reflects the salient features of the process by which it is generated" (Kahneman & Tversky, 1972, *p*. 430).

[46] One reason for this possibility in the case of moral judgment is noted Malle and colleagues, who point out that, at least in the case of blame, "only if people generally have access to the informational basis of their blame judgments [...] can they demand, offer, and negotiate such information as warrants for their acts of blaming" (Malle et al., 2014, *p*. 149,

As an aside, I rather like the way in which seeing theories of moral psychology more explicitly as theories of different kinds of principles and their corresponding representational planes allows for the understanding and framing of old debates in new and potentially constructive ways. Consider, for example, "monist" derivational theories (those proposing a single derivation principle) and their potential implications about the nature of other kinds of principles. I mentioned above that such views, if correct, might obviate the need for faculty or conflict principles. Now we can add a qualification: this elimination would only reach as far as operative principles of those kinds. We clearly have a repository of express—both faculty and conflict—principles, which, if the derivation principle is unavailable to conscious reflection, might be the product of inductive inferences (*post hoc* rationalisation) over the situations in which we have found ourselves judging actions in a certain way. We talk about principles such as the prohibition against intentional homicide, but really, on a monist account it is not *homicide* that is the relevant structural description responsible for our moral judgment, but that, at least prototypically, by killing someone, we commit an injustice, as it might be (e.g. on Sousa and his colleague's deflationary theory, see Sousa et al. 2014, 2018).

Figure 1.2 below has been constructed to help make sense of—and navigate between—the distinctions introduced in this section.

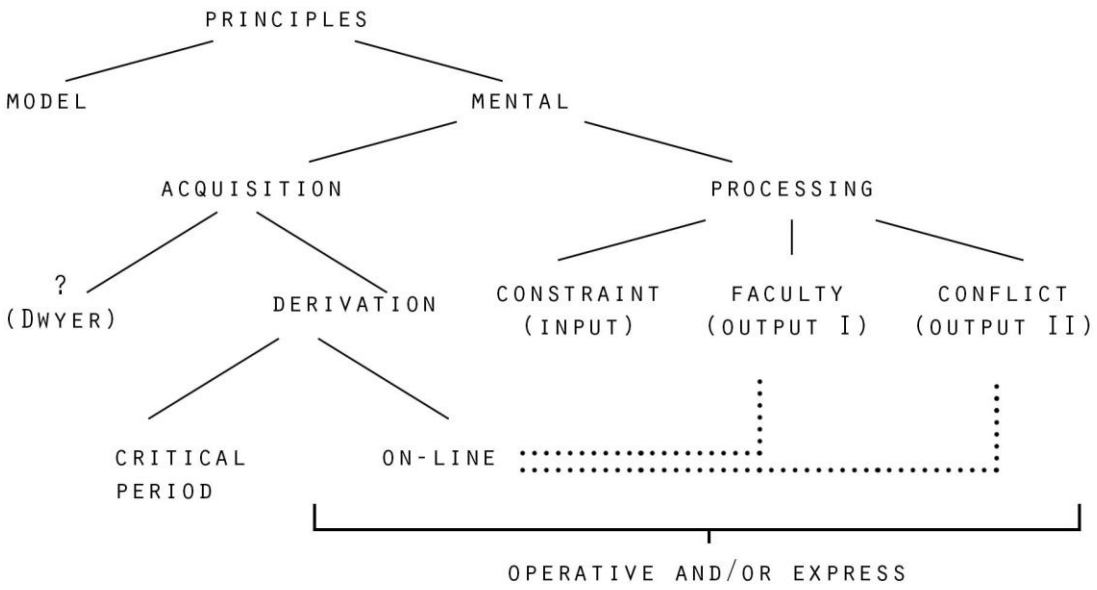

Figure 1.2: The proposed distinctions regarding the term *principle* in the context of the study of moral cognition. The first bifurcation (*model* vs. *mental*) captures the

fn. 2). Indeed, more generally, that many aspects of moral cognition ought to be consciously available more or less follows from some of its presumed functions, such as social persuasion (see e.g. Haidt, 2013; Sterelny, 2010). Nevertheless, such arguments are far from conclusive, of course.

characteristic systematic ambiguity of the Chomskyan framework (*cf*. the same type of ambiguity regarding the referents of terms such as *generative grammar* or *generative rule*; see the text): the scientist's task is to describe internal principles by constructing models of these. Model principles are asymmetrically dependent on internal principles, since the interest in the former is derivative of that in the latter. The second bifurcation (*acquisition* vs. *processing*) concerns the aspect of I-morality that is regulated by the principle. In the case of a processing principle, what is regulated is the (synchronic) state of FM, that is, I-morality. In the case of an acquisition principle, what is regulated is the (diachronic) development of FM. Acquisition principles may be subdivided between the "mysterious" Dwyer-type acquisition principles, and less mysterious derivation principles (the potency of which might not be limited to diachronic acquisition—see the main text). Looking at the third division of the rightmost branch, there are three types of processing principles (characterising I-morality): constraint (which can be understood as governing what representations are potentially subject to deontic valence assignment), output (which can be understood as rules governing the process of *what* deontic value is assigned to an action representation), and conflict (which are principles that guide the resolution of deontic conflicts). This latter trifurcation is based on both presumed temporal order and logical primacy of processing. Finally, each branch of the processing node further splits into operative and express principles, depending on whether they are appealed to in the course of justification of judgment or they actually guide it (or, potentially, both; see text). Note that the *model* branch replicates all the distinctions within the *internal* branch, although the nodes to the left have been omitted for the sake of simplicity.

6. Empirical standards of evaluation

On the basis of the preceding discussion, we might already have a few ideas as to how to assess (aspects of) a theory of the moral faculty. First of all, we noted the productivity problem, and the consequent requirement that a theory of moral competence specify some means by which to account for our ability to assess entirely novel situations in moral terms. We also concluded that the best way of doing so is by appeal to (not necessarily combinatorial) principles, such as the types of principles discussed in the previous section, as well as specifying the representations that those principles are defined in terms of. So we require that we have a set of judgments that humans actually make[47] and the characterisation of a set of principles and representations

[47] Ideally cross-culturally, but I promised I would ignore this complication.

from which these can be derived. This is very much as it should be, with the addition of a further criterion: an explanation of how these principles develop; more accurately, a characterisation of the inherent properties of the mind that enable the acquisition of the relevant principles and representations given the external information that is ontogenetically available to the organism. We shall briefly consider each of these steps in turn.

The first juncture at which the "adequacy" of a theory can be assessed is whether it provides a correct account of the relevant data points. A theory that satisfies this requirement is referred to as *observationally adequate*. In linguistic theory, the relevant data points are judgments of grammaticality. When a generative grammar (in the model sense) satisfies the criterion of observational adequacy, it lists sentences along with their grammaticality status and thus forms a sort of grammatically annotated corpus. In the domain of the study of moral cognition, on one conception of observational adequacy, the sentences are action descriptions, and grammaticality is replaced by a deontic status (obligatory, forbidden, etc.) that is attached to each of those descriptions.[48] Although this level of adequacy may appear rather trivial, it is logically necessary: it is what a theory of I-language/morality must account for.[49]

Another way of thinking of observational adequacy is not as requiring an exhaustive list of data points and grammaticality/permissibility judgments (which would not be possible to achieve anyway), but as a basis from which inquiry can proceed; after all, scientists and philosophers need to agree on what the data that needs to be accounted for are. Thus, observational adequacy can be achieved relative to a closed set of sentences/cases (see the example further below). Once we agree on the content of the relevant judgments, we can carry on to the more interesting part of the inquiry that involves theorising about the processes and mechanisms that synchronically or diachronically give rise to our dataset.

As for the synchronic task, given the productivity of both language and action representation (Section 3.2), what is required is to devise generalisations that not only describe but predict or generate the relevant data points as described by an observationally adequate theory.[50] These generalisations, as argued before, come in the form of a set of (posited) representations and principles defined over them. In the case of language, the representations are a closed set of functional—for example, syntactic or phonological—

[48] Of course, due to the productivity of both domains, the list can never be exhaustive.

[49] This is a narrow conception of the task of moral psychology; there are other legitimate questions to be asked (such as questions about responsibility and virtue, for instance).

[50] It should be rather obvious that a theory that only addresses the problem of observational adequacy can hardly be appropriately referred to as a theory. Maintaining the opposite would be tantamount to claiming that explanation in physics is a list of meter readings (to adapt Chomsky's critique of the practice of referring to psychology as "behavioural science" to a slightly different phenomenon).

categories bound to lexical items and the principles may be combinatorial, such as rewrite rules in syntax or phonotactic principles in the case of phonology (the latter of which define admissible combinations of phonemes taking into account, *inter alia*, place of articulation or position within syllabic structure). In the case of morality, the representations are also arguably from a closed class—often taken to be causally and intentionally structured action and outcome representations (e.g. Cushman, 2008; Greene, 2013; Mikhail, 2011), while the nature of the principles is best understood as derivational, faculty, conflict, or a combination of these, as argued above. When these criteria are satisfied, the theory has reached *descriptive adequacy*.

Two additional points are worth making with respect to descriptive adequacy. The first is that descriptive adequacy concerns the output; that is, it is all about producing the generalisations that predict or generate the data we observe. In principle, there can be many different descriptively adequate theories generating the same data points (in practice, producing *one* such theory for either language or moral cognition is an enormously difficult task). Second, on a fully demanding conception, a theory of language is descriptively adequate if and only if it generates all and only the grammatical sentences of a given language; and *mutatis mutandis* for moral cognition. This is an unrealistically high expectation with respect to language, not to mention moral cognition. It makes sense, therefore, to speak of descriptive adequacy regarding a more circumcised set of cases or judgments (this is the strategy in Mikhail, 2011, for example)—just as I suggested above with respect to observational adequacy.

Sticking with language for the moment, it is perhaps easy to see that a system of principles and representations that satisfies both observational and descriptive adequacy still does not constitute a complete theory of language from a cognitive science point of view. This is because the ultimate explanandum for linguistic theory in this framework is the language faculty, which is not accounted for even if we have an observationally and descriptively adequate grammar for all known languages (which of course we do not). A complete theory would have to explain how FL allows for the development of particular I-languages given the linguistic data children are exposed to (that is, PLD as mentioned before). This is the diachronic task discussed in the previous paragraphs. A successful theory satisfying this criterion (over and above observational and descriptive adequacy) could be said to have achieved *explanatory adequacy*, the deepest constraint on a theory of language. (In the parlance of the previous section, the goal of explanatory adequacy is to discover the acquisition principles that, when conjoined with PLD, explain why we end up with the particular language specific grammars posited by descriptively adequate theories.)

Turning our attention to the domain of moral cognition, informally, aiming for explanatory adequacy requires an explanation of how children develop into creatures having the capacity to think and reason in

moral-deontic terms; or minimally, to make moral judgments concerning permissibility or right and wrong given the relevant data they are exposed to during ontogeny; that is, what we might refer to as the "primary moral data" (or PMD).[51] Less informally, the goal is to discover the principles guiding acquisition (i.e. acquisition principles) that, conjoined with PMD explain and predict the development of particular I-moralities (as characterised by representations and processing principles). As in the case of language, in fact even more so, as things stand, this is more of an ideal given the current state of inquiry (and, again, not for lack of trying).

I mentioned above that it is *in principle* possible to have two (even fully) descriptively adequate theories at the same time, namely, those that generate the "list" as described by observationally adequate theories (and beyond). On Chomsky's view, explanatory adequacy enables us to choose between such a hypothetical set of descriptively adequate grammars in a principled way; that is because a theory satisfying the criterion of explanatory adequacy *explains* why a particular grammar is successful out of a set of extensionally equivalent grammars of a particular I-language.[52] Chomsky certainly appears to hold not only that explanatory adequacy helps select between such extensionally equivalent grammars, but also that the *only* way to arbitrate between them is by reference to an explanatorily adequate theory of language, which, if true, would be unfortunate given the rather unrealistic prospect of developing such a complete linguistic theory.[53] And, as usual, the situation would be even more difficult in the case of moral cognition.

Nevertheless, I would argue that it is possible to disagree in a meaningful way more or less independently of questions about explanatory adequacy over which of two theories describes better the mechanism which results in a set of observations about linguistic or moral judgment (and indeed this has been the case in the

[51] The question of just what it takes to be a creature capable of moral judgment is subject to much controversy, made even more difficult by the dubious legitimacy of determining the answer to this question ahead of inquiry. Here, in line with the discussion in Section 3.3.2 (see also Joyce, 2006), we concentrate on the problem of perception.

[52] In fact, an explanatory adequate theory rules out all the other grammars: such grammars would fail to satisfy UG and would thus constitute descriptions of humanly impossible hypothetical languages despite the hypothesised extensional equivalence. Of course, this is a rather theoretical point, since, as alluded to above, the existence of such extensionally equivalent grammars is highly implausible given the complexity of I-languages.

[53] This is perhaps the most relevant quote from *Aspects*: "on the one hand, the grammar can be justified on external grounds of descriptive adequacy—we may ask whether it states the facts about the language correctly, whether it predicts correctly how the idealised native speaker would understand arbitrary sentences and gives a correct account of the basis for this achievement; on the other hand, a grammar can be justified on internal grounds if, given an explanatory linguistic theory, it can be shown that this grammar is the highest-valued grammar permitted by the theory and compatible with given primary linguistic data. In the latter case, a principled basis is presented for the construction of this grammar, and it is therefore justified on much deeper empirical grounds. Both kinds of justification are of course necessary; it is important, however, not to confuse them. In the case of a linguistic theory that is merely descriptive, only one kind of justification can be given—namely, we can show that it permits grammars that meet the external condition of descriptive adequacy. It is only when all of the conditions [of explanatory adequacy] are met that the deeper question of internal justification can be raised" (Chomsky, 1965, *pp*. 40-41; see also *pp*. 18-27).

literature over moral judgment—see the next chapter). That is, since descriptive adequacy is best described as concerning the functional level of explanation (see Marr's levels in Section 4), and since, as in the case of vision, a full computational explanation of a psychological mechanism simultaneously addresses both the functional and the algorithmic levels (without necessarily relying on solving problems concerning the problem of acquisition), it seems reasonable to expect that a theory (whether of language or morality) go beyond achieving descriptive adequacy in this functionalistic sense. (Although of course, to reiterate, in practice, different theories will not be strictly equivalent, and they will usually make divergent predictions at least in some cases.) To mark this point, in what comes below, I will refer to a theory that is both descriptively adequate in Chomsky's sense and is also successful in terms of the postulation of mechanisms that explain why the theory is descriptively adequate (without thereby implicating a solution to the problem of explanatory adequacy) as having achieved "strict" descriptive adequacy.

To put some flesh on this rather abstruse discussion, let me introduce a toy example; sentences (1)-(4).[54] Consider the following ways in which a theory of language may proceed. First, it requires a specification of the set of sentences that are well-formed or not well formed in a given language (full observational adequacy) or regarding a set of sentences (partial observational adequacy). Thus, an observationally adequate theory will state that sentence (4) is ungrammatical (and so is (1) on the reading on which *her* refers to the subject of the sentence).[55] It will also state that sentences (2) and (3) are grammatical (and so is (1) on the reading on which *her* does *not* refer to the subject of the sentence). A descriptively adequate theory would analyse each of sentences (1)-(4) in terms of the relevant abstract categories instantiated by them as well as their structural arrangements, and posit principles accounting for the grammaticality status of each, and many more besides (fully descriptively adequate theories would do this for all sentences of the given language, and a strictly descriptively adequate one would actually posit the representations and principles that are operative in the mind).[56] Finally, explanatory adequacy requires (acquisition) principles that, in their interaction with PLD, entail the ontogenetic emergence of the principles and representations as described by the (strictly) descriptively adequate theory.

[54] A similar set of examples is also provided by Greene (2006).

[55] Indices in common indicate coreference. I used the exclamation mark to indicate that grammaticality depends on coreference. The asterisk indicates that the relevant sentence is ungrammatical (as is the usual practice in linguistics).

[56] Thus, we can see that full and strict descriptive adequacy concern different dimensions: extent of completeness and degree of mechanistic isomorphy, respectively.

(1) !Beth$_i$ finds her$_{i/j}$ cute

(2) Beth$_i$ finds herself$_i$ cute

(3) Beth$_i$ thinks that she$_{i/j}$ is cute

(4) *Beth$_i$ thinks that herself$_i$ is cute

More concretely, as for descriptive adequacy, the grammaticality of (1)-(4) is determined by reference to principles regulating the behaviour of anaphors (such as reflexive pronouns like *herself* and reciprocals like *each other*) and non-anaphoric pronouns or pronominals (such as pronouns like *her*). What we are interested in with respect to (1)-(4) is what structural elements *her*, *she*, and *herself* can take their reference from, that is, what is a potential "binder" for these words; in these sentences, whether *Beth* can bind *her*/*she* or *herself* (or whether these lexical items can refer to Beth). In government and binding theory (e.g. Chomsky, 1986), the relevant principles are articulated in terms of locality.[57] For present purposes, let us assume that local binding means binding within the same clause (in the first two cases, the anaphor/pronoun is in the same clause as the subject NP, in the last two cases, it isn't). The first principle states that anaphors (e.g. *herself*) must be bound locally (this is more or less Chomsky's "Principle A"). The second principle states that non-anaphoric pronominal expressions (e.g. *her*) must be free within their local domains (Chomsky's "Principle B"). Notice that these principles explain the grammaticality of arbitrarily many sentences beyond the ones under consideration.[58]

The challenge of explanatory adequacy in this case is to propose principles of sufficient abstractness to capture the behaviour of anaphors and pronouns not just in English but in all actual and potential languages, and to do it in a way that, when conjoined with PLD, they provide enough information to the child to acquire the principles regulating the distributional patterns of these linguistic elements in his or her relevant I-language. The introduction of the notion of a single local binding domain is a step in this general direction.

Having considered the standards of evaluation, and before moving on to the next chapter in which I demonstrate the usefulness of LA and defend it against criticisms advanced in the literature, let me return to principles for a moment. The above example illustrates that, in the case of linguistics, the principles specific to (I-)languages are often accounted for by more general (model-acquisition) principles that are supposed to be broad enough to capture the variation exhibited by all (actual and possible) languages—and thus to solve

[57] Additionally, the NP (here, *Beth*) and the anaphor/pronoun have to "agree" in terms of person, number and gender (thus, *Beth* can potentially bind *her*, while it couldn't even potentially bind *him*).

[58] Two examples: (a) 'Sophie$_i$ thinks that Lewis$_j$ should marry herself$_i$' is predicted to be ungrammatical because *herself* is not in *Sophie*'s binding domain; (b) 'Lewis likes him' is predicted to be grammatical only if *him* is not coreferential with *Lewis*, since, as a pronoun, *him* cannot be bound in its local binding domain.

the problem of explanatory adequacy. In this tradition, (processing) principles may be seen as something like special cases of acquisition principles, which is another way of illustrating the general point advanced in Section 5 above, namely that the two kinds of principles (acquisition vs. processing) are intimately related. Whether this is a useful way of thinking of the relation between acquisition and processing principles in the moral domain remains to be seen.

Having shown what LA has to offer, in the next chapter, after considering and criticising one alternative framework for the study of moral cognition, I examine five objections to LA and reply to each of these.

Chapter 2: The Linguistic Analogy—Comparisons, Objections and Replies

Overview

In the previous chapter, I introduced and articulated LA as a viable approach to the study of moral cognition. In the present chapter, I do two things. First, I introduce and evaluate a popular theory of moral judgment—Joshua Greene's dual-process theory—which Greene articulates within an influential alternative framework—namely, the Dual Process framework ("DP"). I criticise both the theory and the framework, and conclude that LA is superior to the latter. Second, having made a case for LA vis-a-vis DP, I defend LA from popular objections that have been levelled against it in the literature.

I proceed as follows. In Section 1, I present Greene's dual-process theory in some detail, including its motivations and how it fits into the broader DP framework. Then, in Section 2, I survey some of the empirical evidence advanced in favour of Greene's theory, and assess the extent to which such evidence makes a good case for the theory—which, as I will argue, is limited. In Section 3, I introduce a central problem for the dual-process theory (as well as other theories of moral cognition), namely, the appraisal problem, which asks a question about the nature of the connection between our intuitive judgments and the stimuli eliciting them. In the same section, I consider Greene's response to this problem and assess whether it is motivated either by his theory, or the DP framework in general. My answer will be in the negative. At the end of this section, I compare LA favourably against DP. In the final section—Section 4—I consider five objections levelled against the Linguistic Analogy and reject them one by one.

1. The dual-process theory: Background and motivations

1.1. The bare bones

According to Joshua Greene's dual-process theory (Greene, 2013; Greene et al. 2001), moral judgment relies on two separate cognitive systems: an emotion system, concerned with how we feel about a particular action, and a reasoning system, concerned with bringing about the best possible outcome. The idea in a nutshell is that when we confront a "moral" scenario containing a potential or actual action, the two systems

may be engaged to different degrees in evaluating that action in terms of whether it is obligatory, permissible or forbidden.[59] Thus, we have two potential outputs emanating from the two systems. These may be identical or divergent. In the former case, that is, when the outputs support the same judgment, other things being equal, that judgment is made by the individual. In the latter case, that is, when the outputs are not convergent, a conflict resolution mechanism resolves the conflict depending on various factors, such as the individual's relative reliance on the respective systems (i.e. the "cognitive style" of the individual), or the "strength" of the respective outputs. In both cases, the integrated output (or decision) is what we normally refer to as a *moral judgment*.[60]

As with theories of moral judgment in general (*cf.* Chapter 1, Section 6), the dual-process theory takes as its primary explanandum a closed set of cases, and it generalises on the basis of these. In the case of Greene's dual-process theory, the set of cases is that mentioned in the previous chapter (Section 3.2.2), namely, the pair of dilemmas making up the so-called "trolley problem". This set of well-known moral dilemmas was initially conceived of by Philippa Foot (1967) and subsequently developed and elaborated by Judith Jarvis Thomson (1985) and others.

In the first scenario, an empty trolley[61] is heading towards five people, who will be killed if nothing is done. However, a bystander (call him 'Hank') happens to be standing next to a switch, which, if thrown, turns the trolley onto an alternative track, where there is one person, who will be killed if Hank decides to act. At issue is whether we judge it permissible for Hank to throw the switch, thus saving five people but killing the person on the alternative track. Most of us do—including lay people as well as philosophers, such as Foot or Thomson themselves.[62] Let us refer to this first scenario as the *switch* case.

In the second scenario, there is no switch or alternative tracks. This time, a person (call him 'Ian') is standing on a footbridge, where there is a very large man standing in front of him with an enormous rucksack. If Ian pushes him off the bridge and onto the track, his body will stop the trolley, and the five people on the

[59] Of course, as per Chapter 1, Section 3.4, there is a perception and a production version of this question. I shall ignore that complication here.

[60] Notice right at the outset that it is not entirely clear what makes such a *judgment* moral, beyond the fact that the situation about which the judgment is made is seen as a morally charged one (e.g. when we are asked to make a decision that has fatal consequences for innocent people—see below). On Greene's theory, this is not necessarily a shortcoming, however, since Greene holds that moral judgment is unified only at the evolutionary-functional level, not at the cognitive level—that is, any of Marr's levels (discussed in Chapter 1, Section 4). (See Greene, 2013, 2015; see also Section 4.2 and Section 4.5 below for related comments).

[61] A trolley is what is referred to as a 'tram' in UK English and so in Foot's original paper. I'll stick to the term 'trolley' though, because that is the description under which the dilemmas became well known.

[62] Thomson later changed her mind, however. See Thompson (2008).

track will survive. However, the large man will die as a result of Ian's action. The question, again, is whether we judge it permissible for Ian to push the man off the bridge, thus saving the five but killing him. Most of us do—including philosophers as well as laypeople. Let us refer to this second scenario as the *footbridge* case.

Explaining the contrasting moral judgments in these two cases constitutes (the original version of) the "trolley problem".[63] Although in both cases, the outcome seems to be exactly identical (one man dead, five saved—or *vice versa*), as mentioned above, it is common to judge that it is permissible to act in the switch case, but it is mostly seen as impermissible to act in the footbridge case (see e.g. Greene et al. 2001; Cushman et al., 2006; Hauser et al. 2007; though see also Ahlenius & Tännsjö, 2012).[64]

The dual-process theory hypothesises that the two proposed cognitive systems may have divergent outputs with respect to the trolley cases. The reasoning system calculates the best possible outcome. In both cases, it calculates that 5 deaths is worse than 1, so in both cases, its output is a rationale in favour of performing the action—or judging that the performance of the action is acceptable or required. However, and here is where the difference lies, the emotional system responds in a different way to the two scenarios. In the switch case, it does not get activated (or it only does so to a small extent); in the footbridge case, it gets activated and shouts "NO!", thus potentially/typically overriding the response of the reasoning system. *Ex hypothesi*, this is because in the footbridge case, but not in the switch case, the action involves being close to and pushing the individual, which is proposed to be emotionally more salient (see Section 3 for more).

Greene observes that the output of the reasoning system happens to be consistent with what is most plausibly prescribed by consequentialism, whereas the output of the emotional system is clearly non-consequentialist. From now on, I follow Greene (2014) in referring to these two types of output as "consequentialist*" and "non-consequentialist*" (with asterisks), which does not imply that these judgments are the result of endorsing such theories. In other words, Greene's terms "consequentialist*" and "non-consequentialist*" are behaviourist constructs: they are neutral as to whether the motivation behind the judgment is in fact consequentialist or non-consequentialist in nature; for example, they are independent from whether or not the judgments have been made in the light of, say, the utility principle.

---

[63] Strictly speaking, this is the moral psychology version of the trolley problem. The normative version asks what is in fact morally permissible to do in these cases. Note also that some authors use the term *dilemma* to refer to the cases separately (hence "trolley problem*s*").

[64] As emphasised in Chapter 1, we should distinguish between questions of moral psychology and those of normative ethics. We may be tempted to ask: OK, but do most people get it right, or are most of us wrong? Such questions are normative questions, however, and we will not be concerned with them here. Instead, in the context of moral psychology, Chapter 1 urges that we should ask questions of the following type: What is the relevant difference between the mental representation of the two cases that, if the relevant principles are applied to them, they yield different judgments?

As mentioned before, the two proposed systems are thought to work independently of each other. Nevertheless, their outputs need not be in direct competition. In many cases, they can be expected to provide identical behavioural or decision-making rationales. For example, if the question were whether it is permissible to throw the switch if the numerosities of the potential victims were reversed, it is very unlikely that anyone would judge any of the actions permissible, since neither system would be expected to produce such a response. (This is indeed the case as confirmed by Mikhail's "disproportional death" scenario—see e.g. Mikhail, 2009.) Note also that the claim is not that both systems are always engaged, and sometimes they compete, sometimes they do not, either. On the contrary, the crucial difference between the switch and the footbridge cases is supposed to be that in footbridge, but not in switch, the emotional system provides a strong negative output. In the latter case, therefore, the "emotional" system is either silent or simply not engaged to a sufficient degree to compete successfully with the reasoning system. (As alluded to above, this formulation allows for individual variation as well; and indeed some people do say that the action is impermissible even in the switch case.)

1.2. A tale of two systems: The DP framework

The general theoretical framework that provides the background for Greene's dual-process theory of moral judgment is a general dual process (or dual "systems") theory of the mind, which I will refer to as "DP" or the "DP framework" below. There are dual-process theories of memory (Atkinson and Shiffrin, 1968; Reber, 1993), attention (Schneider and Schiffrin, 1977), reasoning, decision making and social cognition (Evans, 2008), as well as "broad" dual-process approaches to the human mind in general (Kahneman, 2011). What they all emphasise is the distinction between two different types of cognitive processes (or sometimes systems), often referred to as *automatic* and *controlled*.[65]

Automatic processes are fast, effortless and (not so surprisingly) automatic, meaning that their operation does not require any/much attentional resources, central access or conscious control. Furthermore, they involve parallel—i.e., simultaneously executable—processing, meaning that dissociated (i.e. separate) automatic processes generally do not interfere with each other. For instance, looking at an array of three dots, we realise immediately that we're seeing three dots (*cf.* Chapter 1, Section 3.1.1). There is no need to attend to the dots individually, there is no need to count them, and in fact there is no need to pay attention to them, either. The realisation that there are three dots in front of us just happens to us, as it were, even if we are

[65] The terms "automatic" and "controlled" were first introduced by Schneider and Shiffrin (1977—also, Shiffrin & Schneider, 1997), whose work primarily concerns attention. Their papers have had a considerable effect on other types of dual-process theories (but see Frankish & Evans, 2009), especially those in the domain of social cognition.

engaged in another cognitive task, such as talking to someone on the phone or wondering about what we need to buy in the supermarket.

Controlled processes, in contrast, are slow, often effortful, involve central access and require conscious attention. For example, finding the derivative of a function involves consciously applying a rule. It is by no means automatic, and even after a considerable amount of practice, it may still take a relatively long time (depending, for example, on how complex the function is). Furthermore, it is close to impossible to do any other task requiring controlled processing at the same time. This is thought to be due to the limited capacity and serial (as opposed to parallel) nature of controlled cognition. For example, imagine having to find the derivative of a function while at the same time trying to reconstruct the ontological argument. In contrast to the previous examples, we also have much more access to how we execute tasks requiring controlled processing. (In the parlance of Chapter 1, Section 5, these processes are both operative and express.)

The reason why the two kinds of processes have these different (to some degree opposing) characteristics makes sense in the light of the trade-off between efficiency and flexibility to which they are thought to provide complementary solutions. Greene proposes that automatic processes invariably achieve their efficiency by relying on statistically reliable cues to respond to the things they are designed to process information about.[66] An example is provided by Paul Whalen and his colleagues, who investigated how the amygdala (a collection of nuclei in the medial temporal lobe) rapidly responds to fearful facial expressions (Whalen et al. 2004). They found that the amygdala is engaged even if only eyes (with all other facial features removed) are presented and there is no conscious awareness of the presence of them.[67] Thus, instead of a holistic analysis of faces, the mechanism relies on the presence of enlarged eye-whites as a shorthand. In contrast, controlled processes do not typically rely on such cues, and thus provide us with a less epistemologically dubious source of knowledge. (Enlarged sclera may be useful for detecting fear rapidly in the social environment, but the correlation between enlarged sclera and fear is at best imperfect.)

The contribution of the DP framework to Greene's dual-process theory should be apparent enough. On Greene's theory, two proposed systems function in accordance with the respective operating properties of the two kinds of processes identified above. The emotion system[68] processes information automatically, it is fast,

[66] "All automatic settings rely on specific cues that are only imperfectly related to the things they're designed to detect" (2013, *p.* 227).

[67] This was achieved by "masking" the eye-stimuli, that is, after a 17ms long presentation of the eyes, a masking stimulus was introduced for the duration of 183ms. All subjects reported that they were not aware of the presence of the masked stimulus.

[68] Greene frequently uses the term "automatic settings", which denotes a broader category, however, namely all automatic psychological mechanisms, especially the innate ones. In contrast, the term "emotion system" is used here to

parallel (many tasks at a time), efficient, inaccessible and relies on simple cues for its operation.[69] Hence the frequently documented disconnection between emotionally induced judgments and explicit rationales that could plausibly give rise to them (Haidt, 2001). For example, it is easy to say that pushing the man off the footbridge is wrong, but it is much more difficult to specify what exactly the relevant distinction between the footbridge and the switch scenario is that explains their differential status in terms of judged permissibility (Cushman et al., 2006). This indicates that whatever the explicit rationale, it is not causally relevant from the point of view of the judgment (at least not in its explicit form). By contrast, the reasoning system processes information in a controlled way (i.e. not automatically), it is slow, serial (one task at a time), inefficient and its operation is largely consciously accessible. Thus, we may reason our way to the belief that the two cases (footbridge and switch) are morally equivalent (i.e. equivalent in terms of permissibility), but this may take time and effort.

Importantly, Greene's account of automaticity may be taken to defuse the Argument for Moral Grammar, the central argument of LA (Chapter 1, Section 3.2). This is how: judging novel actions in moral terms might not require an analysis of action in terms of abstract representations and principles. Instead, it may be based on simple cues that some actions manifest. Such a simple knee-jerk-type process could in principle rely on any easily perceivable and reliable cues (whatever reliability means in such a context). For instance, it could be that we judge actions that are performed by agents with three legs as impermissible. By such a procedure, we could judge a potentially infinite set of actions in terms of permissibility. More realistically, there *might* be such a cue based difference between the relevant actions in the switch and the footbridge cases. If so, LA's emphasis on principles defined over abstract representations appears overstated and potentially entirely superfluous.

1.3. Evolution and the dual-process theory: A small detour

A guiding notion of the DP framework (especially Greene's version of it) is that certain apparently complicated problems can have simple solutions. One example is provided by the Whalen study concerning fear detection mentioned above. An interesting parallel to this idea comes from evolutionary game theory. The problem of cooperation is a fundamental problem in such diverse scientific disciplines as evolutionary biology or economics: if agents are in general expected to behave according to their best interests, rather than those of others, how is it possible for cooperation to emerge (either in the animal kingdom or social systems),

denote the system that is engaged in the footbridge as opposed to the switch case, without any claim to the effect that this system can be non-controversially individuated (*cf. fn.* 60).

[69] Unconscious in terms of its operation, not in terms of its output, of course.

given that cooperation (at least the interesting kind) always requires individual sacrifice? One of the most popular ways of illustrating the nature of this problem is the Prisoner's Dilemma (PD).

Imagine that two bank robbers are arrested after robbing a bank. The police, however, cannot prove their involvement in the robbery and can only charge them with a less serious crime, let us say embezzlement. They are taken to two separate rooms, where the police offer both of them a deal: if they testify against the other bank robber, the minor charge against them will be dropped (they can get off scot-free) – at least as long as the other bank robber doesn't also testify, in which case both will receive a relatively serious prison sentence (say, 5 years). If one of them keeps silent while the other testifies, however, the stubborn one will have to take all the blame for the robbery, resulting in 10 years of prison sentence – the worst possible outcome. Finally, if both of them keep silent, then they will both have to face a less significant prison sentence for the lesser charge (say, 2 years).

The dilemma is the following: collectively, they are better off if both of them keep silent, that is, if they *cooperate*—with each other, not with the police, of course. However, individually, they are better off if they do testify, a move that is referred to as *defection* in the literature (see e.g. Axelrod, 1984). Moreover, not only are they better off if they defect, they are better off *irrespective of what the other prisoner does*. That is because if prisoner B cooperates, then prisoner A can go free, while if prisoner B defects, defection for prisoner A is the only way of avoiding the worst possible outcome. If they both succumb to the irresistible force of this piece of reasoning, however, they both find themselves in a situation that is far from the best possible outcome—either individually or collectively. In fact, if only they both cooperated, they would both be better off individually as well as collectively.[70]

This dilemma is not applicable exclusively to bank robbers, of course. In fact it captures the logic of many (if not all) cooperative situations, including collective problems, such as overfishing, use of publically available resources (such as social security), military standoffs, as well as quotidian situations where there is something to be gained from other people's help and much to be lost from being taken advantage of. The crucial factor is the ordering of the outcomes. To be more specific, for a situation to qualify as an instance of PD, taking

[70] This tension is perhaps best captured by the application of two important concepts in game theory that can be used to assess the degree to which a pair of strategies (and the resulting outcome) is "optimal" (at least in a certain limited sense): *Nash equilibrium* and *pareto-optimality*. To simplify a little, a Nash equilibrium is a situation in which no player gains anything from changing their strategy—that is, it is a combination of the players' "best" responses to each other's strategies. It is easy to see that the *only* Nash-equilibrium in situations sharing the logic of the Prisoner's Dilemma is mutual defection. A pareto-optimal outcome is one that is *not* pareto-dominated by any other outcome. An outcome is pareto-dominated by another outcome if the second is an improvement over the other at least for one of the parties involved and does not put any of the other parties at a disadvantage. In the PD, all of the outcomes are pareto-optimal, *apart from mutual defection*.

advantage of the other party (*T* for "temptation") has to be more advantageous for the individual than mutual cooperation (*R* for "reward for mutual cooperation"), which, in turn, has to dominate mutual defection (*P* for "punishment for mutual defection"), which, finally, still has to be preferable to being taken advantage of (*S* for "sucker's payoff"). In sum, T < R < P < S (see Figure 2.1 below).[71] The gains, in the meantime, can be positive, negative, relatively insignificant or of great importance—none of these factors changes the essential logic of the PD.

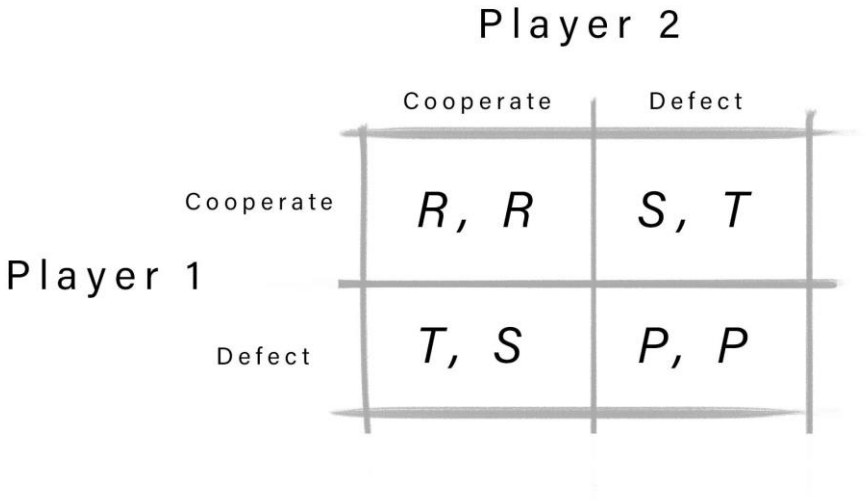

Figure 2.1: The payoff matrix of the Prisoner's Dilemma.

Superficially, no matter how the game is presented, it may seem like it cannot be solved, at least certainly not in a mutually beneficial way, since defection is the only rational strategy. However, when the mathematician and political scientist Robert Axelrod invited scientists to submit strategies for playing PD games in the form of computer programs, not only were a number of programs rather successful at playing the game, but the top ranking eight strategies were all "nice", in the technical sense of never being the first to defect in any encounter (ranking was based on the number of points they gathered in a series of encounters). Not only that, but only the eight best strategies were in fact nice, all the rest were "nasty" (meaning that they defect at times, even if "unprovoked"). The simple reason is that the logic of the game changes when it is iterated, that is, when players engage in encounters repeatedly. In such circumstances (at least when the repetitions are not predictably finite), if a player (be it a computer program or an organism) can monitor the other player's moves and keep track of them, it can take advantage of cooperative relationships by building up "trust" and thus repeatedly reaping the reward for mutual cooperation. On the opposite side of the coin, if the other player is also capable of keeping track of the first's moves, it will also be capable of "retaliation"

[71] Technically speaking, (T + S)/2 also has to be less than (or at most equal to) R.

(i.e. defection in response to defection), when the first player is acting in an uncooperative way. Consequently, the price of defection for the other party increases.

Surprisingly, Tit-for-Tat, the winning program submitted by the psychologist Anatol Rapoport, is not only nice, but also deceptively simple—so much so that its strategy can be captured by a short statement: "Cooperate first, then copy what the other player did last". Thus, Tit-for-Tat does satisfy the criterion that a successful strategy playing the iterated PD should be able to keep track of the other player's moves, but it does so in the simplest possible way. In fact, in Axelrod's competition, it was the shortest program to be submitted (shortest in terms of the number of internal statements of the program). The PD provides a good example of a complex problem being solved by a simple mechanism.

To return to our original concern, Greene's claim is that the emotion system is constituted by a set of devices that collectively solve the (iterated) Prisoner's Dilemma, or more broadly, the problem of cooperation (at least *within* groups). In other words, moral emotions regulate our social interactions with others so that we (tend to) avoid the worst outcomes in cooperative situations (such as punishment for mutual defection, or the sucker's payoff in the Prisoner's Dilemma—or more generally, in situations structurally isomorphic to it). They also do it in a simple way, just as Tit-for-Tat does: instead of having to solve the complex dilemma in an effortful way by reasoning about what the best strategy might be, moral emotions compel us to act (in Greene's characteristic words, they "do this thinking for us"—2013, *p*. 62) in a way that typically benefits us in the long run. (It is worth mentioning that Greene is by no means the first to propose the idea that human moralistic emotions may serve such a regulatory/strategic evolutionary function. Most notably, see Trivers, 1971; Frank, 1988.)

Consider the situation in which the prisoners cooperate mutually. Apart from the material gain (in this case, a relatively minor prison sentence), the prisoners develop mutual trust and gratitude (or even friendship), internally offsetting the temptation to defect. In the case of mutual defection, on the other hand, mutual contempt is expected, both parties ensuring that they cannot be taken advantage of (avoiding the sucker's payoff). The two symmetrical mixed outcomes may trigger shame or embarrassment in one party, while disgust or anger in the other. The former may be understood as a signal to the cheated party, expressing a desire that the cooperative relationship is to be restored, while the latter two may function as an avoidance

mechanism and as a potential threat for the other party (increasing the price of unilateral defection), respectively.[72]

Thus, these strategic (or moral) emotions fit the profile of automatic processes, as described above, just as the emotions elicited by the footbridge dilemma did. We have no insight into how (and why) we experience moral emotions: we just do so, and the fact that we do so seems the most natural thing in the world. Recall the trolley cases: pushing the man off the footbridge is just obviously (intuitively) wrong. Similarly, feeling gratitude when someone helps us is so natural it would be difficult to imagine it could be any other way.

1.4. The efficiency-flexibility trade-off

As suggested above, another point of convergence between such evolutionary accounts of moral emotions and Greene's theory of moral judgment is that both emphasise the simplicity of the processes involved—at least as far as emotions are concerned. Navigating the social world is a complex problem, but as Axelrod's Prisoner's Dilemma competition illustrates, complex problems can have simple solutions.

Recall Greene's distinction between the two systems. It is possible to figure out by conscious reasoning what the best rational strategy is, but this may be costly and inefficient, especially if simple rules of thumb can result in comparable success. Incidentally, this observation applies just as well to situations one would be less reluctant to identify as moral. In such situations, one may consult a normative theory prescribing that a certain moral principle be followed (e.g. "act only according to that maxim whereby you can at the same time will that it should become a universal law," or "always maximise expected utility"). Yet figuring out what acts are consistent with such principles may be a difficult task requiring reflection and conscious deliberation, and this is rarely feasible in many (perhaps most) everyday situations, when we have neither the time nor the capacity to sift through such complex chains of reasoning.

On the other hand, to the extent that from time to time we also need to act more flexibly (e.g. when we encounter novel problems—i.e. ones which no automatic settings were designed to solve), we cannot always rely on such simple solutions, and for a simple reason. As pointed out above, the two systems are thought to provide complementary solutions to a trade-off problem between efficiency and flexibility. If the emotion system provides solutions in the way Greene proposes, the mechanisms involved cannot modify our responses according to changing circumstances. But *we* can. In an alternative world which is full of people with enlarged

[72] These latter emotions are in line with subsequent developments in abstract theorizing about cooperation and successful cooperative strategies, which include the incorporation of "punishment", "forgiveness" or "reputation" into the models (see e.g. Nowak & Sigmund, 1998).

scleras, the social fear detection mechanism described above would presumably be continually responding,[73] but a control mechanism ("I know not all these guys are afraid") could override this, likely producing more adaptive behaviour. To consider a more realistic scenario, take hunger. Hunger is a motivational state produced by an "automatic setting" designed to detect when the body is in need of sustenance. The proximate function of the hunger-producing mechanism is to engage other mechanisms involved in food-seeking behaviour. In some circumstances, food-seeking behaviour can be counter-productive, however. For example, I might need to finish my chapter before the deadline, which, let's say, is in 10 minutes' time. In such a situation, the best response seems to be to suppress the behavioural command temporarily. This requires some degree of flexibility, which is exactly what the automatic settings are incapable of on Greene's account. (I will return to this observation in the next section, where the examples used will be from the domain of moral cognition.)

To summarise the important points of this subsection briefly: the main explanandum of the dual-process theory are the differential judgments elicited by the switch and the footbridge cases (and structurally identical moral dilemmas). The explanans goes along the following lines. There are two proposed systems: the emotion system and the reasoning system. The former is hypothesised to be engaged in the footbridge case but not in the switch case. In contrast, the latter is hypothesised to be engaged in both scenarios, the output being the same rationale (acceptable), because the outcome (what the reasoning system is hypothesised to "care" about) is identical in both cases.

In the following section, I discuss what I take to be Greene's flagship examples of the empirical evidence in favour of the dual-process theory of moral judgment (Section 2.1). I first present the evidence (2.1), then point out some empirical problems with it, on the basis of which I argue that none of these data provide powerful evidence in favour of the dual-process theory of moral judgment (2.2). After that, I consider a general problem for the dual-process theory (as well as other theories of moral judgment), namely, the *appraisal problem* (Section 3.1). Then, I consider Greene's response(s) to the appraisal problem (Section 3.2). Finally, I discuss to what extent Greene's theoretical elaborations are motivated by either the dual-process theory itself, or by the DP framework more generally (Section 3.3). Having concluded that the explanatory and theoretical contributions of either are rather minimal, I move on to the defence of LA from direct criticism (Section 4).

[73] Unless of course it is capable of adaptation.

2. The dual-process theory: An appraisal of the evidence

2.1. The flagship examples

So far, the claim that the switch and the footbridge dilemmas engage the two systems (and particularly the emotion system) differentially has been just that: a claim. To address this concern, Greene et al. (2001) presented subjects in an fMRI scanner with a series of intuitively moral—as well as some intuitively "non-moral"—dilemmas. The subjects were asked to decide whether acting in the relevant case (e.g. pushing the man off the footbridge in the footbridge case) was "appropriate" or not.[74] The moral dilemmas were subdivided into two categories: "personal" and "impersonal", the former being deemed structurally analogous to the footbridge, the latter to the switch dilemma. Greene's lurking hypothesis motivating the division (as mentioned in Section 1.1 above) is that it is the "up-close and personal" nature of footbridge-type dilemmas that explains the heightened emotional activation, which in turn, explains the psychological difference between the two dilemmas, given the presumed lack of such emotional activation in the switch case.

Sure enough, Greene and colleagues did find a differential activation in the medial prefrontal cortex (including the ventromedial prefrontal cortex or VMPFC), the posterior cingulate cortex (PCC), and the superior temporal sulcus (STS) in the case of "moral-personal" dilemmas—all areas previously associated with emotional processing (see e.g. Maddock, 1999).[75] In the "impersonal" (both moral and non-moral) dilemmas, in contrast, the dorsolateral prefrontal cortex (DLPFC) was differentially active—a region previously associated with controlled cognition, such as behavioural inhibition, working memory and planning (e.g. Milner & Cohen, 2001).

A subsequent study using a similar distinction between non-moral, moral impersonal and moral personal dilemmas (Greene et al. 2004) found that when subjects are confronted with personal moral dilemmas, there is increased activity in the anterior cingulate cortex (ACC) as well as the dorsolateral prefrontal cortex. While the former has been implicated in conflict detection involving a concurrent activation of more than one

[74] This probe is problematic for obvious reasons, but I won't press the issue here.

[75] It is common knowledge that different cognitive tasks activate various brain regions differentially. Due to the increased rate of action potentials of neurons in the relevant region, there is an increased local consumption of oxygen leading to decreased levels of oxygen in that region. This, in turn, sets off a haemodynamic reaction whereby local blood flow increases, culminating in a discharge of oxygen as a consequence of which the magnetic properties of the blood are altered (from oxygenated to deoxygenated). This is the change detected by BOLD fMRI (i.e. blood-oxygen-level dependent functional magnetic resonance imaging).

behavioural responses (*cf*. the Stroop task),[76] the latter has been associated *inter alia* with resolving such conflicts (MacDonald et al., 2000). Furthermore, consequentialist* responses were associated with a greater activity in the DLPFC. This, at the very least, is consistent with the dual-process theory, according to which the two systems compete in emotionally engaging situations, such as personal (as opposed to impersonal) moral dilemmas—hence the between condition variability in terms of the activation of ACC (conflict detection) and DLPFC (conflict resolution) in personal dilemmas, and the within condition variability in terms of DLPFC activation in case of consequentialist* moral judgment (conflict resolution in favour of the consequentialist* rationale).

Both of these studies are merely correlational, and so the question of whether emotional or controlled cognitive processes (and/or systems) are causally responsible for the judgments has yet to be addressed. Evidence that the emotion system may indeed be causally efficacious comes from a number of different studies. First, frontotemporal dementia (FTD) is a neurodegenerative disease due to an abnormal build-up of proteins affecting parts of the frontal and temporal lobes of the brain. Its symptoms include the "blunting" of emotions, and it affects social cognition and behaviour. Mendez et al. (2005) compared the moral judgment of FTD patients with that of two control groups (a group of Alzheimer's patients and a group of neurologically unaffected subjects) using Greene et al.'s battery of moral dilemmas. FTD patients (similarly to Alzheimer patients and control subjects) exhibited the usual response pattern in the impersonal condition, however, they proved significantly more likely than either of the control groups to endorse the consequentialist* judgment in personal moral dilemmas, such as the footbridge case.[77]

Second, VMPFC is thought to be necessary for the generation of emotions and, in particular, social emotions. Koenigs and colleagues confronted patients with VMPFC damage with the same set of dilemmas as used in the Greene et al. (2001) study (Koenings et al. 2007). VMPFC patients tend to display an abnormal pattern in terms of emotional reactivity as measured by skin conductance responses (SCRs) to highly emotionally charged stimuli, including pictures of mutilated bodies or social disasters. They also tend to

[76] The Stroop task involves having to name the colour of a written word. The word may be congruent (e.g. "yellow" written in yellow) or incongruent (e.g. "blue" written in yellow) with its colour. If it is the latter, then the task involves suppressing the automatic response (as reading is an automatised process—i.e. literate people "cannot help" reading a word when they see one), which is typically time consuming, hence the reaction time goes up (Stroop, 1935).

[77] "Significant" is a technical term here. It means that *if* there is no difference between these conditions (in this case in terms of consequentialist* vs. non-consequentialist* judgment), then observing such an outcome (i.e. such a distribution of judgments) as has been observed in the study has a probability that is below a predetermined threshold (often $p < .05$). This provides a decision procedure for keeping or rejecting the so-called null hypothesis according to which there is no difference (at least in null hypothesis testing—in more exact sciences, such as physics, the hypotheses tested tend to be more informative).

receive unusually low ratings on scales of guilt, embarrassment and empathy. Koenigs and his colleagues showed that, just like FTD patients, VMPFC patients tend to show a strong preference for consequentialist* judgments. A plausible interpretation of these results is that normal emotional processing is necessary for those judgments in which the non-consequentialist* rationale wins out—i.e. where, according to the dual-process theory, there is a heightened emotional response. Accordingly, the authors concluded that emotions play "a necessary role" in the generation of the typical pattern of judgments at least as far as "personal" moral dilemmas are concerned (*p*. 908), just as predicted by the dual-process theory.

Lastly, Greene et al. (2008) conducted an experiment in which they divided subjects into a control group and a "cognitive load" group. Participants in the cognitive load group were asked to carry out a task designed to engage their cognitive control system while making their judgments about "high-conflict" and "low-conflict" moral dilemmas (a modified set of dilemmas as compared to the previous experiments—see more on why this was necessary below).[78] What they observed is that on average, the reaction time (RT) of subjects in the cognitive load group in the case of high-conflict dilemmas was higher than that of subjects in the control group when they ended up choosing the consequentialist* option. Greene and colleagues maintain that this response pattern is predicted by the dual-process theory, since in cases in which the cognitive control system is engaged, generating the consequentialist* judgment may be met with difficulties due to the serial nature of the reasoning system. In contrast, when the non-consequentialist option is chosen, the cognitive load should not interfere with processing speed, since in that case, it is the emotion system that is responsible for the judgment—and that is exactly what they found.

2.2. Difficulties with the flagship examples

However convincing the collective force of these studies may initially appear, there are a number of problems, including outcomes not predicted as well as shortcomings in terms of experimental design that require closer examination. It is to these that I turn to in this subsection.

First, it becomes clear from a close examination of the reaction time (RT) data that under no cognitive load, the consequentialist* and non-consequentialist* judgments were equally fast: only under cognitive load

[78] "High-conflict" dilemmas included *Sophie's Choice*, for example: "It is wartime and you and your two children, ages eight and five, are living in a territory that has been occupied by the enemy. At the enemy's headquarters is a doctor who performs painful experiments on humans that inevitably lead to death. He intends to perform experiments on one of your children, but he will allow you to choose which of your children will be experimented upon. You have twenty-four hours to bring one of your children to his laboratory. If you refuse to bring one of your children to his laboratory he will find them both and experiment on both of them. Is it appropriate for you to bring one of your children to the laboratory in order to avoid having them both die?"

does RT increase selectively for consequentialist* judgments. This contrasts with what might have been expected on the basis of the theory, since even if there is no cognitive load, in high conflict scenarios when the emotional activation is presumably high, a consequentialist* judgment should prove more difficult (and therefore more time consuming) than a non-consequentialist* one (that of course should also depend on how strong the consequentialist* rationale is—but see more on this below).[79]

Second, there was also no effect of cognitive load on judgment, that is, whether the consequentialist* or the non-consequentialist* option was chosen was not a function of condition. In other words, whether a participant's attentional resources were engaged was inconsequential from the point of view of what option he or she chose. This is not predicted by the theory, either, which would presumably anticipate cognitive load to reduce the frequency with which subjects opt for the consequentialist* option, as the reasoning system is otherwise engaged.

Third, when Greene and his colleagues re-analysed their data by dividing their participants into "high-consequentialist*" and "low-consequentialist*" groups[80] based on the percentage of non-consequentialist* judgments made when confronted with high-conflict dilemmas, the high-consequentialist* group was on average actually *faster* to make consequentialist* judgments that non-consequentialist* ones (this was a noticeable though non-significant effect, $p > .06$—i.e. what is sometimes referred to as a "statistical trend"). The opposite (though slightly less noticeable) effect was observed for the low-consequentialist* group (also non-significant).[81] The first of these effects as well as the observation that generally, consequentialist* judgments tend *not* to be faster than non-consequentialist* ones (as might have been predicted by Greene's theory) suggest that under normal circumstances, thinking in a consequentialist* way can be as fast (or even faster) than doing otherwise. This observation, in turn, raises the possibility that what Greene refers to as "manual mode" thinking may in fact be as automatic as "automatic settings" themselves are, at least as far as

[79] Another potential problem is that the RT increase for consequentialist* judgment under cognitive load is merely three quarters of a second. Whether this is theoretically (rather than statistically) significant is not something we learn either from a detailed examination of the paper in question or from that of the dual-process theory in general. That is, the theory is taken (in the paper) to predict a mere "difference" between the conditions, and no specific (or approximate) prediction is derived from it as to what the size of the difference should be. I should mention that it would be somewhat unfair to criticise Greene too strongly for this shortcoming since the theoretical vagueness that comes hand in hand with null hypothesis testing is a general—though unfortunate—feature of psychological theorizing, see e.g. Dienes (2008).

[80] In the paper, the authors refer to these groups as "high-utilitarian" and "low-utilitarian", respectively, and the judgments as "utilitarian" vs. "non-utilitarian" (without asterisks). However, I opted for "consequentialist*" and "non-consequentialist* here to increase internal consistency (and to be able to ignore what motivational factors are responsible for these judgments).

[81] Neither in the paper, nor in their supplementary materials are exact data provided concerning this latter comparison apart from a graph, but based on the error bars, one can decide whether the effects are significant.

the dilemmas used by Greene et al. are concerned (especially when it comes to people who, judged by their responses, tend to favour the first kind of thinking).[82]

Consider now the increased RT under cognitive load in the case of consequentialist* judgments. Does this increase mean that the cognitive load interferes with consequentialist* thinking, as Greene et al. suggest? Not necessarily. It would be equally plausible to hypothesise that the interference is with conflict resolution, rather than with cost-benefit reasoning. That is, the RT increasing effect of cognitive load may be due to the lack of attentional resources necessary to resolve the conflict detected between the two competing outputs.

This latter hypothesis challenges some of the central assumptions of the dual-process theory, such as the characterisation of the systems. It also raises concerns over the question of how automatic responses can become automatic. Although in the previous section, I put the emphasis on innately determined processes, there is nothing in the model that specifically predicts that only these can be automatic. In fact, Greene frequently emphasises that social or trial-and-error learning may result in developing new automatic responses. The social learning of fear (see e.g. Gigerenzer, 2014) is a good example (which, incidentally, may be aided by the fear detection mechanism briefly outlined in the previous section). In any case, even if in principle, such observations are reconcilable with the dual-process theory, they do appear to blur the distinction between the two systems as well as raise the possibility that ways of thinking (including cost-benefit reasoning) may be automatic to varying degrees, rather than strictly either automatic or controlled.

Finally, browsing through the dilemmas in the supplementary material of Greene et al. (2001), one finds cases such as "the architect":

> You are a young architect visiting one of your construction sites with your boss. Your boss is a despicable individual who makes everyone around him miserable including you. It occurs to you that if you were to push him off of the building you are inspecting he would fall to his death and everyone would think it was an accident. Is it appropriate for you to push your boss off of the building in order to get him out of your life?

The problem is that, presumably, very few people would experience the affirmative rationale as anywhere near strong enough to create a conflict situation (for more criticism concerning the same point, see McGuire

[82] One explanation might be the automatisation of an otherwise controlled reasoning process. Something along the lines of this possibility was also raised by Pizarro and Bloom (2003, especially *p.* 195, *fn*. 2).

et al. 2009). These issues make it difficult to interpret Greene et al.'s data (2001) as either a test of or evidence in favour of the dual-process theory.[83]

3. The appraisal problem: Descriptive adequacy

3.1. The problem

There is a difficulty for Greene's theory that we have so far largely ignored. I said that the footbridge and the switch cases engage the emotion system differentially. However, Greene owes us an explanation as to why this is the case: why is it that being "close up and personal" with the large person on the footbridge triggers an emotional response that is lacking in throwing the switch? This may seem a trivial problem (shoving a person is obviously more emotionally engaging than pushing a lever), but as we shall see, it is far from it. The difficulty of characterising what gives rise to the engagement of the emotion system is an instance of a more general problem, namely, the appraisal problem.[84]

The appraisal problem concerns the question of how certain patterns in the input (i.e. behaviours) are recognised as being instances of certain more general categories. To illustrate the nature of this problem, let me return to Axelrod's Prisoner's Dilemma contest for a moment. I pointed out above that one of the convergences between evolutionary game theory and the dual-process theory was simplicity: the insight there was that difficult problems, such as the (iterated) Prisoner's Dilemma may have simple solutions in the form of a strategy as simple as Tit-for-Tat. As Robert Trivers puts it: "The simplicity of the tit-for-tat strategy bypassed, in one step, the cognitive complexity that was often assumed to be required to get reciprocity going in our own species" (Trivers, 2002, *p*. 54). I also mentioned the possibility that emotions might jointly implement a sort of enhanced Tit-for-Tat.

Nevertheless, there is one problem that has been largely (if not completely) overlooked in the literature on evolutionary game theory, where the focus is usually on strategies rather than the specific ways in which they are carried out—a problem that any organism playing Tit-for-Tat nevertheless has to solve: recognizing the acts of its interactants *qua* instances of cooperative and uncooperative behaviour in order to be able to react

[83] It is only fair to mention that in their 2008 paper, Greene and his collaborators did manage to pit consequentialist* vs. non-consequentialist* rationales more successfully against each other. Still, that study has its own problems, as discussed above.

[84] This problem is traditionally understood to be peculiar to the question of how situations elicit emotions (see e.g. Scherer, Schorr, & Johnstone, 2001), but it can easily be extended—as I will in this section—to instances in which a theory is expected to clarify the link between an eliciting stimulus and a mental state (such as an emotion or the tokening of a concept).

appropriately by employing the relevant strategy (namely, by cooperating or defecting).[85] This, in a nutshell, is the appraisal problem.

The appraisal problem wasn't really a problem for the computer programs in Axelrod's competition, since Axelrod ensured that they could interact with each other by translating their responses into a common language and format. Thus, there was no need for programs to solve the question of how to appraise the behaviour of other programs. On the other hand, for living organisms, especially humans, whose social interactions range from simplest (e.g. kicking someone) to highly complex (e.g. knocking someone out to stop them from committing suicide), the appraisal problem is of paramount importance.

Deciding whether an act is permissible in the context of the trolley dilemma constitutes a specific example of this problem, and understanding how it is solved is exactly the kind of task a theory in cognitive science, and *a fortiori*, a theory of moral cognition has to solve (see Chapter 1). As Mikhail puts it: "the critical issue in the theory of moral cognition is not whether moral intuitions are linked to emotions—they clearly are—but how to characterize the appraisal system that those intuitions presuppose" (2011a, *p.* 39).

It should be evident that the appraisal problem is closely related to the Argument for Moral Grammar, discussed extensively in the previous chapter. Both AMG and the appraisal problem emphasise the variegated nature of the actions and situations that the individual has to cope with in terms of some process of evaluation.[86] Of course, AMG goes further and offers a solution to a version of the appraisal problem (Chapter 1, Section 3.2). I mentioned above that Greene's theory offers another solution: the emotion system relies on certain cues the presence of which sets off an emotional reaction that tends to result in judging the eliciting action as impermissible. In other words, there is no need for a "grammar" and there is no need for abstract representations and principles. Greene's personal vs. impersonal distinction is essentially a proposed solution to the appraisal problem with respect to the trolley problem.

Unfortunately, there are some serious problems with this solution. To begin with, the conditions for a dilemma being classified as "personal" originally included that the agent's action would result in a) serious bodily harm, b) to a particular person or group in a way that c) the harm does not result from deflecting an

[85] An analogous problem, of course, is to align one's behaviour in such a way that it is deemed cooperative by the organism one is interacting with (or uncooperative—although there is a potential asymmetry here, as the individual advantages for displaying defection are different from the converse case for obvious reasons). That is, there are difficult computational questions on both the production and perception sides even of such a simple strategy as Tit-for-Tat (*cf.* Chapter 1, Section 3.4).

[86] There are some differences, of course. For instance, the appraisal problem is not specific to moral cognition. Further, AMG states the variability of action in terms of representation, while the appraisal problem does so in terms of external stimuli.

existing threat (see Greene et al. 2001, *p*. 2107). However, it was (admittedly) the researchers themselves who used these criteria to sort the dilemmas into groups of personal and impersonal, and the question of whether they actually correspond to the psychologically relevant criteria that are either separately necessary or jointly sufficient to evoke strong emotional responses is not addressed in those studies.[87]Furthermore, the personal vs. impersonal distinction generates some erroneous predictions. For instance, without qualifications, it predicts that any "personal" action engaging the emotion system would be judged impermissible, such as pushing someone out of the way of an oncoming trolley (*cf*. Mikhail, 2008b), which is clearly false.

More generally, the problem of what makes an action personal needs to be confronted for the dual-process theory to be able to *explain* why the switch and the footbridge cases are judged differently *if* it attempts to do so by reference to a personal vs. impersonal distinction that is supposed to be exemplified by those cases. What exactly is involved in an action's being personal? Does it involve touching? Does it have to be intentional?

3.2. A second response to the appraisal problem

3.2.1. The bifurcation of the "personal" dimension

In a series of experiments, Greene et al. (2009) attempted to address the above raised questions by separating the variables that could potentially be taken to be relevant with respect to Greene's "personal" dimension. In the first experiment (Experiment 1a) they created three brand new variants of the footbridge dilemma: (a) *remote footbridge* in which the agent has the option of throwing a "remote" switch, causing the man on the footbridge to be dropped on the track via a trap door; (b) *footbridge pole* in which the agent can push the man off the footbridge using a pole (rather than his bare hands); and (c) *footbridge switch*, which is identical to *remote footbridge* apart from the fact that the switch is on the footbridge near the person, rather than far away. These three variants in conjunction with the original footbridge dilemma separate the variables of *spatial proximity* (SP), *physical contact* (PC), and *personal force* (PF), the latter of which involves the agent impacting the patient ("patient" in terms of semantic role) using self-generated force (as opposed to an intermediate force, such as that of a gun). In short: *footbridge* (+SP, +PC, +PF), *remote footbridge* (–SP, –PC,

[87] As McGuire et al. put it: "there are no data demonstrating the necessity of all three criteria [(a)-(c) mentioned in the main text] in eliciting stronger emotional responses, nor do they correspond to well-established philosophical distinctions" (2009, *p*. 577). Although Greene acknowledges the difficulty of determining what exactly it is that renders an action personal as opposed to impersonal (Greene, 2008*, fn*. 2), he does not give up the idea that the distinction (perhaps once properly elaborated) may prove descriptively adequate and therefore solve the trolley dilemma.

–PF), *footbridge pole* (+SP, –PC, +PF), *footbridge switch* (+SP, –PC, –PF), so each possible pairing separates one variable.[88]

The permissibility ratings (phrased in terms "moral acceptability") were the following: *footbridge*: 31% (n=154), *remote footbridge*: 61% (n=82), *footbridge pole*: 31% (n=72), *footbridge switch*: 59% (n=160). The authors' analysis (ANOVA, planned pairwise comparisons)[89] revealed a significant effect of PF, but no effect of either PC or SP. In other words, the only relevant factor from the point of view of experimental subjects' judgments seems to have been personal force (whether or not the specific definition used by Greene and colleagues is on the right track). The results of the statistical analysis can be understood intuitively: the scenarios involving personal force (*footbridge* and *footbridge pole*) received systematically worse ratings than the scenarios lacking it (*remote footbridge* and *footbridge switch*). No other listed factor can be associated with any such trend.

However, in all four versions of the dilemma, the patient is harmed as a means to an end, namely that of saving the five. To put it differently, the agent's goal is only achieved if the patient is harmed. This contrasts with the *switch* case in which whether the patient gets harmed is inconsequential from the point of view of the success of the action (i.e. if the patient were to be able to roll out of the way of the trolley, the five would still be saved). And notice that the switch case also gets systematically better ratings than any of the versions of the footbridge dilemma (Hauser et al. 2006: 85%; Greene 2013: 87%[90]), indicating that personal force may not be sufficient to explain the variability (at least not all of the variability) in terms of the ratings across all the scenarios. Further tests (especially Experiment 2a and 2b, the latter of which was a reanalysis of Cushman et al. 2006) showed that indeed, the effect of personal force depends on intention, that is, whether the patient is used as a means to an end or more generally, whether the harm is intentional.

3.2.2. Modular myopia

The question at this point is (again, taking much of the dual-process theory for granted): why do we have an automatic emotional response to cases in which the agent intentionally harms the patient, especially to thereby achieve an end—even if the goal state is, to some extent, preferable to the outcome that is to be expected without the intervention (i.e. even if the benefits outweigh the costs)?

---

[88] There is no version of the dilemma in which PF is present, but neither of the other two variables. However, it is difficult to conceive of a case with (+PF –SP).

[89] ANOVA (Analysis of Variance) is a statistical procedure for identifying the factors that explain the variability within a given data-set. Planned comparisons, rather than being *post hoc*, are part of the experimental design.

[90] Greene (2013) mentions the 87% figure (*p.* 220*ff.*), however, having browsed through the original papers and their supplementary materials, I could not locate the study from which it was taken.

Greene's answer to this challenge is to propose a mechanism that he refers to as the "myopic module" (MM).[91] To understand how this module is supposed to work, one has to know a bit about Mikhail's theory of action representation (elaborated in Mikhail, 2011a; see also Levine et al. 2018), which he himself partially borrows from Goldman (1970) and Bratman (1987). To simplify grossly, the processing of an action plan is thought to involve representing it in terms of its causal and intentional structure. An action plan representation consists of a primary chain representing the succession of actions necessary for achieving the goal (or end) state in temporal order. Thus, the representation of throwing the switch involves the representation of the initiating movement, the movement of the switch, the movement of the track, the turning of the trolley, and the goal state: the five men saved. Secondary (or tertiary) branches may represent events that are expected to happen as a consequence of carrying out the action (such as the death of the man who happens to be on the side-track), but which are not causally necessary for the attainment of the goal.[92]

Greene's MM is posited as a mechanism for surveying or inspecting the *primary* chain of action representations, actively looking for instances of *harm*. When the MM finds one, it generates the quick automatic reaction that is responsible for the judgment that the action in the footbridge dilemma is morally impermissible. Hence, when the violent action is a means to an end, it becomes active, since only then is the harm part of the primary chain. In contrast, MM does not have the power to also analyse secondary (or tertiary) chains. Therefore, if the harm is a side-effect, MM is blind to it—or in other words myopic (and consequently, "manual mode" wins out in the absence of a salient emotional response). There are two reasons for this. First, the mechanism is held to work in (represented) temporal order, going through the steps of the action representation in a linear fashion. For MM to be able to inspect secondary chains, it would have to have a relatively sophisticated "queue" based memory system, which would allow it to return to the inspection of the secondary task once it is finished with the primary chain, which could be computationally costly (this, according to Greene, would be inconsistent with it being "automatic"). Second, representing an action in terms of goals and means to achieve them is a relatively trivial task, at least as compared to representing the side-effects of an action, which is much less so, mainly because there is no limit to the number and kind of potential side-effects of an action that could be considered by the mechanism (for more details, see Greene, 2013, *pp*. 224–240).

---

[91] On the face of it, this name seems a gem of a tautology. On Fodor's version of modularity theory (Fodor, 1983), modules are encapsulated, meaning that the flow of information into the module is severely restricted. In this sense, all modules are myopic. Nevertheless, Greene refers to a specific kind of myopia to be discussed in the text, which renders the terms slightly better motivated.

[92] Note that technically speaking, intentional and causal chains are understood as separate types of representations, but shall ignore that complication here.

3.3. Evaluation

The myopic module is an attempt at explaining the difference between the moral permissibility ratings of the dilemmas discussed above (as well as others). Whether it is successful is an interesting question that is beyond the scope of this chapter. Here, I am merely concerned with its relationship with Greene's dual-process theory as well as the DP framework more generally.

3.3.1. Greene's dual-process theory

Greene's initial claim that "automatic settings" rely on simple cues is already indirectly questioned by the MM hypothesis.[93] Recall the Whalen et al. (2004) study, which Greene uses as an illustrative example of how the emotion system works (i.e. it is supposed to rely on simple cues). The representation of an action in terms of a primary chain of necessary causal steps will on any reasonable theory count as richly conceptual. Also notice the way in which the dilemmas are usually presented: the subjects are asked to read the stories from a vignette, which results in a conceptual re-description of the events involved (à la Mikhail, 2008b). The "cues" necessary for the engagement of emotion systems will, by necessity, have to be in a format that is available to such systems, but that format is neither low-level nor perceptual, as Greene's theory would suggest.

In fact, the very idea of proposing a mechanism that is supposed to be looking for simple cues seems ill-conceived if the simple cue turns out to be something as elusive as a representation of *harm*, as in the case of MM. My point here is not that such a mechanism (i.e. one looking for instances of harm) is implausible. It is that harms (even "personal" harms) come in all sorts of forms—in fact, a potentially infinite variety of forms. This introduces a new microcosm of the productivity problem. Consequently, there is no reason to expect that a (simple) cue based recognition of harm may work, even in principle, because there is no reason to expect all forms of harm (or even HARM) to share a single cue.[94] Something has to give: either MM does not execute an automatic process as understood by Greene, or automatic processes do not have the property Greene attributes to them. This is no mere "semantic" quibble, either: it entails that the above proposed simple perceptual cue-based solution to the appraisal problem is unsuccessful. Therefore, the Argument for Moral Grammar remains unscathed.[95]

---

[93] To repeat the quote from *fn*. 66, Greene asserts that "all automatic settings rely on specific cues that are only imperfectly related to the things they're designed to detect" (2013, p. 227).

[94] Also note that a perceptual/behavioural definition of PERSONAL FORCE will be at least as hopeless as a perceptual/behavioural definition of FORCE.

[95] There is another interpretation of the dual-process theory that is independent of the simple cue assumption. This merely emphasises the idea that moral judgment is based on (at least) two separate systems (e.g. Cushman et al. 2010). This story is easily reconcilable with LA, however.

3.3.2. The DP framework

In my view, the problematic aspect of Greene's theory is inherited squarely from Greene's adoption of his version of the DP framework. In this framework, automatic processes are mostly innate (potentially learned) knee-jerk like heuristic responses to predictable patterns in the environment. This picture of automatic processes is clearly suggested by Greene's own examples, one of which—namely the fear detection mechanism discovered by Whalen et al.—I also rehearsed above. The DP framework thus understood licences a futile search for those simple cues the system responsible for moral judgment is supposed to exploit. No such search is likely to deliver successful theories. And if they do, it will be an accident (as in the case of MM—that is, assuming it is successful).

That the criterion according to which automatic processes are based on the exploitation of simple or low-level/perceptual cues is moribund becomes obvious by illustrating automatic processes by such cognitive capacities as face recognition or language comprehension. The process of recognising a face does not involve central access or conscious reasoning, such as the following: "judging by the long hair and certain characteristic features including the relative positions of the nose and eyes, this person is a female, so I can narrow my search by eliminating all the men I know as possible candidates, etc." Rather, when we recognise a face, we do so without any conscious effort, and without any insight into how we actually do it. *Mutatis mutandis* for understanding a sentence in one's first language. To illustrate the parallel nature of these processes, we can also understand a sentence in our first language and recognise a face at the same time—think of watching a dialogue in a film and recognising one of the actors or actresses involved. Yet no one would seriously argue that either face perception or language comprehension is based on some mechanisms exploiting simple cues.[96]

Of course, it would be premature to question the appropriateness of the distinction between two types of processes. We may choose instead to purge the characterisation of automatic processes of the "simple cue" assumption. So perhaps the problem is not with DP *per se*, but merely with Greene's take on it. Drawing this conclusion would be missing a by now rather obvious point: there is no clear sense in which the DP framework is a framework of anything in particular. For example, theories of language comprehension or face recognition are not articulated within DP, and nor is it easy to see how they would benefit at all from being so. To take language, for example, none of the fundamental hypotheses and assumptions concerning the nature of language competence guiding contemporary research in linguistics are derivable from DP: for such

[96] This arguably would be more plausible in the case of face recognition; for example, we might recognise a face by virtue of a characteristic nose. Tellingly, although such a strategy is generally available to us, it is only employed by those (so-called 'prosopagnosics') whose face recognition mechanism is deficient.

purposes, more substantive frameworks are required, such as Chomsky's generative programme. In contrast to such rich theoretical frameworks, DP has very little to contribute to the explanatory project of either linguistics or the study of other cognitive domains.[97] So, I submit, is the case with the study of moral cognition.

In conclusion, the dual-process theory offers a reasonably attractive surface explanation of why we judge certain cases of harm morally permissible, while slightly different ones as not permissible. It has the (somewhat dubious) advantage of being intuitively appealing, and it is also held to be supported by a considerable amount of (neuroscientific and psychological) evidence. However, to the extent that it makes falsifiable predictions at all, those predictions are sufficiently vague as to be difficult to evaluate (*cf*. Section 2.2 above). Moreover, Greene's apparent solution to the appraisal problem in terms of a simple cue based mechanism is deemed to failure, as in fact amply illustrated by his very own model. Not only is Greene's elaborated dual-process theory (positing MM) compatible with LA, it also reinforces its central argument, namely, the Argument for Moral Grammar.

In Section 4 below, I move on to considering more direct objections made against the Linguistic Analogy. I will tackle five such criticisms one by one and show that the thesis of first chapter—the case for LA as the general framework for the study of moral cognition—remains largely unaffected by them.

4. LA: Objections and replies

As I suggested in the previous chapter, the most typical objections against LA tend to concern its nativist aspects (e.g. Prinz 2008a. 2008b, 2009, 2014; Sripada, 2008; Sterelny, 2010), especially the moral version of the PoS argument (as best articulated in Mikhail, 2008a), that, together with the Argument for Moral Grammar jointly make the case for a Universal Moral Grammar or UMG (Mikhail, 2007). Such objections are not objections against all versions of LA, and certainly not against the version I endorse in Chapter 1 which understands LA as a research program or explanatory framework, but only against its "strong" version, according to which the success of LA is predicated upon either a very a close correspondence between the respective explananda (language and moral cognition) or the applicability of both of the two crucial

[97] Cognitive scientists and philosophers of cognitive science are well aware that much of our cognitive life (including our moral thinking and reasoning) is automatic, inaccessible, and unconscious (*cf*. e.g. Bargh & Chartrand, 1999). Making such assumptions with respect to a cognitive capacity requires no DP framework.

arguments of the Chomskyan generative framework to the domain of moral cognition. What I hope to be clear on the basis of the previous chapter, however, is that the success of LA as defended here depends on neither of those things.

Therefore, the task of the following section is not all that difficult. In what comes below, I address some of the most salient criticisms levelled against the Linguistic Analogy more or less in increasing order of importance.

4.1. Language vs. moral cognition

The claim that differences between the nature of the psychological underpinnings of language and moral cognition endanger the prospect of basing a theory of moral cognition on the Linguistic Analogy is one that has been made by most critics of LA (e.g. Dupoux & Jacob, 2007, 2008; Prinz, 2008a/b, 2009; Sterelny, 2010 etc.). For instance, Mallon points out (Malllon, 2008) that making the analogy with language invokes a host of properties of language that then will be expected to characterise moral cognition too, even though there is no conclusive evidence that this is indeed the case. For instance, language competence has an innate universal basis, is vulnerable to selective deficits, it exploits combinatorial representations, and operates based on unconscious rules and principles. It is not clear whether these are applicable in the case of moral cognition. It has also been pointed out for example by Jesse Prinz that language may not be the "best" analogy, due to some disanalogies between the language and moral cognition. Capacities other than language, such as vision and motor control, may share relevant features with moral cognition that might render such other analogies more appropriate in certain respects. Thus, one may propose a "vision analogy" or the "motor analogy", which might potentially prove equally as good as or even better than LA (Prinz, 2008a).

These are all interesting points, but besides being inconclusive even against the strongest version of the analogy, they fail to address the reasons why LA was endorsed in the previous chapter. For instance, two of the *potential* disanalogies hinted at by Mallon (i.e. universal innate basis and the existence of selective deficits) affect neither the usefulness of LA nor the applicability of the crucial distinctions and arguments advanced herein.[98] The other potential disanalogies (i.e. combinatoriality and consciousness) are treated elsewhere (Chapter 1, sections 3.2 and 5, and Section 4.3 in this chapter). As for the appropriateness of other analogies, partly for historical reasons perhaps, for better or worse, we have LA, and we have no motor analogy. That a motor analogy *might* be "better" is a weak and unsubstantiated claim. As argued before, ultimately, the success of LA is predicated on whether it provides a correct analysis of the explanandum (Chapter 1, Section

[98] These are addressed in Hauser et al. (2008b), for example.

3), the explanans (Chapter 1, Section 4), and the question of when the latter may be deemed adequate (Chapter 1, Section 6). I argued above that it does indeed do this. Furthermore, LA's success also depends on the extent to which the theoretical concepts and distinctions introduced in Chapter 1 contribute to the inquiry into moral cognition in a fruitful way. This is difficult to assess as things stand, yet I do think we have already encountered some positive signs in this respect in the foregoing discussions.

4.2. "Internal" vs. "external" principles

One principle that is often implicated in discussions of the Linguistic Analogy is the Principle of Double Effect (or PDE), which, *inter alia*, is often held to explain—or at least contribute to the explanation of—the psychological difference between the switch and footbridge scenarios in terms of moral permissibility, as well as numerous others. On the PDE, an act having a bad outcome that otherwise would be impermissible may be deemed permissible if the outcome of the act is not intended only foreseen, and the act is performed to achieve a greater good, which cannot be achieved otherwise.[99] Now, one point that critics of LA have made is that proponents of LA have assumed (or that LA suggests) that PDE is not only descriptively adequate but is also represented in I-morality (Mallon, 2008; Nichols, 2005). This assumption, they point out, is not warranted, nor is the idea that PDE is innate and part of a universal moral grammar (*cf*. the previous footnote).

Nichols goes further and proposes a model on which PDE is a consequence of two separate cognitive mechanisms that are unlikely to be parts of an I-morality—the existence of which both Nichols and Mallon are sceptical of in general. In Nichols's model, what is presupposed are the availability of non-hypothetical rules[100], such as the prohibition against murder, and a general reasoning capacity directed at figuring out how to minimise bad outcomes and maximise good ones. The two systems may be described as deontological and utilitarian, respectively.[101] According to Nichols, these capacities (i.e. non-hypothetical rule reasoning, and reasoning about achieving ends), jointly enable humans to reason in line with PDE, that is, they account for PDE's descriptive adequacy (which we will just assume for present purposes).[102] Furthermore, since neither

[99] As evident from this formulation, this principle is of considerable complexity. With the assumption of descriptive adequacy and the idea that it is untaught (since it is generally not even available to conscious reflection), nor is it obvious how we could "internalise" it based on experience, the PDE is often taken as one of the most potent example of the case for the strongest version of LA, as both the Argument for Moral Grammar and (given the assumed PoS situation) the argument for universal moral grammar may be applicable in its case.

[100] Non-hypothetical imperatives are imperatives the application of which does not depend on the individual's purposes or interests. This contrasts with hypothetical imperatives that do. Non-hypothetical rules are not necessarily moral, since rules of etiquette also belong to this category (*cf*. Foot, 1972).

[101] Note how this model is similar to Greene's model discussed further above.

[102] PDE's descriptive adequacy is not a settled issue. For example, Greene et al.'s findings discussed in Section 3.2.1 indicated that the means/side-effect distinction, which is at the core of the PDE, depends on the presence of personal

of the implicated systems is restricted to the domain of moral cognition, there is no need to posit a *moral* grammar or an I-morality.

In Nichols's rendition (based on Uniacke, 1998), PDE has four conditions. First, the *intended* action is permissible (that is, the intended consequences are the good ones). Second, the foreseen bad effect is not intended. Third, there is no way to achieve the good effect without also causing the bad effect. Fourth, the bad effect is not disproportionate to the good effect. Nichols's two systems account takes it to be the case that if the first two conditions are violated, the action will be judged to be impermissible, because the deontological system gets activated (e.g. in the trolley cases due to the prohibition against intentional homicide). Furthermore, if the latter two conditions are violated, the utilitarian system gets activated and generates a rationale against endorsing the action as permissible.

My point here is not to argue that Nichols's simple quasi-empiricist model doesn't work (although for related points, see Section 3 above or Mikhail, 2013, *pp*. 72-81), but to examine whether and if so to what extent the above criticisms cause problems for the framework presented in the previous chapter. To begin this examination, we must distinguish between two general points in the preceding paragraphs that are especially relevant in this context. First, as Nichols and Mallon emphasise, the PDE's descriptive adequacy (to the extent that we accept this thesis) is no evidence that PDE is represented *as such*. That is, reasoning in line with PDE is not equivalent to reasoning in terms of it (whether this reasoning is consciously accessible or not). Second, Nichols believes that the truthmaker for PDE's descriptive adequacy is not a moral faculty, but a hodgepodge of interacting psychological mechanisms (see also Cushman, 2016, for further possibilities along these lines).

A preliminary thing to notice is that neither of these points pose a serious challenge to either the general framework or the key arguments presented in the previous chapter, since they do not show (a) that the explanatory aims as set by LA are problematic, (b) that the Argument for Moral Grammar (as endorsed here) is mistaken, and (c) that pursuing LA is either likely to be unfruitful or there are better ways of doing moral psychology than pursuing LA. Still, they are clearly relevant and consequently, they do deserve some further discussion.

As per the inference from descriptive adequacy to psychological reality or strict adequacy, there are three observations to be made. First, although I do not wish to enter into gratuitous exegetical analysis, as a matter

---

force. However there are conflicting data in the literature even regarding this (*cf*. e.g. the difference in Mikhail's results regarding the "man in front" vs. the "loop track" scenarios, neither of which involves personal force). This is complicated by the fact that different studies often use different measures (e.g. modal vs. mean ratings) as well as different dependent variables (e.g. acceptability vs. permissibility). See also Cushman (2016), Greene (2013, *pp*. 217-224); Mikhail (2013); Zimmerman (2013).

of fact, rarely have proponents of LA drawn this conclusion as far as I am aware (an exception, mentioned by Nichols himself, is Harman, 2000, but his discussion is brief and admittedly rather speculative).[103] More generally, it is something of a consensus in (at least some areas of) cognitive science that explanations at the functional level do not translate straightforwardly to those at the computational-algorithmic or the implementational level—as pointed out by Mallon himself (*cf*. e.g. Putnam, 1975; Fodor, 1974). This is reflected by our discussion in the previous chapter (Section 3.6) of the empirical adequacies ("mere" descriptive adequacy is not assumed to translate to "strict" descriptive adequacy) as well as of the distinction between model and mental principles, the former of which may only be a working hypothesis for the latter.

As for the second point, that is, whether we need to posit no I-morality to explain PDE's descriptive adequacy, the issues are a bit more complex. One thing to notice is that the inference from PDE's descriptive adequacy to *something* at the psychological level of description explaining it appears to be entirely appropriate (if not deductively valid—as usual in this area of inquiry). Indeed, Nichols himself makes this latter inference, which is apparent from the kind of explanation he provides (this is true of other critics, such as Zimmerman, 2013, for instance, or anyone involved in this debate I can think of). That is, judgments that we *prima facie* would categorise as moral are expected to be explained by reference to psychological processes. This is the most essential part of what Chapter 1 (and hence LA) assumes. Of course, it may turn out that there is no I-morality (or specifically *moral* cognition) to speak of, and this would doubtless be detrimental to LA. However, Nichols's is far from being a strong case in favour of this possibility. First, it only treats one set of examples, which it claims to explain without reference to an I-morality.[104] But from this it does not follow that I-morality in general does not exist. Second, even worse, there is no reason (beyond perhaps some vague notion of simplicity) to prefer Nichols's proposal to other proposals that *do* make reference to I-morality (such as Mikhail's) even regarding this limited set of cases. On the other hand, there are quite a few reasons to disprefer it (see previous section or Mikhail, 2009, 2013). In any case, to come to general conclusions such as

[103] Here are two representative examples: "We can talk about the principles that characterize the computational system as being innate, if we like. However, it is crucial not thereby to imply are represented in any explicit fashion" (Dwyer, 2008, *pp*. 410-411); "we are not concerned here with how the PDE or whatever mental operations it implies are actually implemented in our psychology, nor with whether those operations are modular in Fodor's (1983) sense, or otherwise informationally encapsulated" (Mikhail, 2011, *pp*. 148-9).

[104] Although even this is far from obvious: the prohibition against killing—part of Nichols's story—appears a moral faculty principle *par excellence*. Nichols might retort that this is not in fact a principle specific to moral cognition, as there are other non-hypothetical imperatives that are not uniquely in the domain of moral cognition (see *fn*. 100). We could respond to this by pointing out that although there might be shared mechanisms or representations in various different types of non-hypothetical rules or principles, that does not entail that there is not a special class of them that belong to the moral domain and whose moral-deontic status is generated by I-morality.

the one Nichols and Mallon prematurely makes in this regard, the kind of inquiry endorsed in the first chapter seems best suited (see arguments offered therein).

Finally, notice that PDE is a conflict principle (in the sense of Chapter 1, Section 5), and we said little concerning how such principles relate to I-morality. The minimal job we assigned to the latter (in Chapter 1, Section 2) is the analysis of actions in terms of deontic concepts. Thus, the input to this process is supposed to be (some sort of) action descriptions, while the output is a deontic value "attached" to those representations. With this in mind, it will be important what exactly the kind of judgment is that is provided by conflict principles. It is immediately clear, to begin with, that faculty principles must be part of FM on the current framework (this premise may be empirically mistaken, of course, as discussed above and in Section 4.5 below). If conflict principles produce judgments of acceptability, rather than moral impermissibility, they are not properly characterised as parts of I-morality so conceived. If, on the other hand, they supply moral-deontic judgments (such as *morally impermissible* or *obligatory*, as argued for by Mikhail), then, again, they will be part of the domain of moral cognition on this narrow conception.[105] Hence, the disagreement is not about the necessity of positing principles to explain moral judgments but (a) what the nature of the content of particular purported principles is, (b) whether they are innate (in some sense of the term), and (c) whether they are part of *moral* competence or performance and, relatedly, (d) whether they are operative or merely express. Whatever the outcome of such debates, it will not substantially detract from LA as defended here.[106]

4.3. Is the Argument for Moral Grammar any good?

The observation—discussed in Chapter 1, Section 3.2.2—that the Argument for Moral Grammar is not perfectly isomorphic to its linguistic counterpart is another example of the series of differences between language and moral cognition, which might be regarded as considerations in favour of dismissing LA (see Section 4.1). However, unlike in the previous cases, the particular ways in which the analogy is imperfect may have a crucial bearing on the success of LA. After all, the Argument for Moral Grammar is one of the central pillars of the Linguistic Analogy (as argued in Chapter 1, Section 3.2). I already discussed how I believe this

[105] Notice that it is not obvious that in cases for example where there is a violation of the proportionality condition of PDE (condition 4), why the judgment should be one about moral permissibility, as Nichols appears to assume. If it is the general reasoning system that vetoes the endorsement of the action, then the action is dispreferred in terms of utility, rather than in terms of moral permissibility.

[106] It is true that since Mikhail uses PDE as his flagship example of a universal and potentially innate operative moral principle, if these properties were not to characterise PDE, our subjective assessment of LA could suffer accordingly. Yet strictly speaking, LA *per se* is silent on whether PDE should have these properties.

argument differs from the Argument for Mental (Linguistic) Grammar, and how in spite of the disparity, the part of the conclusion of the argument that is significant for us remains intact.

To reiterate, although at times, Mikhail seems to argue for the idea that the productivity problem cannot be solved without positing a moral *grammar* (in the sense of a combinatorial system), the core of the argument is that representations and processing principles (potentially of all three types) are required to explain how the productivity of action representation does not pose a problem for moral cognition. As discussed before, the difference between language and moral cognition is that in the case of the former, the principles must be (recursive) combinatorial principles, that is, principles that explain how larger linguistic structures (such as phrases) can be constructed out of smaller structures, often of the same kind. In contrast, a theory of moral cognition need *not* posit a combinatorial system of its own, since the combinatorial system responsible for productivity is most plausibly not specific to moral cognition.[107] Nevertheless, importantly, the fact that moral thinking appears designed to meet the productivity challenge posed by the combinatorial nature of the representation of action needs to be accounted for. As far as anyone knows, positing principles (constraint, processing and output) in the domain of moral cognition is the best (only?) way of doing so (see also Section 4.4 below).

Given our rather extensive discussion of this issue, the current analysis of what I take to be the best critique of the Argument for Moral Grammar will be brief and opportunistic. Dupoux and Jacob (2007) point out four reasons why an explanation of moral competence need not presuppose a moral *grammar*, which correspond to four properties that generative linguistic grammars have but moral cognition does not (or does not *obviously*) exhibit, out of which I shall consider three. The starting point is the observation that generative grammars specify the syntactic mapping between phonological and conceptual structure. It is due to this mapping that we can construct a phonological representation on the basis of a thought (production) and *vice versa* (perception). Thus, the first dissimilarity is already apparent: in language, the mapping between the phonetic and the conceptual levels of representation is reversible: if I hear the phrase "Luca ate the brown

[107] Prinz also argues along similar lines: "Are moral rules combinatorial? This is a bit more complicated. As Hauser et al. point out, we certainly need a combinatorial system for categorizing actions. But notice that action categorization is something we do quite independently of morality. Our capacity to tell whether something was done intentionally, for example, operates in nonmoral contexts, and individuals who lack moral sensitivity (such as psychopaths) are not impaired in recognizing actions or attributing intentions. Psychopaths can recognize that someone is intentionally causing pain to another person. Moral rules take these combinatorial, nonmoral representations of actions as inputs and then assign moral significance to them. The distinctively moral contribution to a rule such as that killing is wrong is not the representation of the action (killing), but the attitude of wrongness. It's an interesting question whether moral concepts such as "wrong" have a combinatorial structure; they may. However, by focusing on the combinatorial structure of action representations, Hauser et al. fail to show that representations specific to the moral domain are combinatorial" (Prinz, 2008, *p.* 160).

cow", having constructed the phonological representation based on the sound signal, I can entertain the *thought* that LUCA ATE THE BROWN COW by virtue of the syntactic mapping. The same thought can also occur to me first (for example by remembering seeing Luca eat the brown cow), which is then mapped onto the phonological representational level by virtue of which I can utter "Luca ate the brown cow". This does not seem to be the case with respect to moral cognition: I can evaluate Luca's eating the brown cow as *wrong* (let's assume I'm a moral vegetarian). Parallel to the case in language, this can be seen as establishing a mapping relation between an action description (Luca's eating the brown cow) on the one hand and a moral concept on the other (or a "valence" as Jacob and Dupoux put it). Yet given the output alone, I cannot construct the thought that Luca ate the brown cow.[108]

The second and third points are closely related: encapsulation and domain specificity. As for the former, Dupoux and Jacob point out that while in the case of language, the mechanisms responsible for the implementation of the syntactic mapping function are encapsulated in the sense that what is informationally available to them is strictly limited relative to central cognition and thus they are not even potentially affected by the knowledge and beliefs of the organism. In contrast, moral judgment is potentially sensitive to all sorts of information. For example, the judgment that fracking is morally wrong (ideally) rests on an understanding of what fracking is (fracturing a rock by pressurised liquid to extract gas and oil) and why it is environmentally harmful (e.g. it involves the transportation of huge amounts of liquid that is environmentally taxing). Thus, there is no way of specifying in advance just what kinds of information moral cognition/judgment has access to.

Regarding domain specificity, Dupoux and Jacob observe that while syntactic rules (such as rewrite rules) are specific to language (what use would they have in other domains?), it is unclear whether moral principles are specific to moral cognition. For instance, the rewrite rule we considered above (NP → NP CP) applies in the domain of syntax alone. By contrast, the PDE may be applied to domains other than those of moral cognition. For example, it makes sense to assume that there is a *prima facie* prudential rule against making oneself look stupid. We can also draw a distinction between making oneself look stupid *in order to* do something, for example to get famous, and foreseeing that on the road to becoming famous one has to put oneself through situations in which one looks stupid as a byproduct of this process. We may reason in line with the prudential version of PDE that only the first option is prudentially disfavoured ("prudentially

[108] Unless of course the output is something like "Luca's eating of the brown cow is/was wrong", from which I can trivially construct "Luca ate the brown cow". The point remains that the action description in this latter case will not be formed on the basis of the mapping. As Jacob and Dupoux put it, "morality is an evaluative system, not a generative one" (*p*. 376). (*Cf*. Chapter 1, *fn*. 17.)

impermissible" would sound odd). None of this plausibly involves *moral* evaluation (Nichols also makes a similar point, see 2005, *p*. 361, *fn*. 8). Relatedly, PDE applies to representations such as intentions, side-effects, and so on, many of which seem also not to be specific to moral cognition.

The first thing to notice is that none of Dupoux and Jacob's criticisms appears to render the inference to moral principles endorsed in Chapter 1, Section 3.2.2 problematic. Nevertheless, the points raised are instructive. In what comes below, I shall concentrate on the latter two points of contention.

The version of AMG endorsed here has it as its bottom line that moral cognition is underpinned by principles. Thus, one task for a theory of moral cognition is to posit (model) principles that are descriptively adequate. To take a toy example, consider utilitarianism as a descriptive theory—which of course it is not (the example is also used by Roedder & Harman, n.d.). According to the utilitarian principle, it is obligatory (or at least permissible) to perform an action if it increases overall utility or happiness and not so otherwise. This principle can be seen as an attempt to solve the productivity problem by proposing a calculus: a simple overarching principle that is defined over representations of (possible) outcomes *in terms of overall utility*.[109] The difficulty for utilitarianism as a descriptive theory of I-morality is how to turn representations of events into representations of (potential) overall utility. That is, the interesting theoretical work would be the specification of conversion rules. Once this step is completed, given a set of representations of possible outcomes in terms of overall utility, the selection procedure, which is mandated by the utilitarian faculty principle, would generate straightforward predictions as to which of the available action plans would be regarded as morally obligatory or impermissible to execute.[110]

This toy example indicates a way in which issues of encapsulation and domain specificity may be raised with respect to moral cognition. Concerning encapsulation, we can see how the informational access of the moral principle is limited to a closed set of representational variables, and the operation of the mechanism(s) establishing the input-output relation need be only indirectly affected by the organism's knowledge and beliefs. In regard to domain specificity, that the principles and representations are proprietary to the moral domain is not a requirement. Increase and utility are in no way specific to issues relating to moral cognition of course. Nevertheless, the important point is that the representations over which moral principles are defined need not be promiscuously variable to explain the fact that the kinds of information that are potentially taken into account in moral judgment (broadly defined) may be. On descriptive utilitarianism,

[109] For now, let us now ignore the obvious fact that the principle fails to satisfy descriptive adequacy.

[110] Note in passing that utilitarianism (understood in these terms) eschews the need for conflict principles, which is another way of saying that, at least in its canonical formulation, it doesn't recognise *prima facie* wrongs.

moral cognition can make a judgment regarding the permissibility of fracking once it is processed in a format that represents its perceived overall utility. This format does not need to specify what fracking *is*. Moral cognition only cares about whether it increases overall utility or not. In other words, knowledge about fracking is not input to evaluation by a faculty of moral cognition, only to the analysis that defines how it is conceived of in terms of utility.[111] This might be a difficult point to understand, so let me elaborate on it a little bit further.

One observation about moral cognition is that it involves mandatory processing. That is, when we are exposed to an action (in the form of reading a vignette or perceiving a scene), we automatically evaluate it in moral terms. If I see an innocent person being slapped in the face for no reason, I cannot help but morally judge the action and the agent who performs it. This is similar to language processing and the construction of a 3D visual representation of the world on the basis of 2D input from the retina (Mikhail, 2008b). To sick with language, mandatory processing (one of the many hallmarks of modularity) takes place once the phonetic representation of the sound signal is constructed. If the input is analysed as noise, there is no linguistic processing taking place. The idea is that there is—to use Jackendoff's term—such a representational "plane" in the case of moral cognition also—which is taken to be the very abstract "utility level" by descriptive utilitarianism (or a similarly abstract level by descriptive Kantianism, for example). Although we do not know what this representational plane is, the fact of mandatory processing in this domain appears to imply (or at least suggest) that there is one; otherwise, how would we be able to automatically respond to novel situations rapidly with a moral judgment? An informed (and quite traditional) guess is that it will have to do with the representation of action.

Now, there are many ways in (or representational levels at) which an action can be cognised. For example, drilling a hole in someone's head may be construed as a mechanistic process (this is based on a representation at Jackendoff's "physical plane"), but it can be cognised is terms that we would readily perceive as morally salient: as *X intentionally harming Y* (Jackendoff's "social plane"), as per social domain theory (mentioned in Section 5 of Chapter 1). The idea is that these representations are constructed simultaneously in the same way as the phonetic and syntactic representations of "my car" and "NP" are simultaneously constructed and processed by different mental mechanisms (those dedicated to phonology and syntax) when one utters "my car's been stolen". A difference may be that while we have no conscious access to syntactic representations, for all we know, we do have at least some degree of conscious access to the representational plane that moral

[111] Relatedly, that we make the means vs. side-effects distinction in situations that we do not represent in moral terms is not too illuminating in itself. After all, we don't make *moral* judgments about cases which we don't represent as involving *prima facie* moral wrongs (such as the example in the text).

cognition exploits. To be clear, the idea is not that HARM is an inherently moral concept (indeed it is not, see e.g. Sousa et al. 2014), but that there is a level of mental description of actions (which might be arbitrarily abstract) at which intuitive moral principles operate.

This way of looking at what a theory of moral cognition has to achieve is closely related to Ray Jackendoff's so called *parallel architecture* (Jackendoff, 2002, 2003). The idea in a nutshell is that the traditional components of linguistic analysis (phonology, syntax, semantics) have their proprietary representational levels and principles operating in parallel. For example, syntax only "sees" syntactic phrases (e.g. "NP") and other syntactic features ("Determiner"), and its operations are defined over these representations. (As we saw in Section 3.2.2 of the previous chapter, to avoid the multiplication of rules to the point that they become completely theoretically useless, the principles of syntax need to be defined over representations that abstract away from fully specified lexical items.) This is an efficient arrangement as all the components of I-language can do their mandatory and automatic operations in tandem, with little or no regard as to what is going on at other representational levels (see Figure 2.2 below).

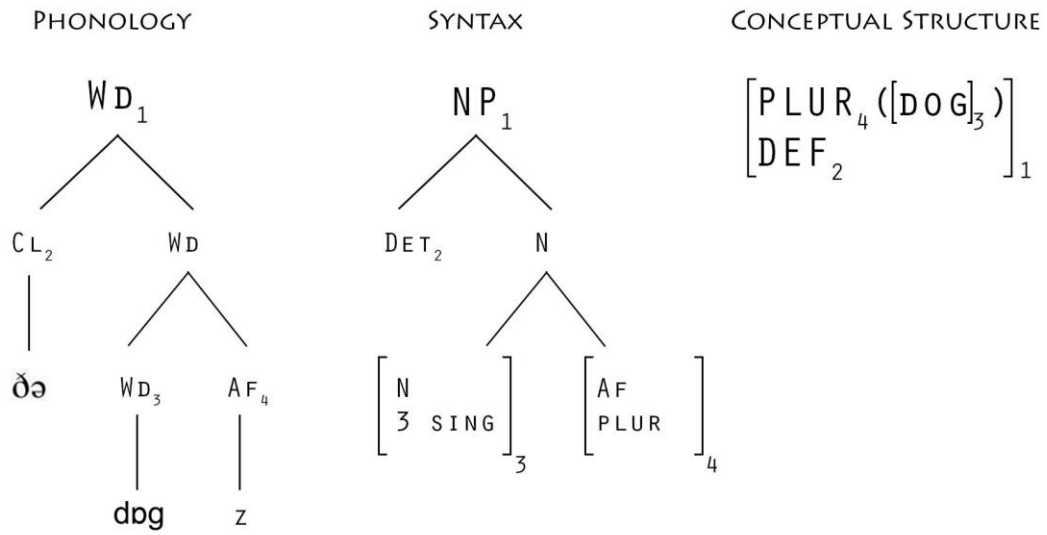

Figure 2.2: The three levels of linguistic processing—phonology, syntax and semantics—involved in the representation of the phrase "the dogs". The indices show how the representations are bound together. For instance, the plural affix ('Af'), which is represented at the phonological level as the phoneme */-z/* is bound to the semantic category *plural* ('PLUR'), which is why we understand the phrase as referring to two or more dogs, rather than only one. Jackendoff calls attention to the fact that the bottom nodes in the syntactic tree are syntactic features rather than lexical items (as in the customary textbook

notation). This is because syntax is blind to the former (*cf.* encapsulation): to create a syntactically well-formed sentence, no reference to the particular lexical items need be made (which is why colourless green ideas may sleep furiously). (The figure is based on Figure 4 of Jackendoff, 2003, *p.* 659.)

To return to moral cognition, the suggestion is that theories of moral cognition may be understood as engaged in the task of discovering the appropriate representational plane for moral cognition and identifying the principles that operate on the representations located at this level of mental structure. It seems obvious to me, for example, that representing the head-drilling as merely a physical process does not result in a moral judgment at all (although this may be difficult to concede, which is likely to do with the fact that representations relevant to moral cognition are automatically constructed). Similarly, a tsunami, however deadly, does not seem to invoke the kinds of judgments we are interested in—unless it is seen in a different light, that is, represented at a different plane; for example, the tsunami may be seen as God's punishment, involving an agent (God) and patients (the sufferers) with mental states and an internal source of energy. There is a rather widespread agreement that the domain of moral judgment is the action of intentional agents (often performed on intentional patients), but there is less agreement as to what the principles are, and what exact representations they operate on, which is part of the reason why the Linguistic Analogy may be particularly useful. Sadly, at this point, we have to leave this difficult problem behind. However, some of the issues raised in subsequent chapters will be related to the concerns raised in this section.

4.4. Do we need principles at all?

As we have repeatedly seen, the Argument for Moral Grammar may be successful even if *what* it is an argument for is not appropriately thought of as a moral *grammar*—in the sense of a system of combinatorial principles or rules, such as natural language syntax. This is because the idea that a potential infinity of actions can be evaluated through a faculty of moral judgment licenses the inference (for want of a better one) that moral judgment is underlain by a set of principles defined over some representations that are (by necessity) more abstract than the fully fledged description of any particular action that can nevertheless be evaluated morally. However, irrespective of the merits of this argument, its conclusion has been questioned on different grounds: some theorists believe that the very idea that moral principles describe something that has psychological reality is mistaken, even if "having psychological reality" is understood in the loose sense of a real reasoning pattern that emerges as a function of the operation of a diverse set of psychological mechanisms,

as suggested by Nichols or Mallon (Section 4.2). A prominent proponent of this line of criticism has been the cognitive scientist, Nicolas Baumard.[112]

Baumard argues against the psychological reality of principles on account of some problems proponents of specific principles (either in descriptive or prescriptive guises) have had to face. These are the following. First, principles always seem subject to exceptions. For example, killing is judged to be forbidden, but killing in self-defence can be seen as permissible (ironically, the PDE was conceived of by Aquinas as an attempt to specify exactly when this is the case). Second, principles apparently guiding moral judgment seem to exhibit a degree of context sensitivity: they seem descriptively adequate in certain cases but not in others. Worryingly, the theories that propose them are rarely able to account for this variability in terms of applicability. It seems that theories merely redescribe situations in more abstract terms and attempt (with fluctuating success) to apply them in other cases. Thus, Baumard argues, principles "might be "folk theories" that people use to describe a situation and argue a position" (Baumard, 2016, *p*. 89).[113] For example, the ownership of a specific object (such as a canoe) and a situation (no one can cross the river without it) might generate new express principles (such as "the owner of the canoe has a duty to help people who want to cross the river"). Tellingly, Baumard notes that "in reality, rather than an infinity of principles, what we have is a single mutualistic logic that applies in an infinite variety of situations" (ibid.).

On the back of such considerations, Baumard criticises theories such as Mikhail's on account of the latter's subscription to—and dependence on—the existence of principles. More constructively, he advances alternative explanations of the results in connection with the trolley scenarios the explanation of which is often proposed in terms of principles such as the Action Principle and the Principle of Double Effect (see above). In line with his mutualistic approach to moral cognition (see also Baumard et al., 2013), he analyses the trolley cases as distributional situations in which a good (survival) is represented such that it must be dispensed as justly as possible for actions to be judged morally praiseworhty, given the relative position of the dramatis personæ.[114] More generally, on Baumard's theory, the judged moral (or deontic) status of an action

[112] Other eminent critiques of principles are moral particularists, such as Jonathan Dancy, who deny the existence of moral principles altogether and argue that moral judgments are made on a case-by-case basis "without the comforting support or awkward demands of moral principles" (Dancy, 1983, *p*. 530). For a critique of Dancy's position, see Mikhail (2011, *pp*. 71-73). See also Williams (1985) for a related argument—which is also a criticism of earlier versions of LA—, and Mikhail (2017) for a reply.

[113] Zimmerman (2013) advances a closely related argument, and Greene's line regarding PDE is also on this track: "people all around the world make judgments that are (imperfectly) consistent with the Doctrine of Double Effect while having no knowledge of the doctrine. This tells us that intuitive judgments come first, and that the doctrine is just an (imperfect) organizing summary of those intuitive judgments." (Greene, 2013, *p*. 223).

[114] As noted by Baumard himself, this analysis may be somewhat more plausible in alternative variants of the trolley problem in which, for example, a buoy has to be distributed between a group of five drowning people and a lone

is a function of an analysis of how the relevant action is taken to affect individuals in terms of their system of values (done by what he refers to as "intutive axiology"). Any action that has the consequence of affecting others' interests is subject to moral evaluation. The particular evaluation will depend on a conception of fairness, thus, it depends on assumptions about merit, costs of the action, magnitude of benefit to parties affected, goods to be distributed, and so on.

To illustrate the explanatory potential of his theory, Baumard proposes a rather intriguing analysis of why "moral leaders"—that is, people who influence and engender what is subsequently seen as moral progress—are special:

> "moral leaders direct other people's attention to phenomena that their own less-developed intuitions had previously been insensitive to. They do not propose new moral principles. They activate our moral sense, showing us that certain people suffer more than we thought, or that helping them is easier than we thought. By changing how we conceive a situation, they show us that we have duties that we did not think we had. If their arguments are convincing and their vision wins out, their way of looking at the situation becomes so natural that their position comes to seem completely legitimate. The difference between the judgments of moral leaders and those of other people essentially results from the fact that, in the light of original thoughts or extra information, they envision costs and benefits in a different, innovative way" (Baumard, 2016, p. 112).

This compelling analysis notwithstanding, there are two serious problems with Baumard's criticism of the idea that principles are of central importance for a theory of moral cognition. First, he fails to distinguish between the many different potential senses of the term *principle*, such as those identified in Section 5 of the previous chapter, which results in him dismissing the relevance of all kinds of principles despite the fact that most of his critique only makes sense if seen as an attack against the proposal that specific types of principles (such as express faculty principles) be understood as the fundamental explanatory tools in the study of moral cognition. Second, in his own theory, Baumard himself appeals to a principle of a very general kind, namely, what we referred to earlier as a *derivational principle*. Therefore, Baumard's attack on the conclusion of the Argument for Moral Grammar (as defended in Chapter 1, Section 3.2.2) is unsuccessful. I will consider these two issues in a bit more detail below.

First, consider Baumard's criticism. He notes that (i) principles are subject to exceptions and (ii) they are context sensitive and (iii) they are potentially infinite therefore useless as an explanatory device. Let us take

---

drowning individual (see Thomson, 1985). However, my point here won't be that the analysis is not plausible, so we might as well ignore this complication.

each of these points in turn. As for the first point, the claim that principles are subject to exceptions makes it sound like Baumard has processing principles in mind. Indeed, as alluded to above, conflict principles (such as PDE) have the function of arbitrating between the outputs provided by different faculty principles in case there is tension between what they require in a particular case. This means that (faculty) principles ought not to be expected to provide the final word on what moral judgment is made. Thus, this point, to me, appears a descriptive statement rather than criticism. I would also point out that the notion that processing principles are never operative would be a radical one, questioning much of the literature on normative judgment (e.g. Nichols, 2004; Sripada & Stich, 2007; Morris & Cushman, 2018; Schmidt & Rakoczy 2018, etc.), that speculate over how "norms" (i.e. rules or faculty principles) function and are acquired, for example.

Second, that (some!) principles are context sensitive, again, does not seem to constitute serious criticism. In the case of context sensitive principles, one explanation might be that they take as input some contextual variable that limits their application. Finally, in line with what was argued in Section 3.2.2 of Chapter 1, we may indeed say that (faculty) principles are infinite, but only in the limited sense that we can attach a deontic value to a potentially infinite number of actions. Then, we may enshrine and expressly preserve each of these action descriptions in terms of "principles" arbitrating over them. This kind of reasoning may indeed produce arbitrarily numerous *express* principles (such as "you mustn't steal my slippers in the morning when it's cold and I'm trying to prepare my breakfast in the kitchen on the tiled kitchen floor"). But express principles are not properly regarded as the fundamental explanatory tool in moral psychology, and few argue that they should be (indeed, they are typically regarded as explananda, not explanans). Thus, none of Baumard's considerations provide the slightest reason to doubt the conclusion reached and endorsed in the previous chapter.

In fact, some of Baumard's reasoning supports it. Consider what he takes to be the fundamental components of moral judgments: representations of costs, benefits and interests of the parties involved, and some notion of fairness that defines why a given action is judged to be moral or not, given such an analysis. Clearly, Baumard is proposing a principle so general as to define the entire domain of moral cognition. I do not aim to take issue with this theory (as a matter of fact, I happen to find it bold and fascinating). Rather, the point here is that Baumard implicitly subscribes to the kind of moral psychology that takes the notion of principles as its fundamental explanatory tool, and for some of the same reasons that Mikhail does, for example. To be specific, he recognises the potential infinity of situations about which we make moral judgments, and realises that this calls for a general explanation (see e.g. the quote above), and the fact that he refers to his proposed mutualistic principle as "logic" makes little difference in this regard. A rose by any other name would smell as sweet. Baumard also recognises that his logic needs to be defined over abstract

representational variables (such as costs, values, benefits). He also correctly notes (elsewhere in his book) that these representations are provided by systems outside of moral cognition (such as the "intuitive axiology" outputting representations of value).

As noted above, Baumard's principle is a derivational principle *par excellence*: it attaches representations of moral-deontic value to representations of actions as a function of the way they are represented at an abstract representational level. For example, prototypical killing may be represented as unfairly depriving an individual from a basic value, namely, his or her own life.[115] So, by Mikhail's standards (and mine), Baumard's theory satisfies the three basic formal requirements a theory of "moral grammar" is expected to meet: it provides a set of structural descriptions, it provides deontic principles (albeit fundamentally only one), and it suggests (if not provides in explicit form) the conversion rules (e.g. the intuitive axiology) that take as input (e.g.) verbal or visual stimuli and has as output the structural descriptions over which his moral principle (or "logic") is defined. True, Baumard's account is clearly a rival of Mikhail's, but it is a rival that shares (if only implicitly) the latter's broad explanatory framework explicated and endorsed in this thesis.[116] Consequently, it would be odd to consider it as a criticism of the Linguistic Analogy—which, paradoxically, Baumard himself does. On the contrary, in spite of his protestations to the contrary, if successful, Baumard's theory would be testament to—rather than an indictment of—the value of LA as a framework for the study of moral cognition.

4.5. Is moral judgment a natural kind?

Since we already addressed this question indirectly in Section 4.2 above, I shall keep the discussion short here. To reiterate, the worry is that there might be no distinctive domain of moral cognition: there might be nothing at the level of psychological description corresponding to the term *moral*, as in "moral judgment" or "moral obligation". We provisionally said in the previous chapter that "moral" in "moral obligation" signifies the fact that the obligation is intuitively judged to be of the moral type. We also treated moral judgment (in the narrow sense) as judgment about the moral-deontic status of actions. Now the current issue is raised by the possibility that at the psychological level, such judgments are not unified, or more appropriately, they fail

[115] Thus, we can provisionally accept the "norm" that killing is wrong. In fact, Baumard's principle thus could be seen as a constraint principle of a general kind.

[116] The question of which theory fares better in terms of descriptive adequacy is an empirical issue that I do not aim to address in this chapter or indeed elsewhere in the thesis.

to constitute a natural psychological kind.[117] This would clearly be detrimental for LA, as indeed for any theory of *moral* cognition in general.

An example of a version of this line of reasoning is advanced by Sinnott-Armstrong and Wheatley (2014), who argue that none of the proposed ways of unifying moral judgments as moral are successful (see also Stich, 2006). These include the form, content, force, phenomenology and neural realisation of what are treated generally in the literature as instances of moral judgment, including judgments concerning fairness, authority, purity, intentional harm, and so on. Sinnott-Armstrong and Wheatley's criticism against psychological theories assuming unification (including by the proponents of LA) is that, although such theories start from a closed set of cases (usually, cases involving some kind of harm, as in the case of Hauser et al., 2006; Mikhail, 2011; or Turiel, 2002), they want to generalise to all and only moral judgments, where where this class (moral judgments) is supposed to capture to some extent what we would intuitively call "moral". But, as they argue, it is not plausible to make such inferences.

At times in their article, Sinnott-Armstrong and Wheatley seem to mistake what the proper explanandum of a theory of moral cognition is. For example, they have the following to say on this topic: "We are interested not in whether some individual could have a unified concept of moral judgment, but rather in whether there is any unified concept of moral judgment that is shared by the many people who engage in debates that are supposed to be about a shared topic in morality" (Sinnott-Armstrong & Wheatley, 2014, *p*. 453). Now, whether there is a shared *concept* of moral judgment is entirely beside the point. If there is such a concept, it is likely to bias the inquiry in obvious ways. But the purpose of a theory of moral cognition is to explain how moral cognition actually works, not to explain what we think about it.

To be fair to Sinnott-Armstrong and Wheatley, they do not consistently make this error, and their point may be valid if understood as applying to the relevant phenomenon, namely, moral judgment itself, rather than our conception of it (as I believe they probably intend it). And of course, as suggested above, our conception of moral judgment is a guide for identifying our explananda. However, pursuing research in any area of inquiry is expected to shape and modify our initial conceptions about the relevant domain, which is exactly why we engage in the study in the first place. For example, if we consider the trajectory of chemistry from Empedocles to Mendeleev, a lot has changed regarding our understanding of what matter is. A similar process is likely to characterise the study of moral cognition, and this is nothing to be afraid of.

[117] In a nutshell, the terms of a perfect psychological theory would be natural kind terms, in that they would identify the psychological mechanisms and processes that support true scientific generalisations about our mental life. There are many ways of defining natural kinds (e.g. Fodor, 1974), but this will suffice for present purposes.

Sinnott-Armstrong and Wheatley themselves suggest that moral psychologists should be "splitters" rather than "lumpers", that is, they should assume that the cases they observe are treated separately at the psychological level. Thus, judgments of moral obligation and permissibility concerning the trolley cases might be entirely different psychologically than the "same" judgments concerning Roy's predicament related at the beginning of the first chapter, for example. Yet as they themselves admit, this is merely a rule of thumb, and it can be taken too far (for example, are the intuitions generated by the different trolley cases treated entirely separately at the level of psychological description?). Even worse, though, it is entirely vague. For instance, it is unclear what similarity metric should be used for lumping judgments together, as *some* lumping will always be required if we want to do science. Additionally, no matter what way we pursue the inquiry, that is, whether we are splitters or lumpers, we will run into errors, although errors of different kinds: missing generalisations vs. making ones that are inadequate. Which one of these errors we want to minimise is an entirely pragmatic question.[118]

In this chapter, I first advanced a case study by reviewing an extremely influential theory of moral judgment, which uses an alternative framework for the study of moral cognition to the one endorsed in this thesis. I have argued that unlike LA, the DP framework does not supply the kinds of distinctions and research questions that are likely to be fruitful for moral psychology. In the second half of the chapter, I discussed some direct criticisms of LA and responded to these. In the next chapter, I move on to defend the descriptive adequacy of a particular principle of moral cognition—namely, the "Ought Implies Can" principle—in the face of putative empirical counterexamples.

[118] Of course, Sinnott-Armstrong and Wheatley advance arguments in favour of moral judgment not being unified, which is their reason for recommending the splitter strategy. Unfortunately, I cannot address these in any further detail here, but suffice it to say that the jury is still out regarding this question (*cf*. e.g. Joyce; 2016; Kumar, 2015).

Chapter 3: The "Ought Implies Can" Principle and Descriptive Adequacy

Overview

In this chapter, I argue that the "Ought Implies Can" principle (OIC)—according to which if an agent ought to perform an action, then they can perform it—, is a descriptively adequate principle of I-morality, that is, it correctly describes ordinary moral judgment. I proceed as follows. In Section 1, I propose that one way of selecting candidate principles of I-morality is by taking moral philosophy as the point of departure. Philosophical consensus regarding the "correctness", conceptual coherence or normative adequacy of a moral principle provides *prima facie* evidence for the descriptive adequacy of that principle. In Section 2, I single out OIC as just such a principle, that is, one in the case of which there is a strong consensus among ethical theorists. Contra this consensus, in recent years, experimental philosophers have directly challenged the hypothesis according to which OIC is descriptively adequate. The first version of this challenge is presented in Section 3. In Section 4, I show how it can be resisted. I develop a methodological criticism of these experimental studies and present novel data that supports OIC's descriptive adequacy. I address a second version of this challenge in Section 5, critically assess it and argue that it is inconclusive. Thus, the hypothesis according to which OIC is descriptively adequate remains plausible in the face of this empirical criticism.[119]

1. Moral Philosophy, Moral Cognition, and Empirical Research

The picture of moral psychology I have been defending so far should be clear enough: moral psychology is in the business of discovering the principles of I-morality and of articulating how they are implemented and acquired. But how can one discover what such principles are? A possible strategy is to collect data on moral judgments at large and attempt to systematise them by introducing principles that appear to capture the data. Then, we may further investigate whether these generalisations are on the right track by applying them to novel situations, and if so, explore what it is about the mind that makes them descriptively adequate. But of

[119] Sections 3 to 5 of this chapter are based on a paper I wrote in collaboration with Holly Lawford-Smith and Paulo Sousa titled *Does Ought Imply Can?* (Kurthy, Lawford-Smith & Sousa, 2017). All the studies presented here were conceived of and conducted collaboratively with Lawford-Smith and Sousa. I have decided to keep the use plural pronouns in the collaborative parts of the chapter.

course, such a method would be far too random. For a start, it is not clear just what kinds of judgment we should collect data on in the first place. Moreover, such data are likely to be compatible with a whole host of putative principles.

Another—far more feasible—option is to start with candidate principles. But which ones? We may wish to proceed by considering introspective evidence, by consulting evolutionary theories or comparative data, by scouring the empirical literature for ideas, and so on. There is not a single correct answer to this question, and all of these methods are understood to be heuristic. One promising strategy—the one I endorse in this thesis—is to take moral philosophy as our point of departure: the fact that normative ethicists or meta-ethicists agree on the correctness, conceptual cogency, or normative adequacy of a moral principle may be taken as *prima facie* evidence that the principle accurately describes ordinary judgment, and thus, potentially, the operations of I-morality.

This strategy requires some clarification. When ethicists argue for and endorse a moral principle, they may do so for various reasons. Amongst these, a prominent one is that the proposed principle accords with the moral judgments the ethicist in question would naturally and intuitively make. In such a case, the ethicist in effect asserts the descriptive adequacy of the proposed principle *as regards his or her I-morality*. To the extent that this generalises to other ethicists, we can treat this agreement as indicative of a trend that may be worth investigating further, since its application might well extend to other I-moralities.

Ultimately, however, consensus amongst ethicists does not obviate the need for the empirical work. One reason for this—besides the obvious issue of generalisability (philosophers are peculiar individuals, aren't they?)—is that ethicists often reason *about* the principle in question, for example in terms of whether it is cogent, or rational, or normatively correct. These types of reasoning may strongly influence the ethicist's verdict about whether the principle has to be accepted or endorsed. However, such reasoning is not relevant to the issue of whether a principle is part of (or, at least, an accurate description of) I-morality. After all, recall that the narrow function we assigned to FM in Chapter 1 was that of appending representations of moral-deontic status to representations of actions. This is clearly not the kind of task reasoning about, say, conceptual coherence involves.[120] Therefore, if one is interested—as we are—in whether the candidate principle characterises I-morality, there is simply no substitute for doing the empirical work. This involves

[120] Another way of putting the same point is this. When ethicists endorse/reject a normative or metaethical principle, they do so not only by testing the principle against their immediate intuitive judgments, but also by reasoning about it. The issue here is that I-morality only concerns one's immediate intuitions, not how such intuitions are elaborated downstream by other cognitive systems, including "central cognition". Thus, what ethicists say about a certain normative principle is only a partially reliable guide as to the operations of FM. It is in fact possible that what ethicists say about a certain principle reflects the working of other cognitive mechanisms.

finding out how ordinary people react to situations that the descriptive version of the principle would generate predictions about. This is why in this chapter I do both things. First, I focus on a principle largely agreed upon by ethicists. Second, I show that this principle correctly describes lay moral judgments. On the basis of these two lines of evidence, I conclude that such a principle is likely to be part of I-morality. What principle? You find the answer in the next section.

2. "Ought Implies Can" as a candidate principle of I-morality

As an example of a prominent case of broad agreement in the field of ethical theory, one of the most promising candidates is the "Ought Implies Can" principle (henceforth, "OIC" or "OIC principle"). According to OIC, if a person ought to perform an action, then he or she can perform it—where *ought* is predominantly understood to refer to a moral obligation, and *can* is taken to denote (physical) ability, or ability plus opportunity (see e.g. Vranas, 2007). OIC is often discussed in terms of its equivalent contraposition, which is arguably where its intuitive grip derives from (e.g. King, 2017). This has it that if a person cannot perform an action, then it is not the case that he or she ought to perform it.

Take, for instance, Roy's predicament familiar from Chapter 1, Section 1. To recall, in this little story, Roy is having a stroll next to a pond in which he notices a small child who is about to drown. This example served as an illustration of a case in which we automatically attribute a moral obligation to a person.[121] Now assume that as soon as Roy notices the drowning child, he suffers a stroke, causing paralysis in one side of his body, as well as partial blindness, lack of speech and dizziness, thus rendering him completely incapable of swimming—or even walking for that matter. Now, according to OIC, Roy no longer has a moral obligation to save the drowning child simply because he cannot.[122]

The OIC principle has been widely accepted by ethicists in various normative and descriptive guises (Griffin, 1992; Haji, 2002; Howard-Snyder, 2006; Sapontzis, 1991; Streumer, 2007; Vranas, 2007;

[121] The fact that the person is fictional need not be problematic from the point of view of moral psychology as long as FM does not have access to this information. This seems reasonable: when we think about stories we obviously know to be fictional, the moral judgments we make in connection with them seem no less automatic and intuitive.

[122] Just a reminder: of course we are interested in the *descriptive* (and ultimately psychological) version of OIC according to which, in such a case, we no longer attribute a moral obligation to Roy. Thus, the question cannot be whether it is *true* (in some sense) that Roy is under a moral obligation in this case, or whether asserting that he is is normatively adequate (for general comments, see Chapter 1, Section 1).

Zimmerman, 1996). It was famously endorsed by Kant,[123] and indeed until recent years, the function of OIC in the moral philosophy literature was often that of an axiom the truth of which was held to be so evident as to require no defence (Howard-Snyder, 2006). OIC also features as a theorem in some standard formulations of deontic logic (Anderson, 1967; Kanger, 1971), where it is sometimes referred to as *Kant's Law* (see generally Hilpinnen, 1971; McNamara, 2018). Furthermore, in the guise of the dictum "impossibilium nulla est obligatio" (i.e. "impossible obligations are invalid"), OIC has been enshrined in certain traditional as well as modern legal systems, such as Justinian's *Corpus Juris Civilis* or *Bürgerliches Gesetzbuch*, the civil code of Germany (Zimmermann, 1990, pp. 686*ff*.), and it has also been endorsed by legal theorists, such as Hans Kelsen (1991) or Lon L. Fuller (1969), among others.[124] Finally, most tellingly perhaps, even the critics of the principle (e.g. Saka, 2000; Sinnott-Armstrong, 1984; Stocker, 1971) endorse *some* version of OIC, if not that which has been more or less canonically assumed or defended (more on which see Chapter 4).

It is the OIC principle with which the rest of this thesis will be concerned. More specifically, I will be concerned with the question of whether OIC accurately describes the way humans intuitively think and reason morally (this chapter), and, if so, how a theory of FM or I-morality can account for this pattern in our judgment (next chapter).

2.1. Shape

The idea that OIC might be a descriptively adequate principle of I-morality should be understood in the following way. First, we do not expect individuals to endorse the principle as such: we merely require that particular moral judgments conform to it. In other words, we ought to avoid the conflation of express and operative principles. Second, there is no presumption of strict descriptive adequacy (*cf*. Chapter 1, Section 6; Chapter 2, Section 2.2). That is, we do not expect individuals to reason *in terms of* OIC, merely that they reason in line with it and thus that their judgment is accurately predicted by OIC. In yet other words, at this point OIC need only provide a functional description of moral judgment, not a computational/algorithmic one. The latter types of problem (algorithmic description) is an objective for future inquiry (but see also Chapter 4).

[123] For example, the following passage is from *Religion Within the Boundaries of Mere Reason*: "if the moral law commands that we *ought* to be better human beings now, it inescapably follows that we must be *capable* of being better human beings" (1973, 6: 50; emphasis original). (See Stern, 2004; Ranganathan, 2010 on the extent of Kant's commitment to OIC.)

[124] In the case of Kelsen, this is all the more significant since he also held there to be no "moral" limitations on the content of legal norms, as suggested by the formula, "Jeder beliebige Inhalt kann Recht sein" (see Grabowski, 2013, *p*. 20, *fn*. 35, *pp*. 302-3).

2.2. Testing

OIC has been largely endorsed by ethicists. So far so good. But, as I have explained earlier on, this is at best *prima facie* evidence for the claim that OIC is descriptively adequate. To make a stronger case for this claim, we need to collect data about everyday moral judgments. But what data exactly? One thing to note at the outset is that, in its standard formulation, OIC is stated in the form of a conditional that is transformable into a universal generalisation. That is, the conditional "if an agent has (or is understood as having) an obligation to ϕ, then the agent can (or is understood as able to) ϕ" should be read as the universal generalisation: "for all cases in which an agent has (or is understood as having) an obligation to ϕ, the agent can (or is understood as able to) ϕ". However, it is logically impossible to *empirically demonstrate* the truth of a universal generalisation (no matter how many confirming instances we find, OIC is never *proven* to be descriptively adequate, since there is no logical reason why the next datapoint should not be a non-conforming instance). Therefore, if we want to know whether OIC accurately predicts how (ordinary) people think and reason, we should rather look at the cases that are most likely to go against it. If the falsifying instances are not forthcoming, we have a good reason to conclude that OIC is descriptively adequate.

The contrapositive formulation comes in especially handy in this context: roughly, it states that "if an agent *cannot* ϕ, then that agent is *not* under obligation to ϕ" (interpreted descriptively, of course). Any instance in which an agent is conceptualised as unable to perform an action, but is understood, at the same time, as having an obligation to do so would be *prima facie* evidence against the principle being operative. (As we shall see later, the best candidates will be constituted by instances of self-imposed inability, that is, cases in which the agent is causally responsible for the emergence of his or her inability—see Section 5 as well as the next chapter).

Fortunately, some data have already been collected regarding judgments in connection with scenarios relevant to the question at hand (Mizrahi, 2015; Buckwalter & Turri, 2015; Chituc et al., 2016; Turri, 2017). The conclusion of these incipient studies regarding the descriptive adequacy of OIC have been rather unfavourable towards it: in general, the conclusion seems to be that OIC fails as a descriptive principle. One of these papers, namely that by Buckwalter and Turri (2015), stands out in terms of significance, since it claims to have demonstrated OIC inconsistent judgment even in cases which have traditionally been thought of as paradigmatic "OIC supporting". I address this study in detail in sections 3 and 4 below and conclude that, due to systematic problems in the study design, Buckwalter and Turri's conclusion, namely, the rejection

of the descriptive adequacy of OIC, is not warranted. In Section 5, I consider another study arguing for a similar conclusion as that of Buckwalter and Turri (2015).

3. First challenge: Buckwalter and Turri (2015)

In a paper titled *Inability and Obligation in Moral Judgment*, Wesley Buckwalter and John Turri have recently presented evidence that ordinary people do not reason in line with the OIC principle (Buckwalter & Turri, 2015—see also Mizrahi, 2015, and discussion in Kurthy & Lawford-Smith, 2015): with a series of studies, they claim to have demonstrated that, in people's judgments, obligations persist irrespective of whether those who hold them have the ability to fulfil them. In their studies, participants had to read stories in which a person is under an obligation but is subsequently described as unable to fulfil it.[125] For instance, participants in one study were asked to consider a case in which an agent ("Walter") promises to pick his friend ("Brown") up from the airport (the promise creating the obligation) but later becomes involved in a car accident and thereby rendered physically unable to keep the promise. Participants were then presented with the OIC probe, asking them to choose one of the following randomly sequenced statements (the numbers are included merely for convenience here):

1. **Walter is obligated to pick up Brown at the airport, but Walter is not physically able to do so.**
2. **Walter is not obligated to pick up Brown at the airport, and Walter is not physically able to do so.**
3. **Walter is obligated to pick up Brown at the airport, and Walter is physically able to do so.**
4. **Walter is not obligated to pick up Brown at the airport, but Walter is physically able to do so.**

In this and other scenarios—varying *inter alia* the source of the obligation involved (e.g. a promise or a social role), the type of inability (e.g. a physical restriction or a constraining feature of the environment), and the seriousness of the consequences of the obligation not being fulfilled (minor or fatal)—participants overwhelmingly chose the first option: "obligated, but not able" (option 1 above). On the face of it, this choice contradicts the OIC principle, since it attributes to the individual both an obligation and the inability to fulfil it.

[125] Buckwalter and Turri were also interested in probing whether people have more difficulty in perceiving inability when the source of the inability is mental rather than physical, e.g. due to clinical depression. Since this issue is tangential to the OIC principle, we leave it completely aside in this chapter.

Moreover, to confirm that participants understood the situation as involving a literal inability to fulfil the obligation, the studies included, after the OIC probe, an inability-comprehension probe, asking subjects whether the person under the obligation was literally unable to fulfil it. The great majority of participants confirmed that there was literal inability, and eliminating the few participants who denied literal inability did not change the general pattern of the results reported in the paper. Thus, Buckwalter and Turri conclude with the claim that "commonsense moral cognition rejects the principle that ought implies can" (2015, *p*. 1)—that is, in terms of our earlier discussions, OIC fails the test of descriptive adequacy by generating judgments that are not actually made by ordinary people.

The studies in the paper testing whether ordinary people make judgments consistent with the OIC principle also included, after the inability-comprehension probe, a blame probe, investigating whether participants would consider the individuals in their stories blameworthy for not fulfilling their obligations. They found that the great majority of participants denied that the individual is to blame in this respect, and suggested on the basis of this finding as well as the results of a separate study focusing directly on the relation between blame and inability that, for ordinary people, "Blame Implies Can". It is important to note that the traditional view of the relation between blame and obligation as far as inability is concerned is that the presence of an inability undermines blame by eliminating the perception of wrongdoing—in particular, by eliminating the perception that someone did something wrong in not fulfilling his or her obligation because in fact the obligation was cancelled by the inability (Hieronymi, 2004; Levy, 2005). Therefore, given that the above results indicate that the presence of an inability undermines blame without cancelling the obligation (and hence without eliminating wrongdoing), Buckwalter and Turri also suggest that the traditional view of the relation between blame and obligation does not appropriately describe the relation between these concepts in ordinary cognition, and may be an invention of philosophers trying to "validate excuses" (Turri & Blouw, 2015; see Section 5 further below).

In the following section (Section 4), we first question the implication of the results reported in Buckwalter and Turri's paper with new evidence based on the same scenarios of inability. We argue first that there are crucial problems with the design of Buckwalter and Turri's studies (Section 4.1). Then, we report two studies indicating the main problem with this design—namely, it does not seem to provide an appropriate test of whether ordinary people reason in line with the OIC principle (Section 4.2). Next, we provide an overview of our new studies with an improved design (Section 4.3). After that, we report four studies showing that the great majority of participants make judgments compatible with the OIC principle and with the traditional view of the relation between blame, obligation, and wrongdoing (sections 4.4-4.7). We summarise our results in Section 4.8. In Section 5, we address and reject another empirical challenge against OIC, namely, that

provided by Chituc el al. (2016). Finally, we consider some broader issues, such as the type of reasoning involved in participants' judgments, the extent to which our results might generalise to cases involving culpable inability, and a possible deflationary explanation of our results in terms of excuse validation.

4. Response to Buckwalter and Turri's challenge

4.1. Potential problems with Buckwalter and Turri's design

Aspects of Buckwalter and Turri's design, in particular the way in which the list of options of the OIC probe are framed, may make the option "obligated, but not able" the sole plausible answer, though in a way that is not inconsistent with the OIC principle.

The stories in Buckwalter and Turri's studies are characterized by an individual under an obligation who is eventually described as unable to fulfil the obligation. There is an obvious but trivial sense in which each story, taken as a whole, involves both an obligation and an inability. The inability creates tension with the expectation of fulfilment generated by the obligation. The option "obligated, but not able" matches this description of the story as a whole, while the other options do not, since they either exclude an obligation ("not obligated, and not able") or include the ability to fulfil the obligation ("obligated, and able"; "not obligated, but able"). Moreover, the option "obligated, but not able" has an ordinary temporal reading (i.e. "obligated, but *subsequently* not able") that mirrors the temporal narrative of the story (i.e. an obligation is made salient early in the story, then later an inability is made salient). This, too, renders the option "obligated, but not able" the best description, because it captures the temporal dimension of the contrast involved in the story as a whole.

In sum, according to our interpretation, when participants choose the option "obligated, but not able", they are not saying that the person is *still* under the obligation even when there is an inability to fulfil it. That would be inconsistent with the OIC principle. Rather, they are saying that the stories involve a contrast between a presumed obligation (made salient first) and an inability to fulfil the obligation (made salient second). This is not inconsistent with the OIC principle, because it may well turn out that the subjects of these studies would accept the obligation for as long as they think there is ability, and reject the obligation after the inability is made evident in the story.

There is another aspect of Buckwalter and Turri's design that may have contributed to the problem we have outlined, and consequently to the predominant selection of the "obligated, but not able" option. The instruction for the OIC probe ("choose the option that best applies") implies that there is a factually correct

alternative among the options, and may suggest to participants that they are being tested on whether they interpreted the story correctly (as if the OIC probe had the same type of function as the inability-comprehension probe—the second probe of their design described earlier). If participants understood the OIC probe in this way, then rather than providing their personal opinion on the relation between the obligation and the inability, they would simply provide the *best description* of what is involved in the story as a whole, which is plausibly the option "obligated, but not able", as discussed above.

Finally, it is important to note that none of the stories in Buckwalter and Turri's studies explicitly state the obligation at stake in the story. In the promise scenario, the story says only that someone makes a promise; in their social-role scenarios, it says only that someone has a social role (e.g. that of a lifeguard); in another scenario, it simply describes a situation in which a small child is drowning and there is a stranger around who could easily help the child. Thus, the participant has to infer from the information given in the initial part of the story (i.e. from the fact that someone made a promise, that someone has a social role, or that someone could easily help) the existence of the corresponding obligations (i.e. the obligation to keep the promise; the obligation related to the social role; the obligation to help the drowning child). True, these inferences are somewhat obvious, and the fact that the obligations are left implicit in the stories is not a problem in itself. However, given the aforementioned problems, it may well be that at least some participants took the OIC probe to be a test on whether they believe that the initial situation described in the story entails an obligation, and chose the first option to confirm that they indeed believe that the relevant aspects in the story warrant an obligation.

4.2. Initial evidence for the relevance of the problems identified: Two studies

4.2.1. Study 1

In this study, we test our main claim about what lead the great majority of subjects in Buckwalter and Turri's studies to choose the "obligated, but not able" option. As we discussed above, we claim that there is an obvious sense in which the option "obligated, but not able" is the correct answer in the context of Buckwalter and Turri's design because of two main factors: (i) the option describes the fact that each story as a whole involves a contrast between an obligation and an inability to fulfil the obligation, and (ii) the option mirrors the temporal narrative of each story (i.e. an obligation is made salient early in the story, then later an inability is made salient).

Two predictions follow from our claim. First, there would be a substantial reduction of "obligated, but not able" responses in the results if one were to simply replace the connectives "but" and "and" in the original options with connectives that more clearly convey the main point of the OIC probe (i.e. that make participants focus on whether there is an inferential relation between the attribution of an obligation and that of a corresponding inability). Second, there would also be such a reduction if one were simply to invert the order of the obligation and inability clauses of the original options (e.g. changing "obligated, but not able" to "not able, but obligated"), thus creating a mismatch between the order of the clauses and the temporal narrative of the story. Our first study tests these predictions.

4.2.1.1. Method

Participants

Participants were 123 adults (56 female, 67 male; M_{age} = 36.84; *SD* = 11.19; range = 52; 98% reporting English as their first language). Our data collection methodology was similar to that employed in Buckwalter and Turri's studies. In all studies to be reported in this chapter, participants were recruited, tested and compensated online. We used Amazon Mechanical Turk and Qualtrics as the online platforms. All participants were US residents. Each participant was paid $0.50 for approximately 4 minutes of their time. Furthermore, in all studies, we collected around 40 responses per condition. Participants were allowed to participate in only one of the studies (or conditions) reported in this chapter.[126]

Design, Materials and Procedure

The study used Buckwalter and Turri's original design of the "Walter promise" scenario (Experiment 1, Physical condition), but without the question asking whether Walter is to blame, and crucially, with three types of between-subjects OIC probes: the original four options of Buckwalter and Turri's design as described in Section 4.2.1 (Original condition); four options using "even if" and "because" as connectives, instead of "but" and "and" (Inferential relation condition); and the original four options with the order of the obligation and inability clauses inverted (Inverted order condition). The OIC-inconsistent and OIC-consistent options of the Original, Inferential relation, and Inverted order conditions were as follows (for the sake of simplicity, we leave aside the two options where Walter was described as able to fulfil his obligation):

[126] Our research design, including the procedure for informed consent, was reviewed by the Research Ethics Committee of the School of History and Anthropology at Queen's University, Belfast, UK and by the Research Ethics Committee of the University of Sheffield, UK. Written informed consent was obtained from all participants in all of the studies reported in this chapter.

1. (*Original*) Walter is obligated to pick up Brown at the airport, but Walter is not physically able to do so.

   (*Inferential relation*) Walter is obligated to pick up Brown at the airport, even if Walter is not physically able to do so.

   (*Inverted order*) Walter is not physically able to pick up Brown at the airport, but Walter is obligated to do so.

2. (*Original*) Walter is not obligated to pick up Brown at the airport, and Walter is not physically able to do so.

   (*Inferential relation*) Walter is not obligated to pick up Brown at the airport, because Walter is not physically able to do so.

   (*Inverted order*) Walter is not physically able to pick up Brown at the airport, and Walter is not obligated to do so.

4.2.1.2. Results

The results of this study are shown in Figure 3.1. In the Original condition, we replicated the results reported in Buckwalter and Turri's paper: 88% chose "obligated, but not able", while only 12% chose "not obligated, and not able" (N = 41). In the Inferential relation condition, we completely reversed the earlier results: only 5% chose "obligated, even if not able", while 88% chose "not obligated, because not able" (N = 42; 7% chose the remaining two options where Walter is described as able). Finally, in the Inverted order condition, the two relevant options were equally chosen: 47.5% chose "not able, but obligated" and 52.5% chose "not able, and not obligated" (N = 40).

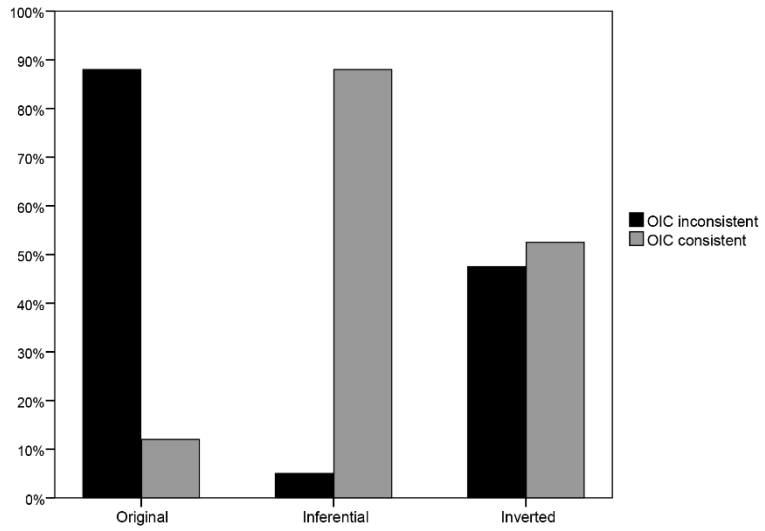

Figure 3.1. Percentage of responses consistent or inconsistent with the OIC principle in each of the three conditions.

Confirming that there was a substantial reduction of the obligated/unable type of response in the Inferential and Inverted conditions, Chi-square tests (with obligated/unable responses coded as "1" and the remaining responses coded as "0") show that these conditions differed significantly from the Original condition: χ^2 (1, 83) = 47.82, $p < .01$, $\phi = .76$, for Inferential versus Original; χ^2 (1, 81) = 15.09, $p < .01$, $\phi = .43$, for Inverted versus Original.

4.2.1.3. Discussion

The results of this study are consistent with our main claim about a problem with Buckwalter and Turri's design—the subjects choosing the option "obligated, but not able" do not interpret it in a way that is inconsistent with the descriptive version of the OIC principle. If this option had been interpreted in terms of *obligated at the time of the inability*, the order of the clauses should not have mattered. It is also worth pointing out that the justifications following selections of the "obligated, but not able" response in the Original condition suggest that, with this option, the participants were merely acknowledging that the story involved both an obligation and an inability, irrespective of whether the obligation is perceived to be in force subsequent to the onset of the inability. The majority of participants provided justifications such as:

> **"[Walter] promised that he would pick Brown up at the airport, which gives him an obligation to pick him up, but he was injured in a serious car accident and is therefore unable to do so."**

"He has committed to do it, and Brown is depending on him. However due to the car accident he won't be able to make it."

"He is obligated because he promised but he is unable to because of the accident."

"He agreed to do it but he subsequently became physically unable."

"He promised that he will pick up Brown at the airport. He was in an accident so he is unable to actually carry out the task."

Finally, the results of the Inferential relation condition, using connectives that arguably make the point of the OIC probe more salient, suggest that ordinary people reason in a way that is consistent with the OIC principle, at least in this type of scenario.

4.2.2. Study 2

Although our previous study suggests that Buckwalter and Turri's design is problematic just in the way we discussed, one could still argue that (a) the Inferential relation condition merely distorted the results due to a different logical framing of the options, that (b) the results of the Inverted order condition do not establish *directly* that the response options of Buckwalter and Turri's design fail to test whether participants' reasoning conforms to the OIC principle, and that (c) the qualitative justifications to the "obligated, but not able" option of the Original condition do not establish *conclusively* that this option is understood in a way that is compatible with the OIC principle. In Study 4 below, we demonstrate that the inferential relation framing does not distort the results. In this study, by asking a follow-up question regarding the relevant response options of Buckwalter and Turri's design, we provide more direct and conclusive evidence indicating that this design is problematic in the way we claim.

4.2.2.1. Method

Participants

Participants were 43 adults (17 female; M_{age} = 32.84; *SD* = 8.47; range = 34; 100% reporting English as their first language).

Design, Materials and Procedure

This study was also based on the story about Walter's promise with one crucial addition. We asked a clarificatory follow-up question in relation to the options "obligated, but not able" and "not obligated, and not able". This follow-up question appeared on a different page, after the participant had provided a response to the original OIC probe (see previous study). Participants choosing the "not obligated, and not able" option were confronted with the following question (Question A):

> **You chose the option "Walter is not obligated to pick up Brown at the airport, and Walter is not physically able to do so." With this choice, do you mean that Walter is *no longer under the obligation* to pick up Brown at the airport *after* he becomes physically unable to do so? (Yes/No)**

On the other hand, participants choosing the "obligated, but not able" option had to answer the following question (Question B):

> **You chose the option "Walter is obligated to pick up Brown at the airport, but Walter is not physically able to do so." With this choice, do you mean that Walter is *still under the obligation* to pick up Brown at the airport *after* he becomes physically unable to do so? (Yes/No)**

The order of the options (Yes/No) was randomised in both cases. Also in both cases, participants responding "no" were asked to explain their choice ("Please explain what you meant, then."). After the follow-up question, participants answered the inability-comprehension probe (i.e. the probe asking whether Walter was "literally unable"). We did not include a blame probe in this study either.

The logic of this study is very simple. If the selection of the "not obligated, and not able" option is to be taken as consistent with the OIC principle, then participants should predominantly answer "yes" to Question A. Concomitantly, if the selection of the "obligated, but not able" option is taken to be inconsistent with the OIC principle, then participants should predominantly answer "yes" to Question B. To spell it out clearly: if Buckwalter and Turri's design and conclusions are sound, then we should expect affirmative answers in both cases—but crucially so in the case of Question B, as it is the high relative frequency with which participants selected the "obligated, but not able" option that was interpreted as constituting the main evidence against OIC-consistent reasoning.

4.2.2.2. Results

The great majority of participants (74.4%) selected the option "obligated, but not able", while only the minority (16.3%) endorsed the option "not obligated, and not able" (the remaining 9.3% chose "obligated, and able"). As in the Original condition of the previous study, and in Buckwalter and Turri's paper in general, the selection of the apparently OIC-inconsistent option (vs. all the other options collectively) is significantly above chance level—goodness of fit against chance: χ^2 (1, 43) = 10.26, $p < .01$, $\phi = 0.49$. The overwhelming majority (93%) agreed that Walter is literally unable to pick up Brown form the airport.

For the rest of the analysis, we exclude participants who denied literal inability. All 7 participants (100%) choosing the "not obligated, and not able" option answered "yes" to Question A, confirming that their reasoning is consistent with the OIC principle. Now, 23 out of 31 participants (74%) selecting the "obligated, but not able" option said "no" in response to Question B—goodness of fit against chance: χ^2 (1, 31) = 7.26, $p < .01$, $\phi = 0.48$—, indicating that, with their response, they did *not* mean that Walter is still under the obligation after he becomes physically unable to pick up Brown.

4.2.2.3. Discussion

Since the great majority of the participants choosing the "obligated, but not able" option answered "no" to the follow-up question (Question B), our results indicate more directly and conclusively that the selection of this option does not track OIC-inconsistent reasoning. In other words, as we discussed above, Buckwalter and Turri's design does not seem to be appropriate to test whether participants reject the OIC principle. The justifications of participants who chose the "obligated, but not able" option and answered "no" to the follow-up question also support this interpretation. Some participants were emphatic that the inability annuls the obligation, suggesting that it did not even occur to them that their response could be taken as a case of obligation ascription after the accident:

> **"Obviously Walter is no longer obligated to pick up Brown from the airport and anyone who tries to philosophically argue the case is limited in their scope of understanding of reality. Walter agreed to pick up someone from the airport but after being severely incapacitated due to a car accident he is no longer able (or obligated) to pick up the person and he should find an alternative."**

Furthermore, many participants pointed out that, in choosing the "obligated, but not able" option, they had intended to express the view that Walter is indeed obligated, but only up to the point at which he becomes incapacitated, which confirms our criticism of Buckwalter and Turri's design:

> **"I meant that Walter was obligated to pick him up until he became physically unable to do so."**

> **"I meant that Walter was obligated to pick up Brown. However, once he was physically unable to, he was no longer obligated."**

> **"He WAS obligated, but cannot physically do it so the obligation is no longer on him."**

> **"He agreed to do it, so he is obligated once he does that. But after getting hurt, he is not still under that obligation."**

> **"He had agreed on picking his friend up. But when he got into a serious accident, the obligation was suspended because he was no longer in the same position to help out his friend."**

4.3. The case against Buckwalter and Turri: Overview of new studies

With our first two studies, we provided strong evidence that the design used in Buckwalter and Turri's experiments does not constitute an appropriate test of whether ordinary people reason in line with OIC. In the studies to follow, we utilised a design that addresses most of the aforementioned problems and makes the task much clearer and simpler for the participants. We modified Buckwalter and Turri's design in the following ways:

(i) We changed some very trivial details of the stories to make it clearer to participants that the characters in the stories are unable to fulfil their obligation, and/or to avoid misinterpretations of the story.

(ii) We changed the instructions of the OIC probe and the inability-comprehension probe to make their different purposes obvious to participants.

(iii) We positioned the inability-comprehension probe before the OIC probe, that is, just after participants read the story. And in case a participant denied that the character in the story was literally unable to fulfil their obligation, we explained to the participant that in fact the character *was* unable

to do so by emphasising the relevant elements of the story; then we asked the participant to assume that there was literal inability before answering the OIC probe. (In our studies, hardly any participants disagreed that the character was literally unable to fulfil their obligation and excluding these participants from the analysis changes nothing in terms of our results and conclusions.)

(iv) We simplified the OIC probe by reducing its four options to two: one consistent with the OIC principle, another inconsistent with it. (Note that the two eliminated options, which say that the character in the story is able to fulfil her obligation, are completely irrelevant to testing whether people make judgments consistent with the OIC principle.)

(v) We phrased the two options of the OIC probe in a way that makes it clearer to participants what the point of the OIC probe is (e.g. using the connectives "because" and "even if" instead of "and" and "but").

(vi) We included a justification probe asking participants to explain their OIC choice, in order to gain some qualitative insight into the reasons motivating participants' choices. (This step was introduced after the OIC option was irreversibly selected, so there is no reason to suppose that it could interfere with the quantitative results of the OIC probe).

The great majority of the above changes should not be controversial, as they merely clarify and/or simplify the task for the participants. Although changing the connectives of the options of the OIC probe may seem controversial, in Study 4, we demonstrate that our usage of "even if" and "because" is not problematic.

Some of Buckwalter and Turri's studies are, arguably, much less central to testing the OIC principle (e.g. Experiment 7, which tests whether the difference between moral and legal obligation is relevant to the principle). Accordingly, our studies focused on those studies that are most central to the OIC principle, namely, Experiments 1, 2, 4 and 5.

4.4. Study 3: Promise

In this study, we used our new design to test whether people make judgments consistent with the OIC principle in relation to obligations generated by promises, using the "Walter" scenario familiar from the first two studies as well as from the first experiment reported in Buckwalter and Turri's paper, where it was found that 80% of participants chose the "obligated, but not able" option, apparently contradicting the OIC principle. In addition, we used different ordinary expressions that are commonly thought to encode the concept of obligation ("obligated", "duty", "ought"), in order to see whether there is variation in judgments as a result of these.

4.4.1. Method

Participants

Participants were 127 adults (60 female; 67 male; M_{age} = 33.95; *SD* = 11.54; range = 53; 97% reporting English as their first language).

Design, Materials and Procedure

After indicating informed consent, participants read the following story:

> **Walter promised that he would pick up Brown from the airport. But on the day of Brown's flight, Walter is in a serious car accident *and is hospitalized*. As a result, Walter is not able to pick up Brown at the airport.**[127]

We added "and is hospitalized" to boost the understanding that Walter is unable to pick up Brown at the airport.

Participants were then presented with the inability-comprehension probe, whose instruction and question were as follows: "First, we would like to ask you a question to check whether you understood the story. According to the story, is the following statement true?" The statement that participants had to evaluate was: "Walter is literally unable to pick up Brown at the airport because Walter is hospitalized". If they answered "yes", they were presented with the OIC probe. If they answered "no", they were given an explanation indicating that Water is indeed unable to pick up Brown because his "injuries are so serious that he requires hospitalization"; then they were asked to assume that this is the case before answering the OIC probe.

The instruction and question of the OIC probe were as follows: "Now, we would like to know your personal opinion about the situation. There isn't a correct answer here. Which statement best reflects your personal opinion about the situation?" Participants had to choose between two randomly sequenced statements, each consistent or inconsistent with the OIC principle. In order to probe participants' judgments with different ordinary expressions that encode the concept of obligation ("obligated", "duty" or "ought"), participants were randomly assigned to one of three phrasing conditions:

[127] Here and elsewhere, the divergences from the wording of the original stories as used in Buckwalter and Turri's studies are shown in italics. For example, in this case, the only change we introduced was the addition of the phrase, "and is hospitalised". See the main text for more.

1. **Under these circumstances, Walter is still obligated to (Walter still has a duty to / Walter still ought to) pick up Brown at the airport, even if he is unable to do so.**

2. **Under these circumstances, Walter is not obligated to (Walter does not have a duty to / it is not the case that Walter ought to) pick up Brown at the airport, because he is unable to do so.**

After choosing one of the above statements, participants were asked to justify their choice: "Please explain why you marked this option".

Finally, participants answered a blame probe, enquiring about the degree to which they believed that Walter deserved blame for not fulfilling the obligation: "To what extent is Walter to blame for not picking up Brown?" Participants answered this probe on a seven-point scale, with "1" indicating "No blame", "4" indicating "Moderate blame", and "7" indicating "Full blame".

4.4.2. Results

Almost everyone (98%) agreed initially that Walter was literally unable to pick up Brown at the airport. The phrasing conditions produced no effect, χ^2 (2, 127) = .01, p = .99, with 100%, 98% and 100% of participants choosing the option consistent with the OIC principle in the "obligated", "duty" and "ought" conditions respectively. Across the phrasing conditions, 126 out of 127 participants chose the option consistent with the OIC principle—goodness of fit against chance: χ^2 (1, 127) = 123.03, $p < .01$, $\phi = 0.98$.

Blame ratings did not differ across phrasing conditions either—$F(2, 124) = 1.04$, $p = .36$. In general, blame ratings were very low (M = 1.47; SD = 1.02), with 92 of 127 participants opting for the "1" rating (i.e. "no blame").

4.4.3. Discussion

With our improved design, we completely reversed the results of Buckwalter and Turri using three ordinary expressions that are commonly thought to encode the concept of obligation, suggesting that there is no variation in judgment due to the examined terminological variation in this domain.

Participants' justifications suggest that, actually, *none* of their answers were inconsistent with the OIC principle. Justifications of participants who chose the "not obligated" option often expressed that, given the inability, it would be unintelligible to attribute an obligation, or that it is self-evident that the obligation does not hold:

> **"It seems silly to say that it's immoral to not keep a promise in extenuating circumstances like this."**
>
> **"It makes no sense to say he should do something he isn't able to."**
>
> **"Because he is unable to do so, it is self-explanatory."**

Sometimes they even explicated the OIC principle literally or in terms of its equivalent contraposition:

> **"'Duty' assumes he will have the ability to implement his duty, just as a soldier is excused from duty when injured."**
>
> **"I think that the existence of a duty presupposes the ability to fulfil that duty. If it is impossible for that duty to be fulfilled, it does not exist."**
>
> **"If someone is unable to do something they can't be obligated to do it."**

Now, the justification of the only participant who chose the "obligated" option suggests that, instead of making a judgment incompatible with the OIC principle, the participant simply shifted the scope of the obligation at stake:

> **"Walter made an agreement with full intention of keeping it and if he cannot fulfill the agreement, notice should be sent and a proxy should be appointed to carry out the agreement as specified."**

In other words, rather than maintaining that Walter is still obligated to pick up Brown at the airport even if he is unable to do so, this participant seems to be saying that even if Walter cannot pick Brown up, he is still obligated to *do something else* to improve Brown's situation. Since our scenario leaves open the possibility that Walter could still do something else in this respect, the response of this participant (and crucially the reasoning behind it) does not necessarily conflict with the OIC principle (this kind of justification will show up in later

studies; we will refer to it as the 'scope-shifting problem', because it involves participants' changing the scope of the obligation to include new or alternative content).

Finally, the great amount of "no blame" answers plus the overall low mean of blame ratings shows that participants think that Walter's inability eliminated his blameworthiness for not picking up Brown at the airport, which is consistent with Buckwalter and Turri's blame results. However, contrary to their results, our results also suggest that participants think that the elimination of blame was linked to the fact that Walter had no related obligation under the circumstances, and, consequently, to the fact that Walter did not do anything wrong in not picking up Brown at the airport. In other words, our results are more consistent with the idea that ordinary cognition is in line with the traditional view on the relation between blame, obligation and wrongdoing.

4.5. Study 4: Playground safety worker

Social roles are normally seen as another source of obligations. In this study, we tested whether people make judgments consistent with the OIC principle in the context of an obligation entailed by the social role of a playground safety worker. The scenario we utilised corresponds to that used in the second experiment of Buckwalter and Turri's paper, where it was found that 98% ("duty" phrasing condition) and 88% ("ought" phrasing condition) of participants chose the "obligated, but not able" option, apparently contradicting the OIC principle. In addition, we tested whether the framing of our options in terms of the connectives "even if" and "because" inadvertently biased participants towards choosing the option that is consistent with the OIC principle.

4.5.1. Method

Participants

Participants were 86 adults (40 female, 45 male, 1 "other"; M_{age} = 37.67; SD = 13.25; range = 53; 98% reporting English as their first language).

Design, Materials and Procedure

Participants read first the following story:

> **Michael is a playground safety worker. He sees some broken glass in an area where kids sometimes play barefoot. But he is stricken by a sudden *full body* paralysis *that immobilizes him to the extent that he cannot even speak*. As a result, Michael is not able to *remove* the *broken* glass.**

As the italics show, we introduced three modifications to Buckwalter and Turri's version of the playground scenario. The first two of these were to boost the understanding of inability and/or to emphasise that there wasn't anything else that Michael could have done to improve the situation (e.g. ask other people to remove the broken glass), and thus to try to avoid the scope-shifting problem identified in the discussion of Study 1. The last modification replaced the phrasal verb "pick up" with the verb "remove," which arguably more clearly describes the content of Michael's obligation in this situation.[128]

Participants were then presented with the inability-comprehension probe, which asked them to evaluate the truth of the following statement: "Michael is literally unable to remove the broken glass from the area because he is completely immobilized." Depending on their truth evaluations, participants proceeded to the OIC probe as specified in Study 1.

The instruction and question of the OIC probe were the same as in the previous study. Since in Study 3, we provided evidence that different ordinary expressions encoding the concept of obligation do not affect the results of the OIC probe, we used only one phrasing for the statements of the probe in this study ("obligated"). However, participants were still randomly assigned to one of two conditions. In the "explicit" condition, participants had to choose between the same type of "obligated" statements of Study 3, while in the "implicit" condition these statements were presented without the inability clauses and their connectives:

1. **Under these circumstances, Michael is still obligated to remove the broken glass, even if he is unable to do so (Under these circumstances, Michael is still obligated to remove the broken glass).**
2. **Under these circumstances, Michael is not obligated to remove the broken glass, because he is unable to do so (Under these circumstances, Michael is not obligated to remove the broken glass).**

We included the implicit condition in this study because one may argue (rather implausibly in our view) that, rather than making more explicit the main point of the OIC probe, the connectives "because" and "even if" inadvertently bias participants to choose the option consistent with the OIC principle, thus distorting the results. Against this "framing" hypothesis, we predicted that there would be no effect of condition, since the fact that we asked the comprehension probe first plus the usage of "under these circumstances" and "still" already makes the main point of the OIC probe clear enough.

[128] The original story (Buckwalter & Turri, 2015, *p.* 5) read as follows: "Michael is a playground safety worker. He sees some broken glass in an area where kids sometimes play barefoot. But he is stricken by a sudden paralysis in his legs. As a result, Michael is not physically able to pick up the glass."

After answering the OIC probe, participants answered the justification probe and the blame probe, similarly to Study 3.

4.5.2 Results

Almost everyone (99%) accepted initially that Michael was literally unable to remove the broken glass. There was no effect of condition, χ^2 (1, 86) = .387, p = .53, with 88% and 84% of participants choosing the "not obligated" response in the explicit and implicit conditions, respectively. Thus, altogether, the overwhelming majority of participants (86%) believed that Michael did not have an obligation under the circumstances—goodness of fit against chance: χ^2 (1, 86) = 44.69, p < .01, ϕ = .72).

Blame ratings remained low (M = 1.79; SD = 1.41), with 59 of 86 participants opting for "no blame". A 2(condition) x 2(OIC option choice) between-subjects ANOVA on blame scores revealed a main effect of option choice, F(1, 82) = 35.6, p < .01, η_p^2 = .303, but no main effect of condition (p = .17) or interaction (p = .30). Thus, participants who chose the "obligated" option saying that Michael was obligated to remove the glass blamed him more (M = 3.67, SD = 1.67) than participants who chose the option that he was not obligated (M = 1.49, SD = 1.11). Accordingly, there was a significant correlation between option choice and blame ratings: r_{pb} = .53, p < .01.

4.5.3 Discussion

Once again, we completely reversed Buckwalter and Turri's results. Furthermore, as we predicted, whether the OIC options involved the inability clauses and their connectives did not affect which option was chosen. This indicates that an argument according to which the effect observed in Study 3 depends on our specific framing of the options, and, in particular, on the usage of the connectives "even if" and "because", is not plausible. Indeed, our results provide corroboration for our contention that it is Buckwalter and Turri's design (rather than ours) that systematically distorts the results.

Justifications for "not obligated" responses again showed that participants' responses were consistent with the OIC principle. In contrast, the justifications of the "obligated" responses (12 in total) were more varied and, overall, did not clearly indicate that these responses were incompatible with the OIC principle. Evincing the scope-shifting problem discussed in Study 1, some participants seem to have shifted the scope of the obligation to the idea that Michael still has the obligation to do (or try to do) something else to improve the situation:

"He has the job of playground safety worker, and he has been presented with an unsafe condition. If he can't remove the glass, he should call out to the kids to avoid the area, call out to another adult, or make some kind of effort to communicate the hazard."

"In some way if he knows there's broken glass and no one else is notified, there needs to be a way he can communicate with someone he can or warn the kids about it."

Since these participants seem to have misinterpreted our scenario in that they still envisaged that Michael could do something else, like informing other people, to ameliorate the situation (or since the description of our scenario does not rule out the possibility that Michael could at least make an effort to improve the situation), their "obligated" responses are not incompatible with the OIC principle.

Some participants seem to emphasise that Michael still has the obligation to remove the glass, not at the time of his paralysis but rather as soon as he recovers:

"Well Michael may be unable to physically remove it himself, but he is obligated to do so in the sense that he should remove it as soon as possible."

"(...) Of course if his condition worsens or doesn't let up then he cannot act on his obligation so he won't clean up the glass, but with the knowledge he should do it, if he can."

This type of justification suggests that in fact the reasoning of these participants is in line with the OIC principle.

Many participants seem to appeal to the connection between the obligation and the nature of Michael's social role (note that the word "responsibility" is often used in the sense of obligation related to a social role [17, 18]):

"It is still his responsibility as a playground safety worker."

"That's his job."

"It's his property. It's his responsibility to get it cleaned up even if he can't do it himself."

"I believe as a worker and having knowledge makes you responsible."

From these justifications, one may take that these participants indeed reason in a way that is not consistent with the OIC principle—the participants seem to believe that obligations related to social roles continue to be in force independent of the circumstances, and hence seem to accept that Michael is still obligated to remove the broken glass in that situation of inability.

However, it is still possible that these participants answered "obligated" simply to emphasise the obligations that are normally entailed by social roles, without necessarily rejecting the OIC principle. Because social roles are deemed to entail obligations, there is a sense in which the entailed obligations do not disappear in cases of inability, since the social role does not disappear with the inability (a playground safety worker does not cease to be a playground safety worker just because he is unable to fulfil his role in a specific situation). Accordingly, people may make a distinction between obligations that are normally entailed by a social role, and obligations that are in force at a specific point in time. This would make it possible for a playground safety worker *qua* playground safety worker to have an obligation to remove the broken glass, and yet this particular *paralysed* playground safety worker to not have that obligation. Thus, the above participants may be interpreting and answering the OIC probe simply in terms of the obligations that are normally entailed by a social role, in which case their responses are not necessarily inconsistent with the OIC principle, given that this principle has generally been assumed to be concerned with whether an obligation is still in force at the time of the inability. (It is important to note that this issue, which may have also prompted participants to choose the "obligated but not able" option in the related studies of Buckwalter and Turri, is different from the main criticism we delineated concerning the way this option is framed: even in the sense of a social-role obligation being in force, there is a trivial sense in which an obligation is involved in the story and leads one to choose the option "obligated but not able".

Finally, the large number of "no blame" answers and low mean of blame ratings, along with the positive correlation between these ratings and OIC responses (i.e. more blame, more "obligated" response) is more consistent with the idea that ordinary cognition is in line with the traditional view of the relation between blame, obligation, and wrongdoing.

4.6. Study 5: Lifeguard

In this study, we tested whether people make judgments consistent with the OIC principle, again in the context of an obligation entailed by a social role, but this time that of a lifeguard. While studies 3 and 4 involved an "internal" inability coming from physical restrictions, this study involves an "external" inability coming from constraints of the environment like distance in space. Furthermore, while studies 3 and 4 involved relatively minor consequences like not being picked up at the airport or stepping on broken glass, this study involves a life-and-death situation. The scenario we utilised corresponds to the one in Buckwalter and Turri's fourth experiment, where it was found that 93% of participants chose the "obligated, but unable" option that apparently contradicts the OIC principle.

4.6.1. Method

Participants

Participants were 42 adults (11 female, 31 male; M_{age} = 38.98; *SD* = 13.13; range = 49; 98% reporting English as their first language).

Design, Materials and Procedure

Participants read the following story:

> **Jessica is the only lifeguard at a remote ocean beach. Two struggling swimmers are about to drown, *and no one else is around except Jessica*. She rushes in to save them, but because of the *great* distance between the swimmers, it is physically impossible for her to rescue both swimmers. Jessica rescues one swimmer but not the other.**

We again introduced several small alterations to Buckwalter and Turri's version. The main modifications of the original scenario were again introduced in order to boost the understanding of inability and/or to emphasise that there wasn't anything else that Jessica could have done to improve the situation (e.g. ask for additional help). (Other minor stylistic modifications, not indicated here, were also introduced to improve readability).[129]

[129] The original story (Buckwalter & Turri, 2015, *p.* 5) read as follows: "Jessica is a lifeguard at a remote ocean beach. Two struggling swimmers are about to drown. Jessica rushes in to save them. But because of the very far distance between the swimmers, it is physically impossible for her to rescue both swimmers. Jessica rescues the one swimmer but not the other."

The rest of the procedure was exactly the same as in studies 3 and 4: inability-comprehension probe ("Jessica is literally unable to rescue both swimmers because they are too far apart"); OIC probe with justification probe; blame probe. In this study, there was only one OIC probe condition, with the following options:

1. **Under these circumstances, Jessica is still obligated to rescue both swimmers, even if she is unable to do so.**

2. **Under these circumstances, Jessica is not obligated to rescue both swimmers, because she is unable to do so.**

4.6.2. Results

Almost everyone (95%) agreed that Jessica was literally unable to save both swimmers. The great majority (79%) of participants felt that the agent was not obligated to save both swimmers—goodness of fit against chance: $\chi^2\,(1, 42) = 13.71, p < .01, \phi = 0.57$.

Blame scores remained relatively low ($M = 1.67$; $SD = 1.18$), with 28 of 42 participants opting for "no blame". However, in contrast with the previous study, participants choosing the "obligated" option did not ascribe significantly more blame to Jessica than participants choosing the "not obligated" one: $t(40) = 1.64, p = .21, d = .49$ ("obligated": $M = 2.11$; $SD = 1.45$; "not obligated": $M = 1.55$; $SD = 1.09$). Accordingly, there was no significant correlation between option choice and blame ratings: $r_{pb} = .19, p = .21$.

4.6.3. Discussion

Yet again, in sharp contrast to the findings reported in Buckwalter and Turri's paper, the "not obligated" option was clearly preferred, even in a case in which the consequences are severe (the death of a swimmer). Moreover, again, while the justifications of the "not obligated" responses show that these responses were consistent with the OIC principle, the justifications of "obligated" responses (9 in total) did not clearly indicate that these responses were incompatible with the OIC principle.

The great majority of "obligated" responses evinced the scope-shifting problem, in this case insisting that Jessica had a further obligation to *try to* save both swimmers:

> **"Even if she thinks and it would be physically impossible, she should still make as much of an effort as possible to try to save both swimmers."**

"She should still make an attempt to do whatever she can do."

"It is her employment obligation to at least attempt to rescue both. One at a time."

"She should at least try to save them since we don't know if she can fail or not."

"It is her duty as a lifeguard to do the best she can with what she has. Despite her being unable to rescue both people, she has to be moral enough to try to save both."

Since our scenario does not rule out the possibility that Jessica can try to save both swimmers, these justifications show that the related responses are not incompatible with the OIC principle.

Again, some participants seemed to appeal to the connection between the obligation and the nature of Jessica's social role:

"The conditions of the rescue could change however her job as a lifeguard does not change"

"She was the only one there, it was her job."

As we discussed in Study 2, these justifications may indicate real inconsistency with the OIC principle. Alternatively, similarly to what we suggested, they may indicate that, with their "obligated" response, the participants are simply emphasising the defeasible obligation that is entailed by the social role of a lifeguard, without yet accepting that the moral obligation was in force in that specific situation—that is, without rejecting the OIC principle.

Finally, although the positive correlation between blame ratings and OIC option choices was not statistically significant, the large number of "no blame" answers and low mean of blame ratings are still more consistent with the view that ordinary cognition aligns with the traditional view of the relation between blame, obligation, and wrongdoing.

4.7. Study 6: Drowning child

Studies 3, 4 and 5 featured obligations created either by the agent through a social action (a promise), or by the social role of the agent (safety worker, lifeguard). In this final study, we feature a case in which the

obligation does not come from a promise or a social role, but from the situation—a drowning child creating an obligation to help, just as in the example we considered at the beginning of Chapter 1 featuring Roy. The scenario corresponds to that in a particular condition ("recent") of Buckwalter and Turri's fifth experiment, where it was found that 88% of participants chose the "obligated, but unable" option that apparently contradicts the OIC principle.

4.7.1. Method

Participants

Participants were 41 adults (12 female, 29 male; M_{age} = 37.29; *SD* = 12.00; range = 42; 100% reporting English as their first language).

Design, Materials and Procedure

Participants first read the following story:

> **Michael is relaxing in the park near a pond when he sees a small girl fall in. She is drowning and definitely will die unless someone quickly pulls her out. This part of the park is secluded and Michael is the only person around. But Michael is stricken by a sudden *full body* paralysis. As a result, Michael is not able to save the girl.**

We used "full body paralysis" instead of Buckwalter and Turri's "leg paralysis" on the premise that this phrasing would be perceived as more of an incapacitating condition, and also as an attempt to preclude the scope-shifting problem (in a pilot study using the scenario with "leg paralysis", a participant with an "obligated" response suggested that Michael should "at least try to crawl to save the girl").[130]

The rest of the procedure was the same as in the previous studies: comprehension probe ("Michael is literally unable to save the small girl because he is completely paralyzed"); OIC probe with justification probe; blame probe. As in Study 5, there was only one OIC probe condition, with the following two options:

[130] Buckwalter and Turri's version (2015, *p*. 10) read as follows: "Michael is relaxing in the park when he sees a small girl fall into a nearby pond. She is drowning and definitely will die unless someone quickly pulls her out. This part of the park is secluded and Michael is the only person around. But Michael is stricken by a sudden paralysis in his legs. As a result, Michael is not physically able to save the girl."

1. **Under these circumstances, Michael is still obligated to save the small girl, even if he is unable to do so.**

2. **Under these circumstances, Michael is not obligated to save the small girl, because he is unable to do so.**

4.7.2. Results

Almost all participants (98%) agreed that Michael was literally unable to save the girl. The great majority of participants (73%) thought that Michael was not obligated when there was an inability to fulfil the obligation—goodness of fit against chance: χ^2 (1, 41) = 8.80, $p < .01$, $\phi = .46$.

Although "no blame" was still the modal rating (18 out of 41 participants), blame scores were noticeably higher in this study ($M = 2.73$; $SD = 2.1$). For example, a *t*-test revealed that the blame scores in Study 5 and Study 6 differed significantly, $t(61) = 2.84$, $p < .01$, $d = 0.72$ (equality of variances not assumed). Moreover, a *t*-test showed that, similarly to Study 4 (but unlike in Study 5), blame scores were significantly higher for participants choosing the "obligated" option than for those choosing the "not obligated" option: $t(39) = 5.15$, $p < .01$, $d = 1.65$ ("obligated": $M = 4.91$; $SD = 2.02$; "not obligated": $M = 1.93$; $SD = 1.48$). Finally, there was a strong, significant correlation between statement choice and blame ratings: $r_{pb} = .64$, $p < .01$.

4.7.3. Discussion

We again reversed the Buckwalter and Turri's results, although, of our studies 3-6, this one had the lowest percentage of "not obligated" responses. However, an analysis of the justifications of "obligated" responses (11 in total) suggests that this study was beset by a major problem. About half of the participants do not seem to have maintained the assumption of literal inability when answering the OIC probe, mostly because they took the full bodily paralysis to be a controllable emotional reaction (involving especially fear):

> **"He needs to overcome his fear and save the girl."**

> **"You have to overcome your fear a person's life is at stake."**

> **"It was just an emotional reaction which he could overcome."**

"Michael is responsible to get control of himself and save the girl. He can control his emotion and reactions and needs to pull himself together."

"He is responsible to save her even if he SEEMS unable to do it. I believe his perception of being paralyzed is not real."

If these justifications indeed correspond to the reason why participants chose the "obligated" response, then their responses are not inconsistent with the OIC principle after all.

Some participants' responses revealed the scope-shifting problem again in terms of obligation to try, which, as we already discussed, is not incompatible with the OIC principle:

"He is obligated to at least TRY. If he can't, he can't. Maybe the water is deep and he can't swim. But he should at least try no matter what."

"I have never heard of a sudden full body paralysis like this, and it seems like Michael should still be trying to help."

A few participants emphasised that there was a (moral) obligation in the situation:

"He had a duty to act, a moral obligation. His fear paralyzed him and he was unable to act."

"He is morally obligated to save the girl."

"Well I assume nothing has changed about the girls [sic] situation just because Michael can't move so the obligation to save her is still there, even if he can't move it still exists."

These justifications may indeed be taken to indicate a type of reasoning that is inconsistent with the OIC principle.

The fact that the overall mean of blame ratings was a bit higher in this study (in comparison with studies 4 and 5) is not incompatible with the view that inability undermines blame, since the mean was substantially

affected by the ratings of the participants with "obligated" responses that did not assume inability as shown by their justifications (with these participants eliminated from the analysis, the overall blame mean drops from "2.73" to "2.25", which is much closer to, and non-significantly different from, the overall mean of studies 4 and 5). Moreover, a large number of participants still chose the "no blame" answer. Finally, these blame ratings plus the strong correlation between blame ratings and OIC choice indicate that ordinary cognition is in line with the traditional view of the relation between blame, obligation, and wrongdoing.

4.8. Summary of results

In studies 1 and 2, we provided evidence indicating that there is a problem with Buckwalter and Turri's research design, namely that it does not unambiguously test whether people reason in line with the OIC principle. In the following four studies, using an improved design, we showed that the great majority of participants judge that a person is not under an obligation if she is not able to fulfil it, completely reversing the results reported in the original paper (see Figure 3.2). Study 3 showed that the obligation to fulfil a promise is deemed annulled when the agent is not able to fulfil it. This study also indicates that this is the case irrespective of the particular term used to express the concept of obligation ("obligated", "duty" or "ought"). Using a different scenario, Study 4 demonstrated that these results do not depend on our particular use of connectives—rather, it is Buckwalter and Turri's results that appear fragile in this respect, as also shown in Study 1. Studies 5 and 6 extended these findings to cases in which the consequences are more serious (the death of a person).

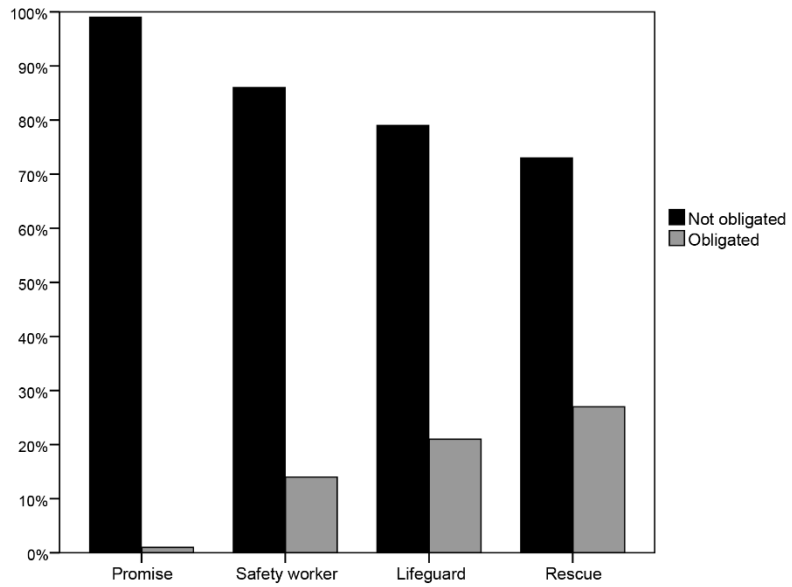

Figure 3.2. Percentage of responses to the OIC probe in studies 3, 4, 5 and 6.

Studies 4, 5, and 6 still saw a relevant minority of participants choosing the "obligated" response, suggesting that there may be some individual variation in this area. However, a substantial part of "obligated" responses still seems to derive from a misinterpretation of the OIC probe and/or the scenarios, as evinced by justifications demonstrating the "scope-shifting" problem, which appeared across all studies, by justifications showing that the participants did not keep the assumption of inability, which appeared in Study 6, and by justifications that seemed simply to emphasise the obligations normally entailed by social roles, which appeared in studies 4 and 5. Of course, if this is correct, it raises the question as to why there was such misinterpretation. The scope-shifting problem may be a result of participants' inclination to blame the agents specifically for not *trying* to do their best to minimise the bad consequences of the situation, something our studies did not control for. The misinterpretation of "full body paralysis" in terms of controllable emotional reaction in Study 6 may have a similar explanation. An interpretation of the OIC probe in terms of whether the obligations are entailed by the social role (instead of in terms of the entailed obligations being in force) may be difficult to avoid completely in contexts involving social roles, since this may always be a possible reading of the statements.

Moreover, one may raise the question of why there may have been an increase in misinterpretation between Study 3 and Study 6 correlated with the increase in "obligated" responses. There is a sense in which the consequence of the scenario in Study 6 (the death of a small girl) is worse than that of the scenario in Study 5 (the death of an adult), which in turn is worse than that of the scenario in Study 4 (the risk of stepping on a broken glass), which in turn is worse than that of the scenario in Study 3 (not being picked up at the airport). (A small study asking participants to rate these scenarios in terms of their seriousness confirmed this hierarchy—$N = 25$, Kendall's $W = .78$, $p < .01$). Thus, it is also possible that this increase in seriousness may have contributed to the increase in the amount of misinterpretation from scenario 3 to 6, by pushing participants to see the situation as less determined and hence to be more hopeful about a positive outcome.

If our take on the minority responses is correct, the range of individual variation suggested by our sample is rather small—almost all participants make judgments consistent with the OIC principle in the types of scenarios that we probed. This raises two broader and complementary issues. The first issue concerns the type of reasoning involved in participants' judgments—in particular, the type of implication connecting the concepts of obligation and ability. The second issue concerns the generalizability of our results to different types of contexts—in particular, to contexts involving culpable inability (Chituc et al., 2016; Henne et al. 2016), the former of which we discuss below and the latter of which we address in the next section.

Since implication can take many forms, the unqualified version of the OIC principle underspecifies the nature of the inferential relation between 'ought' and 'can'. In the philosophical literature, the usual candidates for this relationship are presupposition (Hare, 1963), conversational implicature (Sinnott-Armstrong, 1984), pragmatic-logical inference (Joerden, 2012), and conceptual or analytic entailment (Vranas, 2007; Zimmerman, 1996). This issue is important because each account of implication (insofar as they are understood as hypotheses about ordinary cognition) will entail different predictions about people's judgments. For instance, because on the conceptual-entailment account the inference is logically necessary, attributions of inability (to X at time Y) would preclude attributions of obligation (to X at time Y) across all types of context. In contrast, because on the conversational-implicature account the inference is defeasible, attributions of inability would preclude attributions of obligation in some contexts but not in others. The homogeneity of our results is consistent with any of these accounts—e.g. it may be that our participants reasoned in terms of conceptual entailment or it may be that they reasoned in terms of conversational implicature, but our studies were limited to contexts where the implicature is not cancelled (although see some general arguments against pragmatic accounts in the next chapter—see also King 2017). Accordingly, our results raise doubts about Buckwalter and Turri's claims, whatever interpretation of implication the authors may have in mind (Buckwalter and Turri are not explicit in their article about whether they have a specific version of the OIC principle in mind). However, some results in the literature related to contexts of culpable inability suggest that at least the traditional conceptual-entailment account is not correct, which leads us to the second part of the empirical challenge against OIC.

5. Second challenge: Chituc et al. (2016)

In order to probe whether ordinary people reject the OIC principle in terms of the conceptual-entailment account, Chituc et al. (Chituc et al. 2016) presented participants with two types of scenarios of inability. Some scenarios ("low-blame scenarios") were similar to those of our studies in that their main character did not have control (or had little control) over the source of the inability (e.g. one could not fulfil a promise because one's car broke down unexpectedly). In relation to these scenarios, Chituc et al. obtained results that were overall similar to ours (i.e. the majority of participants gave responses consistent with the OIC principle), which provides further evidence in favour of our claim that Buckwalter and Turr's conclusions are problematic.

However, the other scenarios ("high-blame scenarios") differ from those of our studies in that their main character had total control over the source of the inability (in fact, the inability was intentionally created by

the character himself). In their first experiment, for example, participants were presented with the following vignette:

> **Adams promises to meet his friend Brown for lunch at noon today. It takes Adams thirty minutes to drive from his house to the place where they plan to eat lunch together. Adams decides that he does not want to have lunch with Brown after all, so he stays at his house until eleven forty-five. Because of where he is at that time, Adams cannot meet his friend Brown at noon, as he promised.** (Chituc et al. 2016, *p*. 21)

Participants were then asked whether they agree with the statement "At eleven forty-five, it is still true that Adams ought to meet Brown at noon," which they answered by choosing a point on a scale from -50 ("completely disagree") to 50 ("completely agree"), with 0 as the midpoint ("neither agree nor disagree"). In this condition, 60% of participants provided a response inconsistent with the OIC principle (i.e. answered above the midpoint). Moreover, in their third experiment, they obtained a similar result using a different high-blame vignette (50% of participants provided OIC-inconsistent responses in this new condition). With these results (and others to be discussed below), Chituc et al. claim that the conceptual-entailment account cannot be correct.

Although our studies did not address high-blame contexts, we would like to make two comments about Chituc et al.'s related results. First, as the qualitative data of our studies show, participants are prone to misinterpreting the scenarios and/or the OIC probe in a way that renders their OIC-inconsistent responses of questionable value as evidence concerning whether they reject the OIC principle. Now, it is possible that this tendency to misinterpretation was even more accentuated in Chituc et al.'s high-blame scenarios, given that their cases of self-imposed inability are somewhat bizarre from the perspective of the protagonist's behaviour (namely, making a decision to self-impose an inability after making a promise to a *friend* without even notifying them). Thus, we believe that one has to be cautious about whether Chituc et al.'s high-blame results demonstrate that the conceptual-entailment account is incorrect (for a detailed discussion of cases of self-imposed inability from the perspective of the conceptual-entailment account, see Zimmerman, 1996, *pp*. 96-113—see also Chapter 4, especially sections 2 and 4.1).

Second, even supposing that Chituc et al.'s studies indeed reveal that ordinary people reject the OIC principle *qua* conceptual entailment, it is still plausible to suppose that there is a very stable inferential relation between the concepts of obligation and ability—i.e. that the OIC implication is a core element of the set of

inferential relations normally associated with the folk concept of obligation. Ordinary people seem to understand obligations as having a behavior-regulating role—i.e. obligations are deemed *social or moral constraints* on actions (Beller, 2008). Accordingly, it would seem rather incoherent to think that such a constraint should still be in force when it cannot be effective, namely, when the action in question cannot be carried out (see the next chapter, especially Section 1). Cases of self-imposed inability may simply constitute exceptions to this. If so, OIC needs to be elaborated to accommodate the relevant judgments. (In Section 4 of the next chapter, I discuss a different approach on which the theoretical significance of such examples is diminished.)

We turn now to the discussion of our blame results and of our perspective on how ordinary people understand the relation between blame, obligation/wrongdoing, and inability. In all our studies, a large number of participants attributed no blame to the individual for the fact that the obligation was not fulfilled. The mean blame ratings were low in all studies too. They were highest in Study 6, but this was likely due to the fact that some participants did not maintain the assumption of inability appropriately. Thus, overall, our results suggest that, for ordinary people, inability undermines blame, which is consistent with the results on blame in Buckwalter and Turri's paper. Contrary to their claim that blame attributions are unrelated to obligation attributions, the low percentage of the "obligated" responses plus the correlations between blame ratings and OIC probe choices (i.e. more blame, more "obligated" responses) in our results are consistent with our hypothesis that ordinary cognition is in line with the traditional view that, in cases of inability, blame reduction is mediated by obligation/wrongdoing elimination.

However, there is another set of results in Chituc et al. (2016) that apparently goes against our perspective on how ordinary people understand the relation between blame, obligation/wrongdoing, and inability—indeed, these results apparently go even against the aforementioned hypothesis that, even if not analytical, the OIC implication is a core element of the set of inferential relations normally associated with the operative concept of obligation. (This is something that is not explicit in Chituc et al.'s discussion: while some of their results, as discussed above, go against the conceptual-entailment account of the OIC implication but not necessarily against other accounts, some of their results, to be discussed next, go against a much broader range of accounts.)

In their second experiment, Chituc included only a low-blame scenario of inability (in this scenario, the character cannot keep the promise to meet with his colleague at noon because his car unexpectedly breaks down). They asked participants how much they agreed with statements saying that the character ought to keep the promise, is to blame for not keeping the promise, and can keep the promise (the same agreement

scale was used, as explained before in relation to their first experiment). Restricting the analysis to participants who disagreed with the "can" statement, since these are the relevant cases for our discussion, Chituc et al. found a correlation between blame and obligation responses (r = .24, p < .01), but while they found a correlation between blame and ability responses (r = .24, p < .01), they did not find a correlation between obligation and ability responses (r = .07, p = .37). This suggests that when people give OIC-consistent responses, they are simply engaging in excuse validation (Turri & Blouw, 2015)—that is, they are denying obligation to be consistent with a primary reduction in blame attribution based on the situation of inability, rather than because of an inferential relation between the concepts of obligation and ability. In other words, it suggests that the relation between obligation and ability is completely mediated by blame attributions.

However, the above pattern of correlations was not replicated in their Experiment 3 in the context of its moral/unable conditions, since they found no correlation between blame and ability responses while observing a trend (r = .18, p = .09) between obligation and ability responses.[131] Furthermore, we carried out further analyses of the results of Experiment 3 (based on Supplementary Data S4 available at the publisher's website), showing that the relevant correlations go in the direction of our picture. (The following correlations were not reported in the original article.) If one restricts the analysis to participants who disagreed with the "can" statement—thus including only those subjects whose responses are mostly relevant to our discussion—, one finds that there is a correlation between blame and obligation responses (r = .40, p < .01), but while there is still no correlation between blame and ability responses (r = .03, p = .79), there *is* a correlation between obligation and ability responses (r = .32, p < .01). Thus, although we acknowledge that this is still a contentious issue, and that it is still possible that our results were prompted by an excuse-validation bias, we believe that our overall picture on the relation between blame, obligation/wrongdoing and inability remains more plausible.

6. Conclusion

To conclude, our studies provide strong evidence that despite Buckwalter and Turri's claims to the contrary, people do make judgments largely compatible with the OIC principle, at least certainly in cases in which the inability is not self-imposed. Furthermore, although we acknowledge that this question is far from

[131] It is worth noting that the non-moral conditions of Chituc et al.'s Experiment 3 are completely irrelevant to our issue here, since these conditions do not involve non-moral obligations as Chituc et al. appear to claim; rather they involve what is discretionary—a decision to go to the cinema does not involve a non-moral obligation or an obligation of any description, it simply involves what is under someone's discretion.

settled, we believe that our results are best explained by maintaining that there exists a strong inferential relation between the concepts of obligation and ability in ordinary cognition. Consequently, the empirical results discussed in this chapter do not cast serious doubt on OIC is a descriptively adequate principle of I-morality or FM. Finally, our results are also consistent with the idea that ordinary reasoning is in line with the traditional view that blame reduction is related to obligation elimination via the elimination of wrongdoing. In the next chapter, I consider some of the consequences of taking seriously the main conclusion of this chapter, namely, that OIC is a descriptively adequate principle or moral cognition.

Chapter 4: OIC Meets Cognitive Science—Hypotheses, Old and New

Overview

In this final chapter, I review the ways in which the descriptive adequacy of OIC has been explained in the literature, criticise these hypotheses and propose some of my own based on LA. I proceed as follows. In Section 1, I review some of the ways in which philosophers have rationalised OIC. Aspects of these accounts will be reflected in the hypotheses to be discussed subsequently. In Section 2, I introduce and criticise the hypothesis according to which OIC captures a semantic relation between the lexical concepts OUGHT and CAN. In Section 3, I discuss and reject the view according to which the descriptive adequacy of OIC is due to the pragmatics of communication. In Section 4, I first propose a novel explanation of why such semantic and pragmatic accounts of OIC are deemed to failure, then, I illustrate the way out of this conundrum by proposing some hypothesis sketches of my own. On my favoured hypotheses, OIC's descriptive adequacy is a consequence of the operations of FM. These are not expected to be the final words regarding the question of how to explain OIC's descriptive success, but they do provide a novel conceptualisation of the debate.

1. Three rationales for OIC

In the previous chapter, the central argument was that OIC is descriptively adequate, that is, it accurately describes, generates, and thus predicts ordinary judgments of the relevant type. It is time to consider how OIC is implemented in the human mind, that is, what it is at the level of psychological explanation (*cf*. Chapter 1, sections 4 and 6) that accounts for the descriptive adequacy of the principle. As I did in Chapter 3, I shall proceed from moral philosophy to moral psychology. That is, I will introduce some of the ways in which ethicists have interpreted OIC: (i) as a logical or conceptual principle, (ii) as a normative/moral principle, and (iii) as a principle of practical rationality. In later sections, I revisit these interpretations and translate them into the language of the LA framework.

To begin with, we have a conceptual interpretation of OIC (Haji, 2002; Vranas, 2007; Zimmerman, 1996). According to it, OIC constitutes a conceptual truth, with "S ought to ϕ" entailing "S can ϕ".[132] On this view, contradicting OIC amounts to a sort of "conceptual" confusion.[133] There is some plausibility to this. After all, as pointed out above, when ethicists say that Walter ought to pick up his friend, what they mean is that Walter has a moral obligation to do so (some argue, plausibly in my view, that *must* expresses this relation better than *ought*—McNamara, 1996; von Fintel & Iatridou, 2008). Now obligations (whether moral or otherwise) are often understood both in the philosophical literature (Zimmerman, 1996) and in ordinary thought (Jackendoff, 1999) as constraints on what actions/options are available to the agent, given a certain code of conduct (e.g. morality, the law, and so on). If so, what sense exactly could we possibly make of a requirement that *constrains* an *unavailable* action?

Another way of conceiving of OIC appeals to the notion of fairness (e.g. Copp, 2008; Fischer, 1999, 2003; for criticism, see van Someren Greve, 2014).[134] According to this view, OIC is a moral principle grounded in the relation between obligation and blame. Recall Walter's predicament from the previous chapter. Imagine that Walter's wife and lifelong critic Susan insisted that Walter is *still* under obligation to deliver his friend from the airport in spite of knowing full well that this has now been rendered impossible: let us assume that both of Walter's legs are broken and he's currently in a coma. Wouldn't this be inconsiderate and unfair of Susan? Suppose that Walter still has the obligation to pick up his friend. Violating an obligation constitutes wrongdoing and warrants blame. Since we already know that Walter will not fulfil this obligation, the persistence of it would necessitate Walter's future wrongdoing in spite of the clear sense that he has done nothing wrong. Thus, by insisting on the obligation, Susan would be indirectly blaming Walter for something that is clearly beyond his control, which, on the face of it, seems completely unfair towards him.

---

[132] Where S and ϕ are placeholders for an agent and an action, respectively. With this formula, we also assume that the referents of *ought* and *can* are fixed (Kurthy and Lawford-Smith, 2015)—they are usually taken to refer to a *moral obligation* and some form of *ability* or *physical possibility* (see more on this in Section 2 below).

[133] Kant also defended a similar view: "duty commands nothing but what we can do [...] For if the moral law commands that we ought to be better human beings now, it inescapably follows that we must be *capable* of being better human beings" (Kant, 1793, *pp*. 47-50; original emphasis—though see Stern, 2004).

[134] *Copp*: "It would be unfair to expect a person to do something, or to require that she do it, if she cannot do it. Similarly, morality would be unfair if it allowed that a person might be [...] morally required to do something that she cannot do" (2008, *p*. 71). *Fischer*: "what justification could be offered for [OIC]? It is most natural, I think, to say that [OIC] is valid because if it were not, then there could be cases in which an agent ought to do X but in fact cannot do X (and never could do X). Thus, given the connection between its being the case that an agent ought to do X and the agent's being blameworthy for not doing X, there could be cases in which an agent is blameworthy for not X-ing and yet he cannot X. And this seems unfair" (1999, *p*. 124). Note, however, that Fischer subsequently rejects OIC, although he does so somewhat hesitantly.

Third, one might read OIC as a systemisation of rational deliberation. This view puts the emphasis on the realisation that insisting on the obligation in cases such as Walter's would be entirely *pointless*, even irrational (see e.g. Sapontzis, 1991; Griffin, 1992; Joerden, 2012)[135]—unless of course Susan's intention is to blame Walter (see above). Picture Susan standing beside Walter's bed in the hospital, telling Walter (who is unconscious, remember) that he *ought to/must* go and pick Brown up from the airport. Her behaviour could be fairly described as completely futile and irrational—on the reasonable assumption that her goal in insisting on Walter's obligation is to make sure that Brown will not be stranded at the airport. In any case, Susan certainly seems to be wasting her time: she could go and pick up Brown herself, she could call a cab for him, she could give him a call to let him know Walter won't be able to get him, or she could go home and organise her stamp collection. According to this latter interpretation, OIC is thus a principle of practical rationality.[136]

Let me explain why these rationales for OIC are relevant in the context of the present project. Having argued in the previous chapter in favour of the descriptive adequacy of OIC, given the LA framework expounded and defended in the first half ot the thesis, the problem at hand is to find some ways to account for the results in terms of our moral competence. The first step in this direction is the generation of a hypothesis space: a space of possibilities as to the psychological facts (mechanisms, representations and processes) that account for OIC's descriptive adequacy. The above proposals indirectly offer some preliminary ways of creating such a hypothesis space. Why do we reason and make judgments in line with OIC? Perhaps because our concepts force us to. Or perhaps because we deem it fair to do so. Or perhaps we recognise that it is the rational thing to do. These accounts begin to look like psychological accounts of the descriptive adequacy of OIC. However, they are not detailed enough to be of sufficient use relative to the

[135] *Sapontzis*: "it would [...] be pointless to hold that moral agents are obligated to do things they are constitutionally incapable of doing" (1991, *p.* 391). *Griffin*: "Action-guiding principles must fit human capacities, or they become strange in a damaging way: pointless" (1992, *p.* 123). *Joerden*: "Whoever commands a person P to perform a certain act must assume that P *can* actually perform this act. Otherwise, the commander finds himself in a pragmatic self-contradiction that can be described in short with the following clause: "I know that you *cannot* lift this rock, nonetheless, I command that you lift it" (2012, *p.* 205, emphasis original).

[136] Beyond such considerations concerning fairness and practical reasoning, the OIC principle is significant in philosophical theorising for a variety of other reasons. I will only mention two. First, it has been used to argue against the possibility of genuine moral dilemmas. In a nutshell, since performing two incompatible actions is impossible, one cannot be obligated to perform both (see Mason, 1996). Second, it has been held to have repercussions for the correct formulation of consequentialism/utilitarianism (Mason, 2003). Roughly, it may be argued that any form of consequentialism that requires us to do what is beyond our capabilities is in violation of OIC and is or should be discarded. Suppose it is psychologically impossible to abandon our families and promote the welfare of those the improvement of whose condition has a larger impact on overall well-being. According to the argument under consideration, a version of consequentialism entailing such obligations is mistaken.

standards of LA. In the following sections, these interpretations will resurface in the guise of somewhat more elaborate hypotheses.

In the next two sections, I review and reject the two dominant accounts of the descriptive success of OIC. I argue that given their obvious shortcomings, other accounts are should be offered, namely, ones that take the import of the first half of the thesis seriously. I set to this task in Section 4.

2. The Semantic Hypothesis (SH)

The most general, least computationally or mechanistically committed, and arguably boldest hypothesis about the psychological underpinnings of OIC is what Leben (2018) refers to as the *Semantic Hypothesis* (henceforth, SH). On this account, OIC captures a meaning relation between the lexical items 'ought' and 'can'. More specifically, this relation is understood to be that of entailment. Consider sentences (1) and (2) below: (1) is generally understood to entail (2).

(1) Dimebag was murdered

(2) Dimebag is dead

On the prevailing, truth-based definition of entailment, *p* entails *q* ('p' and 'q' being sentential variables), if and only if the truth of *p* necessitates the truth of *q*. In the example above, if (1) is true, then so must (2) be. Thus, because, as specified in the definition, and as also illustrated by the example, entailment is generally seen as a relation between sentences, the way in which to frame OIC as per SH is something like this: the sentence *type* '*S* ought to ϕ' entails the sentence *type* '*S* can ϕ'.

This formulation of OIC is in fact overly simplistic, since OIC is often taken to concern the relation between a moral obligation and ability (plus opportunity) (Vranas, 2007), yet *ought* and *can* admit of many other interpretations, depending upon context. Whether this variation constitutes a case of true polysemy or not is debated in the literature (*cf*. Kratzer, 1977), but that some *ought*s do *not* entail *can* is not controversial. For instance, when one says "children ought not to suffer", it does not seem to be the case that the person is attributing an obligation to anyone, and indeed the OIC inference is not customarily made in such situations: it is perfectly OK to insist that children ought not to suffer even if we accept that their suffering is inevitable (for whatever reason). Such uses of 'ought' are referred to as *evaluative* as opposed to the *deliberative* use we are after, such as when 'ought' refers to an obligation. The hypothesis should be then formulated as follows: the sentence type '*S* ought_{D} to ϕ' entails the sentence type '*S* can_{A} ϕ'—where 'ought_{D}' and 'can_{A}' signal that these two modals receive a deliberative-moral-deontic and ability reading, respectively. Or, equivalently: the

sentence type 'S has the moral obligation to ϕ' entails the sentence type 'S has the ability (and the opportunity) to ϕ'.

It is already worth paying attention to the fact that although such hypotheses of semantic entailment are not necessarily psychologically "innocent", they are nevertheless psychologically underspecified. That is, when we agree that (1) entails (2), for example, we had better assume that there is *something* about the lexical concept MURDER(*X*,*Y*)—where *X* and *Y* are argument places for the murderer and the murdered, respectively—that makes it the case that its instantiation licences the inference: MURDER(X,Y) → DEAD(Y). Thus, we cannot think "Dimebag was murdered" without thinking—or at least being disposed to think—that "Dimebag is dead". However, beyond the claim that such inferences are in fact made, the semantic *qua* conceptual hypothesis has little to say about what makes this the case from the point of view of psychological systems and processes. Similarly, although we may hypothesise (as in fact Jackendoff, 1999 does) that from $\text{OUGHT}_{\text{D}}(S,\phi)$ licences the inference $\text{CAN}_{\text{A}}(S,\phi)$, this is hardly more than a mere restatement of the semantic relation (defined at the level of sentences) at the lexical-conceptual level (defined in terms of conceptual relations).[137]

2.1. Against SH

SH suffers from a very serious *prima facie* shortcoming, namely that it fails what I like to refer to as the "bachelor test", that is, a standard test for semantic entailment (*cf*. Henne et al. 2016). Consider (3) and (4):

(3) John is a bachelor

(4) John is married

Since (3) entails the negation of (4), competent speakers of English would deem the assertion of (3) incompatible with that of (4). Thus, (3) and (4) are contradictory.

Now consider (5) and (6):

(5) Walter ought to pick up Brown

(6) Walter cannot pick up Brown

[137] This is not the way Jackendoff states the inference. Instead, he asserts that the expectation that an agent is able to carry out an action that is deemed obligatory derives from a constraint on the first argument place of OUGHT_{D} (i.e. what we might refer to as the *ability constraint*). It is worth noticing, however, that this is not so much a novel hypothesis about OIC as the reformulation of conceptual OIC in terms of conceptual argument constraints.

On the face of it, (5) and (6) do not have the status of contradiction: it seems more or less all right to say "Walter ought to pick up Brown, but he cannot" (and this is not a matter of lexical choice either: substituting "ought to" for "is obligated to" or "must" does not seem to make much of a difference in this respect).[138] Granted, this much is based on pure introspection, but recall that towards the end of the previous chapter (Chapter 3, Section 5), I suggested that Chituc et al.'s data (Chituc et al., 2016) can be seen as a significant empirical challenge against such semantic entailment theories (Henne et al., 2016). Their data shows that, in cases of self-imposed inability, not only is the conjunction of sentences like (5) and (6) not seen as contradictory, such statements are in fact endorsed simultaneously.

This is not to say that SH is thereby *refuted*. Indeed, there are ingenious defences of the semantic-conceptual view in the literature, for example in terms of an elaboration of OIC specifying the referents of temporal indices of the modals 'ought' and 'can' (Zimmerman, 1996, *pp*. 95-113; see also Streumer, 2003). Peter Vranas also makes the point that the fact that a contradiction follows the conjunction of sentences such as (5) and (6) does not entail that the contradiction is psychologically transparent: we might need to think deeply to realise that there is a contradiction.[139] Although such arguments may have some merit, this does not change the fact that OIC does not fit the prototypical entailment model comfortably. Let me put it this way: although sentence pairs such as (3)-(4) are customarily used in semantics textbooks as illustrations of semantic entailment, it may be safe to assume that sentence pairs such as (5)-(6) will never be. This suggests that we might have to look elsewhere for an explanation of the descriptive adequacy for OIC.

3. The Pragmatic Hypothesis (PH)

It is typically assumed that if SH is rejected then some version of what Leben refers to as the *Pragmatic Hypothesis* (or PH for short) should be adopted. The starting point of (all versions of) PH is that, as purportedly illustrated by the lack of contradiction in judging both (5) and (6) true at the same time (or in asserting them simultaneously), there seems to be no semantic entailment relation between the terms 'ought'

[138] Leben makes the point that although "it sounds normal to say, 'I ought to help, but I can't' [...], it sounds odd to say, 'I can't help out, but I ought to'" (2018, p. 161), suggesting that the lack of incompatibility with respect to (5) and (6) might be due to the fact that, if they presented in succession, the 'ought' in (5) is understood to refer to a so-called prima facie (as opposed to an overall or all-things-considered) obligation, or it may even be read as an evaluative rather than a deliberative 'ought' (see above). Although the argument has some merit, this does not change the fact that OIC does not fit the prototypical entailment model comfortably.

[139] "Conceptual entailment need not be transparent: it may take some thought to realize that 'A is sufficient for a condition which is necessary for B' implies 'B is sufficient for a condition which is necessary for A'" (Vranas, 2007, p. 170).

and 'can'. Thus, proponents of PH typically hold that SH is too strong as an account of OIC.[140] Nevertheless, as a testament to the intuitive call of OIC, they maintain that there is *some* kind of inferential link between the terms involved (see the previous footnote), but this link, they contend, is one that is relative to communicative contexts. Roughly, the idea is that a statement of the form:

(7) *S* ought/is obligated to ϕ

is often taken to carry an "assumption" (shared by the interlocutors, at least in successful communicative exchanges) to the effect that the agent (*S*) in fact *can* perform the action (ϕ); this could be schematically represented as:

(8) *S* can ϕ

As mentioned above, there are two main versions of PH. They disagree over what the nature of this assumption is, that is, they have different ideas as to the nature of the pragmatic inference between the 'ought' and the 'can' sentence. On the presupposition view, (7) presupposes (8), and (6) presupposes (7), in the same way as (9) presupposes (10):[141]

(9) The King of France is bald

(10) There is a King of France

On the other main version of PH, the relation between (7) and (8) is understood as (generalised) conversational implicature.[142] To borrow from Sinnott-Armstrong, "saying *p* conversationally implies *q* when saying *p* for a certain purpose cannot be explained except by supposing that the speaker thinks that *q* and thinks that the hearer can figure out that the speaker thinks that *q*" (1984, *p*. 256). Thus, for instance, (11) conversationally implicates (12), because, although (12) is not entailed by (11),[143] uttering (11) would be odd and unhelpful if the speaker did not think (12) also to be true, which is something both speakers are aware of.[144]

---

[140] As also noted by Leben (2018), versions of PH are typically understood not so much as versions as denials of OIC (in the form of SH). However, strictly speaking, versions of PH are explications of the link between 'ought' and 'can', and, as such, they are properly understood as accounts of OIC.

[141] Besch (2011), Cooper (1966), Driver (2011), Hampshire (1951), Hare (1951, 1963), Martin (2009). The originators of this view are Hampshire (1951) and Hare (1951).

[142] Forrester (1989), Littlejohn (2009), Oppenheim (1987), Saka (2000), Sinnott-Armstrong (1984), Turri (2017), Vallentyne (1989), Vogelstein (2012). The originator of this view is Sinnott-Armstrong (1984).

[143] For example, (11) is consistent with the negation of (12).

[144] The relation between (11) and (12) is a *generalised* implicature, because it does not depend on specific aspects of the context. This contrasts with *particularised* implicatures, in which special reliance on the context is necessary. Consider, for example, the following exchange: Speaker 1: "Is Frank coming to the dinner?" Speaker 2: "He's ill". Speaker

(11) John has two brothers

(12) John does not have more than two brothers

On this view, the relation between 'ought' and 'can' statements is tied to the purpose of the 'ought' statement.[145] Sinnott-Armstrong postulates three main uses of 'ought' statements: (a) advising, (b) blaming, and (c) deliberation.[146] Sinnott-Armstrong's point is that only in contexts in which an 'ought' sentence (or statement) is used for the purpose of advising is a corresponding 'can' sentence conversationally implicated.

This view is attractive at first sight because not only does it account for the failure of the entailment view, but it also systematically predicts the contexts in which the implication is not made. For example, in contexts in which an agent is culpably responsible for his or her inability to perform an action, it seems all right to insist that the agent has an obligation to perform it in spite of the inability (*cf*. Chapter 3, Section 5).

3.1. Against PH

Just like in the case of SH, both versions of PH are deeply problematic. King (2017) mounts an incisive attack on all available pragmatic accounts of OIC, showing that 'ought' and 'can' sentences fail both the paradigmatic test for presupposition—i.e., the constancy under negation test (King, 2017, *pp*. 4-13)—as well as (*pace* Sinnott-Armstrong) the calculability and cancellability tests for conversational implicature (ibid., *pp*. 13-20).

But the crucial point from our perspective is that, as King observes, such accounts fail to capture the fact that OIC is not first and foremost a communicative principle, and it is (or seems to be) operative in contexts not involving communicative exchanges of any kind at all, such as in private deliberation. Consider the following case:

> Suppose Nicole is teaching a discussion-based class. Everyone in the class speaks up a lot, except one very quiet student. She thinks to herself, this student really ought to speak up more. He does himself and everyone else a disservice by not contributing! Not only that, it seems like he's always passing notes with the student sitting next to him. How incredibly rude! She eventually discovers that this student is mute. Those notes? That was him occasionally asking comprehension questions to the

2's utterance implicates that Frank is not coming, but it wouldn't obviously do so without reference to the particular communicative context (namely, Speaker 1's inquiry concerning Frank's involvement).

[145] Which, I believe, renders somewhat problematic the idea that the implicature is generalised (see the previous footnote). I shall not be concerned with this potential difficulty, however.

[146] Though, as noted by King, Sinnott-Armstrong's use of the latter term is somewhat non-standard, referring to something like hypothetical reasoning.

> student sitting next to him. Of course she no longer thinks he ought to speak up more. In fact, she is mortified that she ever thought he ought to. (King, 2017, *pp*. 20-21)[147]

Now, it seems exceedingly far-fetched to explain such inferences as being due to some unspecified conversational principles, and the prominent pragmatic accounts are certainly powerless to provide much help.

The third serious problem versions of PH are confronted with is that of making sense of the all important contraposed version of the OIC principle—that is, "Cannot Implies Not Ought" (see also Kurthy and Lawford-Smith, 2015; Southwood, 2016, *p*. 70, *fn*. 2). Assume that OIC is accounted for in terms of conversational implicature. The main issue is this: if "S ought to ϕ" conversationally implicates "S can ϕ", then it is unclear why one *can* infer "it's not the case that *S* ought to ϕ" from "*S* cannot ϕ". King's example is "I went to a birthday party yesterday", which conversationally implicates that the speaker did not go to her own birthday party (King, 2017, *pp*. 17-19). It is easy to see that one cannot infer "I did not go to a birthday party yesterday" from "I went to my own birthday party yesterday". More generally, pragmatic inferences do not support contraposition, only conceptual or logical entailment does.[148]

4. OIC and FM

It seems that we are facing a conundrum. In the literature on the OIC principle, there are two major options. Some interpret OIC *qua* SH, while others endorse some version of PH. However, neither SH nor (either version of) PH appears very plausible on closer inspection: it seems that OIC is neither best understood as a semantic nor as a pragmatic principle. At least *prima facie*, this indicates that something has gone seriously wrong in the OIC literature, and that some other interpretation is called for. In this section, I first provide an analysis of why such a situation might have emerged, then, I put forward an alternative interpretation of OIC that avoids the problems with the extant accounts.

[147] King's scenario is in terms of "ought", but it is easy to paraphrase it as one involving (non-contractual) moral obligations—for instance by adding the assumption that unless children in Nicole's class verbally contribute to the discussion, newborn puppies will be killed by an evil wizard.

[148] For instance, "S is a bachelor" entails "S is not married" and, contrapositively, "S is married" entails "S is not a bachelor". And of course "if *p* then *q*" is famously logically equivalent to "if not *q* then not *p*".

4.1. Problems with extant accounts of OIC

Let us consider SH first. The "bachelor test" argument (Section 2.1) is what we might refer to as an internal critique: it suggests that SH fails on its own terms. But there are some external reasons to question the very idea that SH can provide the right *type* of account for OIC's descriptive success. To wit, the problem here is that, as an explanation, SH does not interface too well with the study of moral cognition, especially as pursued within an LA framework. As mentioned in Chapter 1, the empirical evidence supports the existence of a moral faculty (or FM); that is, a specialisation for moral judgment—as indicated by its ontogenetic trajectory, automaticity and effortlessness, as well as other types of evidence. Our relevant lexical concepts are likely to tap into (the outputs of) FM, but the nature of this link is far from being well understood. Take, for instance, the argument is Section 2.1, namely that OIC fails the semantic entailment test. The current point is that *if* the descriptive adequacy of OIC is due to the properties of FM or I-morality, then there is no strong reason to believe that it shouldn't fail it, that is, that OIC should have a semantic counterpart.

As so often has been the case in this thesis, an analogy from language helps illuminate the problem at hand. We considered some of the (operative) representations linguists have posited to explain aspects of I-language, such as its syntax, which included representations of phrase types, such as NP or CP (e.g. Chapter 1, Section 3.2.1).[149] It is quite obvious that, as laymen, we have no intuitions about the semantic relations the *concept* NP enters into (assuming we have one); perhaps it enters into none. If that is how we use the term *concept*, and entailment is understood as due to a certain relation between concepts (*qua* mental representations), it is far from obvious whether semantic relations, such as entailment, will be of any interest at all from the point of view of the study of I-morality. The operative representation of NP evidently enters into (non-semantic) relations with other representations and principles hosted by the language faculty, namely, syntactic ones (such as the rewrite rules mentioned in Chapter 1). There is no reason whatsoever to expect there to be express semantic relations mirroring these operative ones.

As another example, compare the case of moral-deontic reasoning to the case of depth perception. The problem the visual system faces is that of producing a 3-dimensional representation of the world on the basis of 2-dimensional retinal images. Stereopsis is the disparity between the retinal images of the two eyes and it is one of the cues the visual system exploits to solve this problem (i.e. to compute depth). Humans are capable

[149] As also mentioned before, linguists need to posit the existence of technical terms such as 'NP', because without doing so, they would be powerless to explain fundamental aspects of our language competence. The best explanation of the utility of positing such terms of art is that there is something like a *functionally equivalent* representation operative in the mind. See e.g. Jackendoff: "It is obvious that speakers don't have a direct counterpart of the symbol NP in their heads. Rather, what is significant about the symbol is only that it differs from the other syntactic categories, not how it is labeled" (2002, *p.* 24).

of estimating depth relying on this cue alone (Julesz, 2006), which means that—although depth perception is much more complex involving many separate cues, such as motion parallax or oculomotor cues—information about binocular disparity is input to a mechanism which computes a 3-dimensional "image" partly on the basis of information related to stereopsis. Now, the point is that any statement about the discrepancy of retinal images will be *semantically* compatible with any statement about the 3-dimensional nature of an object (*cf*. e.g. "there's no binocular disparity, but the object is/seems 3-dimensional"). That is because, although we may have concepts about such things as stereopsis, binocular disparity or depth (*cf*. NP), those concepts do not constitute the information/representations over which the computations of the (relevant part of the) visual system are defined. Rather, they are concepts *about* the information that enters or leaves it.

As indicated by the empirical evidence mentioned earlier, our competence with reasoning in terms of moral obligations (and other moral-deontic representations) indicates a broadly parallel situation in the case of moral cognition.[150] This competence may not stem from semantic entailments between such concepts as OUGHT_{D} (or OBLIGATION)—again, were the term 'concept' is understood along the lines indicated above. Thus, it may be that we do in fact judge an agent as unable to perform an action as not morally obligated to perform it (as suggested by the evidence presented in the previous chapter), and yet find nothing semantically wrong with a sentence that states the opposite. This is even consistent with an *overall* judgment according to which the agent is morally obligated to perform the action, because what we are hypothesising about are representations output by a specialised faculty, which might be consciously accessed to various degrees or even overwritten by processes downstream it—for instance, by a reasoning system à la Greene (2013) or central cognition à la Fodor (1983, 2000). Indeed, the blame validation hypothesis with respect to OIC-consistent reasoning (according to which subject might assert an obligation in some cases to indirectly blame them—see Chituc et al. 2016) is most naturally understood in such a way.[151]

Nevertheless, we also have to be careful to avoid seeing the parallel between language (or depth perception) and moral cognition as too strong in this instance. To wit, drawing the distinction between operative and express representations in the case of language does not appear at all problematic: syntactic representations

[150] Of course, there is the thorny issue of whether such competence is accurately characterised as *moral*—after all, the human adeptness with deontic reasoning is not limited to moral contexts, however one delineates what exactly the latter involves. On the other hand moral-deontic reasoning does not obviously reduce to "pure" deontic reasoning either synchronically or ontogenetically. I will ignore this issue here (but see also Chapter 2, sections 4.2 and 4.5), because it is not easily resolved and would lead us too far off the track.

[151] This is the converse of the excuse validation hypothesis mentioned in the previous chapter, which asserts that the reason why we *deny* the obligation in non-culpable situations is because we wish to indirectly excuse the relevant agent, so cases in which OIC holds are not due to competence factors.

simply do not feature in our practical reasoning, or indeed in any other type of quotidian reasoning.[152] Thus representations of syntactic phrase types are purely operative (and *mutatis mutandis* for depth perception). By contrast, the very reason why moral considerations may enter into our deliberations is presumably to do with the fact that I-morality (narrowly defined) interfaces with action planning in some ways (*cf*. Sterelny, 2010). Otherwise, judging actions to be morally obligatory or forbidden would have no consequences for our actions, which is clearly not the case.[153] Although we did draw a distinction between perception and production in the first chapter (Section 3.4), clearly, the connection (just as in the case of I-language) must be non-zero.

Still, the point remains that the relation between I-morality or FM, on the one hand, and lexical concepts on the other is so ill-understood that relations such as semantic entailment should be treated with caution in the context of the study of moral cognition. For example, even if, as it might be, in talking about moral obligations and prohibitions (that is, when we token the lexical concepts OBLIGATION and PROHIBITION), we rely directly on the representations the principles and operations of the FM are defined over, it is not clear that OIC must be (reflected in) a semantic relation.[154] This is because it does not follow from this hypothesis that in ordinary reasoning, we access all operative aspects of these representations as well as all their causal-inferential roles (where 'inferential' is understood loosely).[155] Thus, once again, the usefulness of semantics for the study of moral cognition (and *ipso facto* for the study of OIC as a principle of I-morality) must not be assumed or overstated.[156]

Given the rather extensive discussion above, we can afford to give short shrift to OIC *qua* PH. After all, if semantic relations are not expected to mirror the workings of a domain specific faculty (as I argued above), it is even less apparent why we should expect communicative principles do the same. More generally, the idea

[152] For one, I have no qualms with referring to syntactic processing as a form of reasoning: this is one of those issues that are called "purely semantic" in ordinary discussions. The point is that even if syntactic processing is a form of reasoning, it is not the ordinary type of reasoning we pre-theoretically refer to as such.

[153] As an aside, I mention that in our earlier understanding of psychopathy, the idea was that it is psychopaths' moral competence that is damaged or missing—both perception and production-wise (e.g. Blair, 1995). More recently, evidence has indicated that the problem may be with the interface between judgment and practical reasoning (*cf.* e.g. Aharoni et al. 2014; Cima et al. 2010).

[154] This possibility is not out of the question. Moral-deontic reasoning is universal, and so are concepts such as OBLIGATION or PERMISSION (as documented by Brown, 1991). This contrasts with the *concept* NP, which is very far from universal, as opposed to the operative syntactic representation, which is universal.

[155] Yet another way of making the same point is to say that there is no guarantee that lexical concepts about the information presumed to be available to (or computed by) a putative cognitive system will provide us with a good characterisation of the computational properties of that system. Put this way, the point appears uncontroversial as far as contemporary (philosophy of) cognitive science is concerned.

[156] Although this point might appear trivial (especially with the help of LA), rarely does one see it asserted or even mentioned in the moral psychology literature.

that OIC be understood as a principle of language or (even worse) language use is an unwarranted assumption that has nevertheless been taken for granted in the literature on OIC. I suggest that this assumption should be suspended.

4.2. New hypotheses

As we have just seen, in the debate on OIC, theorists have proposed to explain the OIC principle either in terms of semantic relations among lexical concepts, or in terms of the pragmatics of language use. Both of these proposals fail. Instead, motivated by the theoretical approach defended in the first half of this thesis, I propose to look elsewhere. In this section, I propose an account of OIC according to which the descriptive adequacy of this principle is due to the operations of I-morality or FM. As suggested by previous discussion (Chapter 1), there are (at least) two general versions of this account. First, OIC may be understood as a processing principle, governing the synchronic operation of FM (as we shall see, there are different versions of this hypothesis, too). Second, OIC may be understood an acquisition principle, governing the diachronic functioning of FM. Currently, evidence cannot obviously adjudicate between such proposals, but they do fare better in terms of the explanation of the descriptive adequacy of OIC than any of the extant accounts.

4.2.1. OIC as a processing principle (descriptive adequacy)

Let us first consider the hypothesis that OIC is a processing principle. Since I distinguished between three different types of processing principles, there are three potential versions of this hypothesis. However, for pragmatic reasons (because it seems *prima facie* implausible), I ignore the possibility that OIC is a conflict principle. Thus, the two other remaining options are either that (i) OIC is a constraint principle (regulating input to FM and shaping the extent and kind of its domain specificity), or that (ii) OIC is an faculty principle, contained "within" FM, regulating the output of FM.

In this first case, we would be assuming something along the following lines: the mind computes something functionally equivalent to (the contrapositive version of) OIC as a constraint on obligation attribution so that whenever an agent (*S*) is *understood*[157] to be unable to carry out an action (φ), or whenever φ is deemed impossible for *S*, the output is something like '*φ is not obligatory*'. Thus, it may be that moral-deontic status is always assigned to action representations irrespective of whether or not they are represented as impossible.

[157] The italics are to indicate that what matters is how our idealised subject *mentally represents* the situation, rather than how the situation *is* as a matter of fact. (Contrast this with some other accounts of OIC that are concerned with whether or not the agent is *actually* unable to perform the action, see e.g. Zimmerman, 1996).

Second, it may be that OIC is a *constraint principle* on what kind of representations moral-deontic computations are performed on. This may be illustrated by assuming that the action representation is indexed with a binary variable representing the relevant value of the possible/impossible (able/unable) dichotomy. If the action is indexed with the value "impossible", then deontic computations are not performed over it—in this case, the result would be something like "deontically unvalenced" or "not deontically determined" (or simply, "no output") rather than "not obligatory".[158] In contrast, on the previous model, OIC is a principle *of* FM: in that case, to stick to the tag metaphor, FM computes "not obligatory" on all representations with the impossibility tag (of the requisite kind), rather than such representations not satisfying the input constraints of FM.[159]

Such accounts can be interpreted as a substantial hypotheses about FM/I-morality, or of how FM outputs OIC-consistent intuitions in the form of the assignment of moral-deontic status to action representations. That is, they both assume that the computation of impossibility is prior to or simultaneous with the assignment of deontic status to action representations. The first can be understood as specifying the representations on which *deontic* computations (i.e. the computations of FM) are performed. That is, the first proposal is an elaboration of our model of FM, and as such, we may refer to it as an "intra-faculty" principle. Meanwhile, the second hypothesis can be understood as a "cross-modal" or "inter-faculty" principle, that is, as an informational constraint principle on the input to FM.

Yet another possibility is that the computations of FM are independent of modal concerns, such as whether something is possible: FM computes representations of, say, the logical form 'O(S,ϕ) at t_1',[160] but in some cases, the outputs are blocked or overwritten by some other reasoning mechanisms. For instance, it is rather straightforward to interpret the intuition behind the fairness-based defence of OIC in such a way. To wit, the moral faculty does its proprietary computations and outputs a certain moral-deontic status (of action

[158] Of course the modal tag is a functional assumption. For instance, one way of "impossible" representations (acts represented as impossible) may fail to be inputs to FM is by FM operating on an action representation tree defining a possibility space for the agent. FM could operate on such action trees, assigning deontic status only to potential actions. The question of how the potential actions are selected would have to be answered by a theory more general than that of FM.

[159] The parenthesised clause is important, because even on this hypothesis, we ought not to predict that FM accepts representations of *all* kinds (however modally tagged). For example, an object flying faster than the speed of light may be represented as an impossible *event*, but presumably, such representations are not subject to deontic status at all: FM does not accept them as input. I would presume the same to be true of all non-action event representations (and many many others besides), but I won't argue that point here.

[160] Standing for something like "action ϕ is obligatory for agent S at time t". (Here, I am assuming that the argument structure of deontic concepts minimally consists in agent and act representations—such complex mental representations are supposed to be at least part of the output of FM. This assumption is somewhat controversial, but one can substitute one's favourite candidate.

ϕ for agent *S* at a time *t*). However, due to an independent conception of fairness, we deny that there is an obligation because we believe it would be unfair to make/insist on such a judgment. But the computations influencing this (overall) judgment are downstream from the processing that results in computing moral-deontic status for actions. So in a sense, "we" (i.e. more central reasoning mechanisms) end up rejecting the output of the mechanisms dedicated to the attribution of moral-deontic status. On this type of account, OIC is not a principle peculiar to FM (unlike in the case of the previous two hypotheses), rather, it derives or results from other components of the mind.

As an analogy, consider watching a cartoon. As a matter of automatic processing and more specifically due to the engagement of the "theory of mind" mechanism (e.g. Leslie, 1994; Baron-Cohen, 1995), we interpret the cartoon as depicting agents with intentions, beliefs and other mental states ("Tom intends to avenge Jerry's insolence"). On reflection though, we know of course that these mental states we attribute to the characters are not strictly speaking appropriate:[161] we reject their objective reality as well as that of the characters for that matter. (This is also an instance of what Pylyshyn [1984, 1999] refers to as *cognitive impenetrability*: the phenomenon whereby the computations carried out by a mental mechanism or module cannot be altered by the operations or representations of central cognition, such as the explicit beliefs and desires of the agent.)

4.2.2. OIC as an acquisition principle (explanatory adequacy)

I shall briefly mention another interpretation of OIC in the context of moral psychology, just because I am not entirely convinced that it is a non-starter. Note first that there has been an ambiguity inherent in our discussion of OIC as a descriptively adequate principle of moral cognition. First, OIC may be understood as applying at the level of synchronic moral judgment.[162] This is the type of account we have been assuming so far. Second, it also seems plausible to assume that moral rules (or faculty principles)[163] tend to concern things (actions) that we are generally capable of performing. So might OIC be a diachronic influence on possible

[161] And neither is the impression that the characters are physically continuous three (or even two) dimensional entities—another trick our visual system plays on us (*cf*. Pinker, 1997, Chapter 1).

[162] This classification may be rendered more complex by making the Chomskyan distinction between the problems of *production* and *perception* that is standard in modern psycholinguistics (i.e. in this case, how our behaviour is driven or guided by deontic reasoning and how we predict, understand and evaluate other agents' behaviour in terms of moral concepts). Although it is reasonable to expect shared mechanisms and representations, it is also reasonable to expect significant dissimilarities (see the application of this distinction in reasoning research—producing and evaluating arguments—e.g. in Mercier, 2016, where the difference between the two is rather striking).

[163] In this subsection, I revert to referring to faculty principles as "rules" (as per Dwyer, 2008). By "norm", I mean a pattern in a group or a community that tends to result in the acquisition of rules in the relevant group or community.

moral rules? An affirmative answer to this question would suggest an understanding of OIC as an acquisition principle.

Again, like in the previous case, this proposal need not be tied to a strong nativist framework. For instance, it might be understood as a specification of Sripada and Stich's moderate or mild nativist model of rule acquisition and the psychology of rule based thinking (Sripada & Stich, 2007—they refer to it as "norm acquisition", but see *fn*. 163). Stripada and Stich posit what they refer to as the *Acquisition Mechanism*, which provides the input to the so-called *Rule Database* (what they refer to as a "Norm Database"). Whatever gets in to the Rule Database (as a function of the Acquisition Mechanism) will procure the characteristic functional role of (normative) rules in the cognitive economy of the individual—such as an intrinsic (i.e. non-instrumental) compliance motivation, and an intrinsic punitive motivation (in case of violation).[164] Their theory leaves the question open as to what informational constraints the Acquisition Mechanism and the Rule Database have, but that is all right, after all, it is but an empirically informed theory sketch. As Stich himself puts it elsewhere, "one of the components in [their] theory is a norm database, and *it is the job of the theory to tell us what can and cannot end up in that database*. In so doing, the theory will give us an increasingly informative account of the natural kind that we call 'norms'" (Stich, 2009, *p.* 224, emphasis added). Thus, if one is in favour of the Stich-Sripada model, one can interpret OIC as characterising or being due to the operations of the Acquisition Mechanism. This will be only the beginning of a fully explanatory account (e.g. how is OIC implemented and by what algorithm), but if OIC satisfies the expectations of explanatory adequacy, then it can drive research on such more delicate issues.

Of course the most obvious doubt for this interpretation becomes apparent when considering how strange a system of norms (or rules for that matter) having no regard for performability would be. In line with the proposed "rationales" for OIC discussed in Section 1, such a system would be pointless and/or unfair, since agents could not help but violate the prevailing norms requiring actions that are impossible to perform. Still, the question remains: is the the fact that there are no *rules* that require actions that are impossible to perform due to some general concern for avoiding pointlessness and unfairness, and to that extent, do we assume norm acquisition to be a "rational" process, or is it due to a representational constraint on the mechanism(s) dedicated to the acquisition of rules, and to that extent, do we regard norm acquisition to be a process that is rational to a much more limited degree (*cf*. Fodor, 1981)?

[164] Together (database + device(s) implementing motivational profile), these are referred to as the *Execution Mechanism*. Thus, for better or worse, Sripada and Stich do not draw a sharp distinction between the production and the perception problems (see Chapter 1, Section 3.4).

The motivation behind the second option may be that we already know that rule acquisition can be automatic and may rely on input only minimally informative about the presence or absence of norms, indicating a poverty of stimulus situation. An example is the phenomenon of "promiscuous normativity" in children, whereby they infer the presence of a norm on the basis of the performance of a single action (Schmidt et al., 2016). Of course this is in no way a definitive argument in favour of the hypothesis that OIC is an acquisition principle. After all, if norm acquisition operates via observing others perform actions (let's say when certain cues are simultaneously present), then at no point would there be an opportunity for inferring and acquiring norms/rules prescribing actions that are impossible to perform (this would account for the presumed explanatory adequacy of OIC in a "cheap" way). Yet no one assumes that rule acquisition (whether moral or nonmoral) proceeds with exclusive reliance on actions seen performed in the presence of certain cues. For instance, not only do we acquire rules about what is obligatory to do, but also (perhaps even more importantly) about what is forbidden, that is, what is obligatory *not* to do, only a subset of which are we likely to see performed.

Whatever the merits of the above speculations, one thing is obvious: we can *learn* any rule whether or not it regulates a possible or an impossible action. Let us say there is a norm in tribe X according to which agents of type A have to jump 5 metres high every time they hear the sound of a whistle. Unfortunately for them, though, type A agents cannot jump even 2 metres high. I've just come up with this scenario, and presumably the reader will have found it strictly speaking conceivable and will have learned it upon first exposure. The question, from the point of view of the study of moral cognition, and more particularly, FM, is how this learning happened and whether it is analogous to the way moral rules are normally acquired.

5. Concluding remarks

In the literature on the OIC principle, there has been a theoretical impasse: there are two general hypotheses of the presumed descriptive adequacy of OIC, namely, SH and PH, and yet both of them fail on their own terms. This is a problem if we take the bottom line of Chapter 3 seriously; that is, the conclusion that ordinary people do in fact reason in line with OIC. In this chapter, I proposed an analysis of why this situation obtains and provided a way out of the conundrum. This involves (i) assuming the theoretical framework of LA (as defended in chapters 1 and 2), and (ii) hypothesising that OIC is to be understood vis-à-vis the (synchronic or diachronic) operations of FM. It rather straightforwardly follows from the marriage of these proposals that we have no good reason to maintain either SH or PH *qua* explanations of OIC's

descriptive success. In the final part of the chapter (Section 4.2), I provided two general hypotheses as to the ways in which (ii) can be cashed out.

Conclusion

In this thesis, I hope to have achieved two things. First, I endorsed a version of the Linguistic Analogy as the best extant framework for the study of moral cognition, explicated my version of it in some detail, compared it favourably to another popular framework, the Dual Process framework, and defended it against influential criticism. Second, I provided a case study that demonstrates some of the main strengths of the LA by considering OIC as a candidate principle of the moral faculty. To this end, I first defended the sustained plausibility of the claim that OIC is a descriptively adequate principle of the human mind. Then, I considered the two predominant extant hypotheses—SH and PH—for why OIC is descriptively adequate. Although problems associated with these accounts are well known in the literature on OIC, LA offers a novel analysis of why these proposals are ultimately unsuccessful. Furthermore, LA also provides some novel hypotheses of OIC.

References

Aharoni, E., Sinnott-Armstrong, W., & Kiehl, K. A. (2014). What's wrong? Moral understanding in psychopathic offenders. *Journal of Research in Personality*, *53*, 175–181.

Ahlenius, H., & Tännsjö, T. (2012). Chinese and westerners respond differently to the trolley dilemmas. *Journal of Cognition and Culture*, *12*(3–4), 195–201.

Anderson, A. R. (1967). Some nasty problems in the formal logic of ethics. *Noûs*, *1*(4), 345–360.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *Psychology of Learning and Motivation*, *2*, 89–195.

Axelrod, R. (1984). *The Evolution of Cooperation*. New York: Basic Books.

Baillargeon, R., Scott, R. M., He, Z., Sloane, S., Setoh, P., Jin, K., ... Bian, L. (2015). Psychological and sociomoral reasoning in infancy. In M. Mikulincer & P. R. Shaver (Eds.), *APA Handbook of Personality and Social Psychology, Volume 1: Attitudes and Social Cognition.* (Vol. 1, pp. 79–150). American Psychological Association.

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, *54*(7), 462–479.

Baron-Cohen, S. (1995). *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge MA: MIT Press.

Barrett, H. C. (2015). Modularity. In V. Zeigler-Hill, L. L. M. Welling, & T. K. Shackelford (Eds.), *Evolutionary Perspectives on Social Psychology* (pp. 39–49). Dordrecht: Springer.

Barrett, H. C., Bolyanatz, A., Crittenden, A. N., Fessler, D. M. T., Fitzpatrick, S., Gurven, M., ... Laurence, S. (2016). Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences*, *113*(17), 4688–4693.

Baumard, N. (2016). *The Origins of Fairness: How Evolution Explains Our Moral Nature: How Evolution Explains Our Moral Nature*. Oxford: Oxford University Press.

Baumard, N., André, J. B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, *36*(1), 59–78.

Beller, S. (2008). *Deontic norms, deontic reasoning, and deontic conditionals*. *Thinking & Reasoning 14*(4), 305-341.

Besch, T. M. (2011). Factualism, normativism and the bounds of normativity. *Dialogue-Canadian Philosophical Review*, *50*(2), 347–365.

Blair, R. J. R. (1995). A cognitive developmental approach to morality: Investigating the psychopath. *Cognition*, *57*(1), 1–29.

Bratman, M. (1987). *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Buckwalter, W., & Turri, J. (2015). Inability and obligation in moral judgment. *PLOS ONE*, *10*(8), 1–20.

Carruthers, P. (2006). *The Architecture of the Mind: Massive Modularity and the Flexibility of Thought*. Oxford: Oxford University Press.

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

Chomsky, N. (1964). *Current Issues in Linguistic Theory*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge MA: MIT Press.

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin and Use*. New York: Praeger.

Chomsky, N. (2000). *New Horizons in the Study of Language and Mind*. Cambridge University Press.

Chomsky, N. (2018). Two notions of modularity. In R. G. de Almeida & L. R. Gleitman (Eds.), *On Concepts, Modules, and Language: Cognitive Science at Its Core* (pp. 25–40). New York: Oxford University Press.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*, 489–558.

Cima, M., Tonnaer, F., & Hauser, M. D. (2010). Psychopaths know right from wrong but don't care. *Social Cognitive and Affective Neuroscience*, *5*(1), 59–67.

Collins, J. (2004). Faculty disputes. *Mind & Language*, *19*(5), 503–533.

Cooper, N. (1966). Some presuppositions of moral judgments. *Mind*, *75*(297), 45–57.

Copp, D. (2008). "Ought" implies "can" and the derivation of the Principle of Alternate Possibilities. *Analysis*, *68*(1), 67–75.

Cummins, R. (2000). "How does it work?" versus "what are the laws?": Two conceptions of psychological explanation. In F. C. Keil & R. A. Wilson (Eds.), *Explanation and Cognition* (pp. 117–144). Cambridge MA: MIT Press.

Cushman, F. A. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380.

Cushman, F. A. (2016). The psychological origins of the Doctrine of Double Effect. *Criminal Law and Philosophy*, *10*(4), 763–776.

Cushman, F. A., Young, L., & Hauser, M. D. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*, *17*(12), 1082–1089.

Cushman, F. A., Young, L. L., & Greene, J. D. (2010). Multi-system moral psychology. In J. M. Doris (Ed.), *The Moral Psychology Handbook* (pp. 47–71). Oxford: Oxford University Press.

Dancy, J. (1983). Ethical particularism and morally relevant properties. *Mind*, *92*(368), 530–547.

de Waal, F. (1996). *Good Natured: The Origins of Right and Wrong in Humans and Other Animals*. Cambridge MA: Harvard University Press.

Descioli, P., & Kurzban, R. (2013). A solution to the mysteries of morality. *Psychological Bulletin*, *139*(2), 477– 496.

Dienes, Z. (2008). *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference*. Basingstoke: Palgrave Macmillan.

Donagan, A. (1979). *The Theory of Morality*. Chicago: University of Chicago Press.

Driver, J. (2011). Promising too much. In H. Sheinman (Ed.), *Promises and Agreements: Philosophical Essays* (pp. 183–197). Oxford: Oxford University Press.

Dupoux, E., & Jacob, P. (2007). Universal moral grammar: A critical appraisal. *Trends in Cognitive Sciences*, *11*(9), 373–378.

Dupoux, E., & Jacob, P. (2008). Response to Dwyer and Hauser: Sounding the retreat? *Trends in Cognitive Sciences*, *12*(1), 2–3.

Dwyer, S. (2008). How not to argue that morality isn't innate: Comments on Prinz. In W. Sinnott-Armstrong (Ed.), *Moral Psychology Volume 2: The Cognitive Science of Morality* (pp. 407–418). Cambridge MA: MIT Press.

Dwyer, S., & Hauser, M. D. (2008). Dupoux and Jacob's moral instincts: throwing out the baby, the bathwater and the bathtub. *Trends in Cognitive Sciences*, *12*(1), 1–2.

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*(1), 255–278.

Evans, J. S. B. T., & Frankish, K. (Eds.). (2009). *In Two Minds: Dual Processes and Beyond*. Oxford: Oxford University Press.

Fede, S. J., Borg, J. S., Nyalakanti, P. K., Harenski, C. L., Cope, L. M., Sinnott-Armstrong, W., ... Kiehl, K. A. (2016). Distinct neuronal patterns of positive and negative moral processing in psychopathy. *Cognitive, Affective and Behavioral Neuroscience*, *16*(6), 1074–1085.

Fessler, D. M. T., Barrett, H. C., Kanovsky, M., Stich, S., Holbrook, C., Henrich, J., ... Laurence, S. (2015). Moral parochialism and contextual contingency across seven societies. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1813), 20150907.

Fitch, W. T., Hauser, M. D., & Chomsky, N. (2005). The evolution of the language faculty: Clarifications and implications. *Cognition*, *97*, 179–210.

Fodor, J. A. (1974). Special sciences (Or: The disunity of science as a working hypothesis). *Synthese*, *28*, 97–115.

Fodor, J. A. (1981). The present status of the innateness controversy. In *RePresentations* (Harvester, pp. 257–316). Brighton.

Fodor, J. A. (1983). *The Modularity of Mind*. Cambridge, MA: MIT Press.

Fodor, J. A. (2000). *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. Cambridge, MA: MIT Press.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, *5*, 1–7.

Foot, P. (1972). Morality as a system of hypothetical imperatives. *The Philosophical Review*, *81*(3), 305–316.

Forrester, J. W. (1989). *Why You Should: The Pragmatics of Deontic Speech*. Hanover NH: Brown University Press.

Fox, C., & Feis, G. (2018). 'Ought implies Can' and the law. *Inquiry*, *61*(4), 370–393.

Frank, R. H. (1988). *Passions Within Reason: The Strategic Role of the Emotions*. New York: W.W. Norton & Company.

Frankish, K., & Evans, J. S. B. T. (2009). The duality of mind: An historical perspective. In J. S. B. T. Evans & K. Frankish (Eds.), *In Two Minds: Dual Processes and Beyond* (pp. 1–29). Oxford: Oxford University Press.

Fuller, L. L. (1969). *The Morality of Law (Revised Edition)*. New Haven: Yale University Press.

Gigerenzer, G. (2014). *Risk Savvy: How to Make Good Decisions*. Allen Lane.

Goldman, A. I. (1970). *A Theory of Human Action*.

Grabowski, A. (2013). *Juristic Concept of the Validity of Statutory Law: A Critique of Contemporary Legal Nonpositivism*. Springer.

Greene, J. D. (2013). *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. New York: Penguin Press.

Greene, J. D. (2015). The rise of moral cognition. *Cognition*, *135*, 39–42.

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*(3), 364–371.

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

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

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*, 2105–2108.

Greene, M. (2006). Levels of adequacy, observational, descriptive, explanatory. In *Encyclopedia of Language & Linguistics Volume 1* (2nd ed., pp. 49–51). Elsevier Science.

Griffin, J. (1992). The human good and the ambitions of consequentialism. *Social Philosophy and Policy*, *9*(2), 118–132.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 83–95.

Haidt, J. (2013). *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. New York: Vintage Books.

Haji, I. (2002). *Deontic Morality and Control*. Cambridge: Cambridge University Press.

Hamlin, J. K. (2015). Does the infant possess a moral concept? In E. Margolis & S. Laurence (Eds.), *The Conceptual Mind: New Directions in the Study of Concepts* (pp. 477–517). Cambridge, MA: MIT Press.

Hampshire, S. (1951). Symposium: Freedom of the Will. *Proceedings of the Aristotelian Society*, *25*, 161–178.

Hare, R. M. (1951). Symposium: Freedom of the Will. *Proceedings of the Aristotelian Society*, *25*, 201–216.

Hare, R. M. (1963). *Freedom and Reason*. Oxford: Oxford University Press.

Harman, G. (2000). *Explaining Value and Other Essays in Moral Philosophy*.

Hauser, M. D. (2006). *Moral Minds: The Nature of Right and Wrong*. Harper Collins.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it , who has it, and how did it evolve? *Science*, *298*(5598), 1569–1580.

Hauser, M. D., Cushman, F. A., Young, L., Jin, R. K.-X., & Mikhail, J. (2007). A dissocation between moral judgments and justifications. *Mind & Language*, *22*(1), 1–21.

Hauser, M. D., Young, L. L., & Cushman, F. A. (2008a). Reviving Rawls's linguistic analogy: Operative principles and the causal structure of moral actions. In W. Sinnott-Armstrong (Ed.), *Moral Psychology Volume 2: The Cognitive Science of Morality* (pp. 107–143). Cambridge MA: MIT Press.

Hauser, M. D., Young, L. L., & Cushman, F. A. (2008). On misreading the linguistic analogy: Response to Jesse Prinz and Ron Mallon. In W. Sinnott-Armstrong (Ed.), *Moral Psychology Volume 2: The Cognitive Science of Morality* (pp. 171–179). Cambridge MA: MIT Press.

Henne, P., Chituc, V., De Brigard, F., & Sinnott-Armstrong, W. (2016). An empirical refutation of "Ought" Implies "Can." *Analysis*, *76*(3), 283–290.

Hieronymi, P. (2004). The force and fairness of blame. *Philosophical Perspectives*, *18*, 115–148.

Hilpinen, R. (Ed.). (1971). *Deontic Logic: Introductory and Systematic Readings*. Dordrecht: Reidel.

Howard-Snyder, F. (2006). "Cannot" implies "not ought." *Philosophical Studies*, *130*(2), 233–246.

Huebner, B., Dwyer, S., & Hauser, M. D. (2009). The role of emotion in moral psychology. *Trends in Cognitive Sciences*, *13*(1), 1–6.

Hume, D. [1739] (1964). *A Treatise of Human Nature*. Oxford: Clarendon Press.

Jackendoff, R. (1994). *Patterns In The Mind: Language And Human Nature*. New York: BasicBooks.

Jackendoff, R. (1999). The natural logic of rights and obligations. In R. Jackendoff, P. Bloom, & K. Wynn (Eds.), *Language, Logic, and Concepts: Essays in Memory of John Macnamara* (pp. 67–95). Cambridge MA: MIT Press.

Jackendoff, R. (2002). *Foundations of Language*. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press.

Jackendoff, R. (2003). Précis of Foundations of Language: Brain, Meaning, Grammar, Evolution, *26*(1), 651–707.

Jackendoff, R. (2007). *Language, Consciousness, Culture: Essays on Mental Structure*. Cambridge, MA: MIT Press.

Jackendoff, R., & Pinker, S. (2005). The nature of the language faculty and its implications for evolution of language (Reply to Fitch, Hauser, and Chomsky). *Cognition*, *97*(2), 211–225.

Joerden, J. C. (2012). Deontological square, hexagon, and decagon: A deontic framework for supererogation. *Logica Universalis*, *6*, 201–216.

Joyce, R. (2006). *Evolution of Morality*. Cambridge, MA: MIT Press.

Joyce, R. (2016). *Essays in Moral Skepticism*. Oxford: Oxford University Press.

Julesz, B. (2006). *Foundations of Cyclopean Perception*. Cambridge MA: MIT Press.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454.

Kanger, S. (1971). New foundations for ethical theory. In R. Hilpinen (Ed.), *Deontic Logic: Introductory and Systematic Readings* (pp. 36–58). Dordrecht: Reidel.

Kant, I. [1793] (1998). *Religion within the Boundaries of Mere Reason and Other Writings*. (A. Wood & G. Di Giovanni, Eds.), *Cambridge Texts in the History of Philosophy*. Cambridge: Cambridge University Press.

Kelsen, H. (1967). *Pure Theory of Law*. New Yersey: Lawbook Exchange.

Kelsen, H. (1991). *General Theory of Norms*. Oxford: Clarendon Press.

King, A. (2017). 'Ought Implies Can': Not so pragmatic after all. *Philosophy and Phenomenological Research*, *95*(3), 637–661.

Koenigs, M., Young, L. L., Adolphs, R., Tranel, D., Cushman, F. A., Hauser, M. D., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, *446*(7138), 908–911.

Kohlberg, L. (1969). Stage and sequence: The cognitive-developmental approach to socialization. In D. Goslin (Ed.), *Handbook of Socialization Theory and Research* (pp. 347–480). Chicago: Rand McNally.

Kohlberg, L. (1984). *Essays on Moral Development: The Psychology of Moral Development*. San Francisco: Harper & Row.

Kratzer, A. (1977). What "must" and "can" must and can mean. *Linguistics and Philosophy*, *1*(3), 337–355.

Kumar, V. (2015). Moral judgment as a natural kind. *Philosophical Studies*, *172*(11), 2887–2910.

Kurthy, M., & Lawford-Smith, H. (2015). A brief note on the ambiguity of 'ought'. Reply to Moti Mizrahi's 'Ought, can and presupposition: An experimental study.' *Methode*, *4*(6), 244–249.

Kurthy, M., Lawford-Smith, H., & Sousa, P. (2017). Does ought imply can? *PLoS ONE*, *12*(4), 1-24.

Leben, D. (2018). In defense of "Ought Implies Can." In T. Lombrozo, J. Knobe, & S. Nichols (Eds.), *Oxford Studies in Experimental Philosophy. Volume 2* (pp. 151–166). Oxford: Oxford University Press.

Leslie, A. M. (1994). ToMM, ToBy, and Agency: Core architecture and domain specificity. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture* (pp. 119–148). Cambridge: Cambridge University Press.

Levine, S., Leslie, A. M., & Mikhail, J. (2018). The mental representation of human action. *Cognitive Science*, *42*, 1229–1264.

Levy, N. (2005). The good, the bad and the blameworthy. *Journal of Ethics & Social Philosophy*, *1*(2), 2–16.

Lewis, D. (1975). Languages and language. In K. Gunderson (Ed.), *Language, Mind, and Knowledge* (pp. 3–35). Minneapolis: University of Minnesota Press.

Littlejohn, C. (2009). "Ought", "can" and practical reasons. *American Philosophical Quarterly*, *46*(4), 363–372.

MacDonald, A. W., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science*, *288*, 1835–1839.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147–186.

Mallon, R. (2008). Reviving Rawls's linguistic analogy inside and out. In W. Sinnott-Armstrong (Ed.), *Moral Psychology Volume 2: The Cognitive Science of Morality* (pp. 145–155). Cambridge MA: MIT Press.

Marr, D. (2010). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge: MIT Press.

Martin, W. (2009). Ought but cannot. *Proceedings of the Aristotelian Society*, *109*, 103–128.

Mason, E. (2003). Consequentialism and the "Ought Implies Can" principle. *American Philosophical Quarterly*, *40*(4), 319–331.

Mason, H. E. (Ed.). (1996). *Moral Dilemmas and Moral Theory*. Oxford: Oxford University Press.

McGuire, J., Langdon, R., Coltheart, M., & Mackenzie, C. (2009). A reanalysis of the personal/impersonal distinction in moral psychology research. *Journal of Experimental Social Psychology*, *45*(3), 577–580.

McNamara, P. (1996). Must I do what I ought? (or will the least I can do do?). *Proceedings of the 3rd International Workshop on Deontic Logic in Computer Science (DEON 1996), Sesimbra, Portugal, January 11-13, 1996*, 154–173.

McNamara, P. (2018). Deontic logic. In *Stanford Encyclopedia of Philosophy* (Fall, 2018). Retrieved from https://plato.stanford.edu/entries/logic-deontic/

Mendez, M. F., Anderson, E. D., & Shapira, J. S. (2005). An investigation of moral judgement in frontotemporal dementia. *Cognitive and Behavioral Neurology*, *18*(4), 193–197.

Mercier, H. (2016). The argumentative theory: Predictions and empirical evidence. *Trends in Cognitive Sciences*, *20*(9), 689–700.

Mikhail, J. (2000). *Rawls' Linguistic Analogy: A Study of the "Generative Grammar" Model of Moral Theory by John Rawls in A Theory of Justice (PhD Thesis)*. Cornell University.

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 (JD Thesis)*. Stanford Law School.

Mikhail, J. (2008a). The poverty of the moral stimulus. In *Moral Psychology Volume 1: The Evolution of Morality* (pp. 353–359). Cambridge, MA: MIT Press.

Mikhail, J. (2008b). Moral cognition and computational theory. In W. Sinnott-Armstrong (Ed.), *Moral Psychology Volume 3: The Neuroscience of Morality* (pp. 81–91). Cambridge MA: MIT Press.

Mikhail, J. (2009). Moral grammar and intuitive jurisprudence: A formal model of unconscious moral and legal knowledge. *Psychology of Learning and Motivation*, *50*(C), 27–100.

Mikhail, J. (2011). *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. Cambridge MA: Cambridge University Press.

Mikhail, J. (2013). New perspectives on moral cognition: Reply to Zimmerman, Enoch, and Chemla, Egre, and Schlenker. *Jerusalem Review of Legal Studies*, *8*(1), 66–114.

Mikhail, J. (2017). Chomsky and moral philosophy. In J. McGilvray (Ed.), *The Cambridge Companion to Chomsky* (2nd ed., pp. 235–254). Cambridge: Cambridge University Press.

Miłkowski, M. (2013). *Explaining the Computational Mind*. Cambridge MA: MIT Press.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202.

Miller, G. A., & Chomsky, N. (1963). Finitary models of language users. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of Mathematical Psychology, Vol. II* (pp. 419–492). New York: Wiley.

Mizrahi, M. (2015). Ought, can, and presupposition: An experimental study. *Methode*, (6), 232–243.

Moro, A. (2008). *The Boundaries of Babel: The Brain and the Enigma of Impossible Languages*. Cambridge, MA: MIT Press.

Morris, A., & Cushman, F. A. (2018). A common framework for theories of norm compliance. *Social Philosophy & Policy*, 35 (1):101-127

Navarro, P. E. (2013). The Efficacy of Constitutional Norms. In L. D. D'Almeida, J. Gardner, & L. Green (Eds.), *Kelsen Revisited: New Essays on the Pure Theory of Law* (pp. 77–99). Oxford: Hart Publishing.

Nichols, S. (2004). *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford: Oxford University Press.

Nichols, S. (2005). Innateness and moral psychology. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The Innate Mind Volume 1: Structure and Contents* (pp. 353–369). Oxford: Oxford University Press.

Nowak, M. A., & Sigmund, K. (1998). The dynamics of indirect reciprocity. *Journal of Theoretical Biology*, *194*(4), 561–574.

Nucci, L. P. (2001). *Education in the Moral Domain*. Cambridge: Cambridge University Press.

Piaget, J. (1932). *The Moral Judgment of the Child*. London: Routledge and Kegan Paul.

Piazza, J., Sousa, P., Rottman, J., & Syropoulos, S. (2018). Which appraisals are foundational to moral judgment? Harm, injustice, and beyond. *Social Psychological and Personality Science*, 1–11.

Pinker, S. (1997). *How the Mind Works*. London: Penguin Books.

Pinker, S., & Jackendoff, R. (2005). The faculty of language: What's special about it? *Cognition*, *95*(2), 201–236.

Pizarro, D. A., & Bloom, P. (2003). The intelligence of the moral intuitions: Comment on Haidt (2001). *Psychological Review*, *110*(1), 193–196.

Prinz, J. J. (2008). Resisting the linguistic analogy: A commentary on Hauser, Young, and Cushman. In W. Sinnott-Armstrong (Ed.), *Moral Psychology Volume 2: The Cognitive Science of Morality* (pp. 156–170). Cambridge MA: MIT Press.

Prinz, J. J. (2009). Against moral nativism. In *Stich: And His Critics* (pp. 167–189).

Prinz, J. J. (2014). Where do morals come from? – A plea for a cultural approach. In M. Christen, C. P. van Schaik, J. Fischer, M. Huppenbauer, & C. Tanner (Eds.), *Empirically Informed Ethics: Morality between Facts and Norms* (pp. 99–116). Springer.

Putnam, H. (1975). The nature of mental states. In *Mind, Language, and Reality: Philosophical Papers, Vol. 2* (pp. 51–58). Cambridge: Cambridge University Press.

Pylyshyn, Z. W. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science*. Cambridge MA: MIT Press.

Pylyshyn, Z. W. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioural and Brain Sciences*, *22*, 341–423.

Ranganathan, S. (2010). Does Kant hold that ought implies can? In J. Sharma & A. Raghuramaraju (Eds.), *Grounding Morality: Freedom, Knowledge and the Plurality of Cultures* (pp. 60–87). Routledge.

Rawls, J. (1971). *A Theory of Justice*. Cambridge MA: Belknap.

Reber, A. S. (1993). *Implicit Learning and Tacit Knowledge: An Essay on the Cognitive Unconscious*. New York: Oxford University Press.

Richard J. Maddock. (1999). The retrosplenial cortex and emotion: New insights from functional neuroimaging of the human brain. *Trends in Neurosciences*, *22*(7), 310–316.

Robinson, C. D. (2016). *Multilingual Law: A Framework for Analysis and Understanding*. London: Routledge.

Roedder, E., & Harman, G. (n.d.). *Moral grammar*.

Roedder, E., & Harman, G. (2010). Linguistics and moral theory. In J. M. Doris (Ed.), *The Moral Psychology Handbook* (p. 493). Oxford: Oxford University Press.

Saka, P. (2000). "Ought" does not imply "can." *American Philosophical Quarterly*, *37*(2), 93–105.

Sapontzis, S. F. (1991). "'Ought' does imply 'can.'" *The Southern Journal of Philosophy*, *29*(3), 382–393.

Scherer, K. R., Schorr, A., & Johnstone, T. (Eds.). (2001). *Appraisal Processes in Emotion*. Oxford: Oxford University Press.

Schmidt, M. F. H., Butler, L. P., Heinz, J., & Tomasello, M. (2016). Young children see a single action and infer a social norm: Promiscuous normativity in 3-year-olds. *Psychological Science*, *27*(10), 1360–1370.

Schmidt, M. F. H., & Rakoczy, H. (2018). Developing an understanding of normativity. In A. Newen, L. C. de Bruin, & S. Gallagher (Eds.), *The Oxford Handbook of 4E Cognition* (pp. 685-706). Oxford: Oxford University Press.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*(1), 1–66.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, *84*(2), 127–190.

Sinnott-Armstrong, W. (1984). 'Ought' Conversationally Implies 'Can.' *The Philosophical Review*, *93*(2), 249–261.

Sinnott-Armstrong, W., & Wheatley, T. (2014). Are moral judgments unified? *Philosophical Psychology*, *27*(4), 451–474.

Sousa, P., & Piazza, J. (2014). Harmful transgressions qua moral transgressions: A deflationary view. *Thinking & Reasoning*, *20*(1), 99–128.

Southwood, N. (2016). "The thing to do" implies "can." *Nous*, *50*(1), 61–72.

Sripada, C. S. (2008). Nativism and moral psychology: Three models of the innate structure that shapes the contents of moral norms. In *Moral Psychology Volume 1: The Evolution of Morality* (pp. 319–343). Cambridge MA: MIT Press.

Sripada, C. S., & Stich, S. (2007). A Framework for the psychology of norms. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The Innate Mind Volume 2: Culture and Cognition* (Vol. 2, pp. 280–301). Oxford: Oxford University Press.

Sterelny, K. (1990). *The Representational Theory of Mind: An Introduction*. Oxford: Basil Blackwell.

Sterelny, K. (2010). Moral nativism: A sceptical response. *Mind & Language*, *25*(3), 279–297.

Stern, R. (2004). Does "ought" imply "can"? And did Kant think it does? *Utilitas*, *16*(1), 42–61.

Stich, S. (2006). Is morality an elegant machine or a kludge? *Journal of Cognition and Culture*, *6*(1–2), 181–189.

Stich, S. (2009). Replies. In D. Murphy & M. Bishop (Eds.), *Stich: And His Critics* (pp. 190–252). Chichester: Wiley-Blackwell.

Stocker, M. (1971). "Ought" and "can." *Australasian Journal of Philosophy*, *49*(3), 303–316.

Streumer, B. (2003). Does "ought" conversationally implicate "can"? *European Journal of Philosophy*, *11*(2), 219–228.

Streumer, B. (2007). Reasons and impossibility. *Philosophical Studies*, *136*(3), 351–384.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology: General*, *18*, 643–662.

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal*, *94*(6), 1395–1415.

Thomson, J. J. (2008). Turning the trolley. *Philosophy and Public Affairs*, *36*(4), 359–374.

Trivers, R. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, *46*(1), 35–57.

Trivers, R. (2002). *Natural Selection and Social Theory: Selected Papers of Robert Trivers*. Oxford: Oxford University Press.

Turiel, E. (1983). *The Development of Social Knowledge: Morality and Convention*. Cambridge, England: Cambridge University Press.

Turiel, E. (2002). *The Culture of Morality: Social Development, Context, and Conflict*. Cambridge: Cambridge University Press.

Turri, J. (2017). How "ought" exceeds but implies "can": Description and encouragement in moral judgment. *Cognition*, *168*, 267–275.

Turri, J., & Blouw, P. (2015). Excuse validation: A study in rule-breaking. *Philosophical Studies*, *172*(3), 615–634.

Uniacke, S. (1998). The principle of double effect. In E. Craig (Ed.), *Routledge Encyclopedia of Philosophy, Vol. 3*. Routledge.

Vallentyne, P. (1989). Two types of moral dilemmas. *Erkenntnis*, *30*(3), 301–318.

van Lier, J., Revlin, R., & de Neys, W. (2013). Detecting cheaters without thinking: Testing the automaticity of the cheater detection module. *PLoS ONE*, 8(1), 1-8.

van Someren Greve, R. (2014). "Ought", "can", and fairness. *Ethical Theory and Moral Practice*, *17*(5), 913–922.

Vogelstein, E. (2012). Subjective reasons. *Ethical Theory and Moral Practice*, *15*(2), 239–257.

von Fintel, K., & Iatridou, S. (2008). How to say ought in foreign: The composition of weak necessity modals. In J. Gueron & J. Lacarme (Eds.), *Time and Modality* (pp. 115–141). Springer.

von Wright, G. H. (1983). Norms of higher order. *Studia Logica*, *42*(2–3), 119–127.

Vranas, P. (2007). I ought, therefore I can. *Philosophical Studies*, *136*(2), 167–216.

Vranas, P. B. M. (2018). I ought therefore I can obey. *Philosopher's Imprint*, *18*(1), 1–36.

Wellman, H. M., & Miller, J. G. (2008). Including deontic reasoning as fundamental to theory of mind. *Human Development*, *51*, 105–135.

Whalen, P. J., Kagan, J., Cook, R. G., Davis, F. C., Kim, H., Polis, S., ... Johnstone, T. (2004). Human amygdala responsivity to masked fearful eye whites. *Science*, *306*, 2061.

Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, *358*(6389), 749–750.

Zimmerman, A. (2013). Mikhail's naturalized moral rationalism. *Jerusalem Review of Legal Studies*, *8*(1), 44–65. Zimmerman, M. J. (1990). Where did I go wrong? *Philosophical Studies*, *59*, 55–77.

Zimmerman, M. J. (1996). *The Concept of Moral Obligation*. *Cambridge studies in philosophy*. Cambridge: Cambridge University Press.

Zimmermann, R. (1990). *The Law of Obligations: Roman Foundations of the Civilian Tradition*. Cape Town: Juta & Co.

Zinchenko, O., & Arsalidou, M. (2018). Brain responses to social norms: Meta-analyses of fMRI studies. *Human Brain Mapping*, 39(2), 955–970.