Library of Ethics and Applied Philosophy 31

Bert Musschenga
Anton van Harskamp *Editors*

What Makes Us Moral? On the Capacities and Conditions for being Moral

Springer

What Makes Us Moral? On the capacities and conditions for being moral

LIBRARY OF ETHICS AND APPLIED PHILOSOPHY

VOLUME 31

Editor in Chief

Marcus Düwell, *Utrecht University, Utrecht, NL*

Editorial Board

Deryck Beyleveld, *Durham University, Durham, U.K.*
David Copp, *University of Florida, USA*
Nancy Fraser, *New School for Social Research, New York, USA*
Martin van Hees, *Groningen University, Netherlands*
Thomas Hill, *University of North Carolina, Chapel Hill, USA*
Samuel Kerstein, *University of Maryland, College Park, USA*
Will Kymlicka, *Queens University, Ontario, Canada*
Philippe Van Parijs, *Louvaine-la-Neuve (Belgium) en Harvard, USA*
Qui Renzong, *Chinese Academy of Social Sciences, China*
Peter Schaber, *Ethikzentrum, University of Zürich, Switzerland*
Thomas Schmidt, *Humboldt University, Berlin, Germany*

For further volumes:
http://www.springer.com/series/6230

Bert Musschenga • Anton van Harskamp
Editors

What Makes Us Moral?
On the capacities and conditions for being moral

Springer

Editors
Bert Musschenga
Faculty of Philosophy
VU University
Amsterdam
The Netherlands

Anton van Harskamp
Faculty of Philosophy and Department of Social and Cultural Anthropology
VU University
Amsterdam
The Netherlands

ISSN 1387-6678
ISBN 978-94-007-6342-5 ISBN 978-94-007-6343-2 (eBook)
DOI 10.1007/978-94-007-6343-2
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2013939263

© Springer Science+Business Media Dordrecht 2013
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

‘What makes us moral’? is an intriguing question which lends itself for various interpretations. Who are the ‘we’? What does it mean to be moral? Are we made into moral beings by external, causal factors? Many other questions can be added. This plurality of interpretations made ‘What makes us moral?’ into an excellent theme for an academic conference. The conference was an initiative of the Section for Practical Philosophy of the Faculty of Philosophy of VU University, Amsterdam, and was held on June 22–24, 2011. The majority of the chapters of this book were first presented and discussed during the conference, and reviewed anonymously and revised after the conference. We are most grateful to the reviewers who did an excellent job.

Conferences, even small and modest ones, cost money. A grant from the legacy of VU University’s former Blaise Pascal Institute made it possible both to organise the conference and to prepare this volume for publication. We thank Johan Kraay for careful editing of the texts.

Amsterdam
September 2012

Bert Musschenga and Anton van Harskamp
The editors

Contents

Part I Morality, Evolution and Rationality

Part II Morality and the Continuity Between Human and Nonhuman Primates

Part III Nativism and Non-Nativism

Part IV Religion and (Im)Morality

Part V Morality Beyond Naturalism

Chapter 1
What Makes Us Moral? An Introduction

Bert Musschenga

1.1 Why Be Moral; Why Are We Moral; What Makes Us Moral?

Human beings learn a lot of things and develop many skills during childhood. They learn to walk, speak, read, write, ride a bike, and also to ‘behave’. They learn that it is improper to upset their plate if they don’t like the food, and proper to thank their grandparents for gifts, to offer a hand to visitors, that it is nice to help their little sister, and wrong to lie to their parents, kick other children, and steal their Nintendo. Growing up, they come to see that there is a difference between the rules of the schoolteacher and rules such as not to beat up other children or not to hurt animals. Children who don’t obey the rules are ‘disobedient’ or ‘naughty’; children who regularly beat up other children, steal things and hurt or kill animals are ‘bad’ or ‘little criminals’. Children who never give things away or never help others are ‘selfish’. For most children learning to be good and to do the right thing is a ‘natural’ part of growing up. Although they sometimes feel strangely attracted to children who are naughty and bad, becoming like them is not a real option for them.

Most of the Moroccans and the Turks in the Netherlands are Muslim. They don’t ask themselves why they are Muslim rather than Christian. Being a Muslim is part of being a Moroccan or a Turk. What worries them is not that they are Muslim and not Christian, but how to be a good Muslim. Being a Muslim is self-evident for them, being a good Muslim is not. Most adults see themselves as moral persons.

B. Musschenga (✉)
Faculty of Philosophy, VU University, De Boelelaan 1105,
1081 HV Amsterdam, The Netherlands
e-mail: a.w.musschenga@vu.nl

B. Musschenga and A. van Harskamp (eds.), *What Makes Us Moral? On the capacities and conditions for being moral*, Library of Ethics and Applied Philosophy 31,
DOI 10.1007/978-94-007-6343-2_1, © Springer Science+Business Media Dordrecht 2013

What matters to them is not why they are moral, but that they don't always know what is the right thing to do in a particular situation and, when they do know, why they don't do it.

Besides this commonality between moral persons and religious believers, there is also a difference. I am quite sure that no one ever wrote a book with a title such as 'Why I am moral', while there are books titled 'Why I am a Christian', 'Why I am a Muslim' or 'Why I am an Atheist'. While reflective Christians and Muslims sometimes feel the need to account for their faith in God or Allah, moral persons never feel an urge to explain why they are moral – with the exception, of course, of moral philosophers.

In the last century, many philosophers took up the centuries-old question 'Why be moral?' This wasn't an existential question for them, they wouldn't stop being moral if they didn't find a satisfying answer. They were interested in finding out whether morality is rational, whether there are good reasons for being moral. The question wasn't purely academic for most of these moral philosophers. Something important was – and is – at stake. If morality were rational, it should be possible to convince people who, for one reason or another, do not play the moral game – selfish people, fanatics, and moral perpetrators – that they must participate in the game. Unfortunately, even among philosophers there is no consensus on whether morality and being moral is rational. Surely not when 'rational' is interpreted as 'required by reason'. Most of them will agree that being moral is not irrational, that reason allows but does not require being moral, that it is irrational not to be moral. The debate on 'Why be moral' was illuminating because it provided deeper insight into morality's (weak) foundations, but it didn't give the friends of morality the argumentative weapons by which they could knock down morality's outsiders. Those among the outsiders who take pride in being rational wouldn't be impressed by arguments suggesting that being moral is not less rational than staying outside the morality game. More important, part of these outsiders not only don't play the morality game, they are also lacking in rationality.

While the question 'Why be moral?' asks for a justification of morality and being moral, the question that is central to this book, 'What makes us moral?', asks for an explanation. This question presupposes that we already know what being moral is. However, part of most fundamental discussions on morality regards what being moral means. Let's say that being moral means that one isn't solely focused on one's own interests, but also takes account of others who have interests and are capable of well-being, and from time to time even makes sacrifices for their sake. Still, the question is open to diverse interpretations. The first is what capacities and competences we as moral beings have that enable us to be moral. Capacities and competences are different things. I don't have the competences needed for being a surgeon, although I might have the capacities of becoming one. We have to start identifying moral competences, look for the underlying capacities, and find out what steps have to be taken for developing capacities into competences. But someone who has a driving licence, who officially has the competence to drive a car, may still cause an accident. The same applies to beings with moral competences. We cannot take for granted that moral beings always act morally. This is the second interpretation of the question ‘What makes us moral?’ What causes moral beings to make, from time to time, minor and major mistakes? The causes can be internal as well as external. External causes for road accidents are, for example, a flat tyre, ice on the road, or an error on the part of another road user. Internal causes are, for example, a heart attack or distraction by reading a map. Some moral agents often make moral mistakes, while others almost always do the right thing. This observation also asks for an explanation. Does one agent have more moral luck than another? Or are the latter ones more strongly motivated and/or do they have a stronger capacity to resist temptations?

It is common sense that one question often triggers a lot of other questions. Knowledge of the main moral capacities and competences is not sufficient for explaining immoral actions. Immoral action is not the only category that asks for an explanation. Moral philosophers usually make a distinction between immorality and amorality. The average moral person violates from time to time, intentionally or unintentionally, moral rules. Moral violators usually have feelings of guilt, shame or remorse. If that is the case, they continue to be seen as participants in the morality game. But if people don’t show such feelings, doubts will arise whether they are moral at all. People who don’t care at all for the interests and the well-being of others are said to be amoral. They either lack the relevant capacities or something has gone wrong on the developmental path from capacity to competence. Knowledge of the capacities and competences for being moral can suggest directions in which the causes of amorality should be sought. Reversely, insight into characteristics and deficiencies of amoral persons may also contribute to our knowledge of the capacities and competences required for being moral.

I took a big leap from ‘Why be moral?’ to ‘What makes us moral?’ Actually, there is one step in between. The question ‘Why are we moral?’ forms, as it were, the transition. Again, this question also lends itself to different interpretations. Does it make sense that we are moral? What is the point, the purpose of being moral? In the language of business the question would be what the ‘added value’ of being moral is. Evolutionary biologists, who regard everything as a product of evolution, would rephrase it as ‘What is the evolutionary advantage of being moral?’ (In a way, we can say that they already know what makes us moral: It is evolution.) ‘Why be moral?’ is a purely philosophical question. It is a question moral philosophers can discuss among themselves. The other two questions, ‘Why are we moral?’ and ‘What makes us moral?’ force them to leave the field of moral philosophy and to turn to empirical social and human sciences. Finding answers to these questions is a truly interdisciplinary enterprise.

1.2 Part I: Morality, Evolution and Rationality

Until now, I just gave an overview of questions that arise when reflecting on what makes us moral and why we are moral. It’s time now to introduce the chapters to this volume and to explain how they relate to these questions.

The chapters of *Part I* discuss the relation between evolution and morality. The basic tenet of evolution theory is that those behaviours and dispositions will be selected that maximize an individual's fitness for survival and reproduction. Evolution puts, one could say, a premium on selfishness. How then is it possible that cooperative behaviour, which requires co-operators to restrain their selfish behaviour and from time to time even to make sacrifices, is so widespread in the animal world? If, as is argued, it is the function of morality to facilitate cooperation between individuals, how could it emerge from evolution and persist? One might answer by arguing that the benefits of cooperation exceed the costs. But selfish individuals will always try to cash the benefits and not to pay the costs. This requires that they succeed in deceiving their partners about their true intentions. Cooperation can only be successful if co-operators are able to detect individuals with a disposition to cheat.

This issue is the subject of Alejandro Rosas's chapter. He argues that selfishness as such is not the biggest threat to cooperation. The biggest threat is the capacity to deceive. Deceit can only succeed if motivations and dispositions are not transparent, which is indeed the case. He hypothesizes that cooperative behaviour could only persist because co-operators developed a capacity for recognizing which potential partners possess what he calls 'a disposition for fairness', a genuine cooperative attitude. A precondition for detecting this cooperative attitude is the presence of mind-reading abilities.

In her chapter, Katharine Browne argues that an explanation for cooperative behaviour such as Rosas's is not valid for human cooperation. What is distinctive about human cooperation, she says, is its scale and scope. These characteristics make it particularly difficult to reconcile morality with the usual evolutionary explanatory mechanisms. She defends an explanatory account of the emergence of cooperative behaviours that appeals to cultural group selection. There is, she says, an analogy to the problem that cooperation poses for evolution theory in the normative context. Cooperative behaviour requires that individuals constrain self-interested pursuits. But 'rationality' as it is most commonly understood requires that agents act as selfish utility-maximizers. It thus seems that cooperative behaviour is at odds with rational behaviour. In the normative context she argues that David Gauthier (1986) is successful in his attempt to show that adopting a disposition to constrained maximization is rational. His argument consists of two moves. The first is to show that the disposition to constrained maximization will yield a greater utility than an alternative disposition. The second is to make the case that the rationality of a disposition entails the rationality of the actions that the disposition recommends. It is Browne's conviction that Gauthier's view offers reconciliation between moral action and rationality.

Rational Choice Theory (RCT) is especially dominant within economics. Many economists take for granted that parties on the market place are disposed to maximize their preferences. The theory doesn't exclude altruism and morality; it just treats them as contingent preferences on a par with other preferences. Already in 1977, Amartya Sen criticized the behavioural foundations of economic theory in his seminal article 'Rational fools'. The alternative view that Sen develops goes in another direction than Gauthier with his morally constrained maximization of preferences. On Sen's view, actions that are chosen on the basis of a moral judgment rather than on the basis of preference-satisfaction are hard to accommodate for RCT, because moral behaviour is to be understood as 'committed behaviour', which means that it is action resulting from counter-preferential choice.

In their chapter, Catherine Herfeld and Katrien Schaubroeck analyse Sen's concept of commitment and argue that it is impossible for RCT to accept the Kantian interpretation of practical rationality that underlies Sen's concept of commitment. They tie in with Sen's alternative conceptualization of moral behaviour in terms of a choice that implies a meta-ranking, i.e. a ranking over preference-rankings. Herfeld and Schaubroeck argue that enriching the concept of meta-ranking with Harry Frankfurt's concept of care would enable the rational choice theorist to accommodate moral behaviour into the economic picture in a way that is compatible with the most fundamental concepts of economics (Frankfurt 1999).

Even in theories that do not see rationality as the basis of morality, rationality is still an important aspect of moral agency. We expect that moral persons show consistency and coherence in their judgments and actions. Consistency and coherence are demands of rationality. Coherence plays an important role in theories of justification. In their chapter, Markus Christen and Thomas Ott deal with the practicability of coherence as an instrument to analyse the behaviour of moral agents. They present a descriptive notion of coherence and show how it can be used for understanding the way the structure of someone's moral beliefs may influence his behaviour. In particular, they show how their approach is able to integrate different types of coherence relationships between single beliefs. In this way, their notion of coherence allows them to analyse how cognitive and affective similarities between reasons used in moral decision making may interrelate. Furthermore, it can give novel insights into phenomena like practical irrationality in decision making.

1.3 Part II: Morality and the Continuity Between Human and Nonhuman Primates

There is another evolutionary approach to morality that doesn't ask for the evolutionary advantages of moral behaviour, but starts with assuming that, because of evolutionary continuity, the basic capacities for being moral were already present in humans' evolutionary predecessors and are present in their evolutionary next of kin, the nonhuman primates. Some students of animal behaviour contend that the systems for the regulation of social behaviour found with nonhuman primates cannot be called moralities, although they make use of the same capacities as required for being moral. Other students hold that these systems already are moralities, though less complex than the ones found with humans. In his chapter, Bert Musschenga states that rules constitute the central mechanism of human morality. A species whose moral behaviour is regulated by rules can be said to possess a morality.

We may assume that a species has a morality if it can be shown that members of that species can have moral motives and if social disapproval occurs after the violation of rules. Musschenga thinks that only nonhuman primates such as chimpanzees and bonobos have a morality. Animal morality regulates behaviour automatically and unconsciously. While human morality to a large extent also functions automatically and unconsciously, it also makes use of conscious and reflective processes.

The issue of the (dis)continuity between human morality and animal (proto-)morality is also discussed by Florian Cova. Continuists think we share with animals the psychological requirements for morality, and discontinuists think that we are endowed with unique moral capacities. Cova argues that the two perspectives could be reconciled by distinguishing two components of our moral life: moral agency (we are morally responsible for our actions) and moral judgments (we are able to evaluate our behaviour and those of others). To side with the discontinuist, Cova says, it is hard to deny that moral life is much richer in human beings than in any other moral animals: we are able to ask tough moral questions and can reason about difficult moral situations (such as moral dilemmas). Nevertheless, the continuist can also cite cases in which it is hard to deny that other species can have a real moral life. Cova concludes that humans are both moral agents and moral judges, while nonhuman primates are only agents.

Evolutionary continuity is also central to Andrés Luco's chapter on the motivation of moral action. He starts with an overview of Humean and anti-Humean theories of moral motivation. Anti-Humeans claim – and Humeans deny – that a cognitive state, or a state with both cognitive and affective features need not be accompanied by an independently existing affective state, to generate motivation. In his defence of the Humean theory of motivation, Luco appeals to a 'continuity constraint' that requires that the origin of an observed psychological trait should be explicable as a product of descent with modifications of pre-existing traits. He argues that the Humean theories do, and anti-Humean theories do not, meet this constraint. The motivations which Humeans posit to account for moral action are the same as, or very similar to, the motivations that researchers have invoked to explain proto-moral behaviours in nonhuman primates. Luco adds another element to his defence of Humean theories. He appeals to 'Morgan's Canon'. In 1894 the comparative psychologist C. Lloyd Morgan formulated this principle, stating that we may not interpret an action as the outcome of the exercise of a higher faculty if it can be interpreted as the outcome of exercises of a capacity which stands lower in the psychological scale (Morgan 1903).

Empathy is considered to be the core or at least the most essential building block for morality. Empathy, however, has many evolutionary and developmental stages. No one will deny that from a certain age onwards young children have a morality, although adults clearly have more developed empathic capacities than children. Thus, it is not the most developed form of empathy that is a precondition for morality. The relevant form of empathy is the one which enables to 'read' another's mind. Without that capacity, it is impossible to understand what other beings feel, think, and why they behave the way they do. In his chapter, Harry Wels makes a distinction between empathy as object of research and as research methodology.

He suggests that many scientists interested in empathy with animals are themselves using empathy as research methodology. The obvious example is Frans de Waal. Interestingly enough, Wels says, it seems that 'empathy-as-an-approach-to-fully-understand-others-especially-non-human-animals' has penetrated popular culture nowadays far more than that it is recognized or accepted as a route to knowledge in science. Popular culture abounds with stories, articles, documentaries and books about people who through sheer empathy learn the ways of an animal and are able to literally live with them. Think for example of the rather popular (wild) animal whisperers. His chapter should, according to Wels, also be read as a first attempt to try and reflect on formulating a trans disciplinary research methodology, essentially trying to 'merge' the social and biological sciences, although other disciplines are equally included like psychology, literary criticism and philosophy, as are other knowledge producers, the various 'whisperers', from the domain of popular culture.

1.4 Part III: Nativism and Non-nativism

In the second half of the last century there was a peak in the part of the 'nature–nurture debate' focused on the origins of morality. Impressed by the wide variety and great diversity of moral customs and practices, anthropologists took the stance that moralities must – at least for the greater part – be cultural constructions. Evolutionary biologists, however, concluded from their studies on the evolution of social and cooperative behaviour that human morality must have a biological, thus universal, core. The majority of anthropologists and some philosophers found that moral relativism (epistemic and normative) was the only theory compatible with the data on moral diversity. Other philosophers argued that the unmistakable diversity of morals should not blind us to underlying universal structures and categories. Morality may be context sensitive, but not entirely context relative. Diversity is compatible with universality on the level of basic principles. In this century, the universalist tree got two new shoots, the 'moral grammar' theory of which Susan Dwyer (2008), John Mikhail (2008) and Marc Hauser (2006) are the most important proponents, and the 'moral foundations' theory of Jonathan Haidt et al. Adherents of the moral grammar theory build on John Rawls's suggestion that, analogous to Noam Chomsky's innate universal linguistic structures, there are also innate universal moral structures and principles (Rawls 1971). Haidt et al. speak of universal moral intuitions on top of which cultures build incommensurable moralities (Haidt and Joseph 2004). Unlike nativists' formal universal structures and principles, these intuitions have content. One of most outspoken critics of moral nativism undoubtedly is Jesse Prinz (2007), who rejects it as 'constructive sentimentalism'.

Moral nativism is the subject of *Part III* of this book. In the first chapter of that part Jessy Giroux takes moral grammar theory and moral foundationalism together under the heading 'Output Model' which he compares with the 'Input Model'. The Input Model sees moral norms as social constructions transmitted to individuals through socialization; the Output Model views moral norms as naturally emerging from human dispositions. Giroux agrees with anti-nativists in saying that one does not need a nativist framework to account for moral universality. Unlike most anti-nativists, however, he shares the nativist's essential intuition that moral universality speaks in favour of a conception of morality as emerging *from within* human individuals. As alternative, he presents a model of his own, the ‘Direct Outgrowth’ model: moral norms should be seen as the natural extension of human dispositions rather than as social constructions. In this perspective, he says, moral norms are directly shaped by humans' emotional preparedness and by a variety of other innate dispositions. These innate dispositions are not specifically moral, which means that they did not evolve for the purpose of morality. But they do lead ‘normally constituted’ individuals to develop naturally certain basic moral principles.

The discussion on nativism is continued by Carsten Fogh Nielsen. Nielsen is not an anti-nativist who denies that some innate learning mechanism is necessary to fully explain moral development. His aim is ‘to issue a warning against a lingering reductive tendency found among certain contemporary moral nativists: a tendency to greatly exaggerate the importance of such innate mechanisms for moral development while simultaneously downplaying the importance of other factors in this process’. Moral nativists such as Susan Dwyer, Marc Hauser and John Mikhail have, according to Nielsen, a tendency to implicitly ignore, downplay and underestimate the actual complexity and breadth of the moral stimuli which children are exposed to and encounter during upbringing. Referring to the work of Kim Sterelny (2010) and of Hegel (2008), Nielsen stresses the importance of human culture for the transmission of moral norms and the influence of social structures for the moral development of human beings.

In her chapter, Julia Hermann entirely shifts the focus from the innate basis of morality or innate learning mechanisms to the process of moral learning itself. The end of that process is the morally competent person. A morally competent agent has a number of capacities, including a capacity for moral judgment, a capacity for empathy, and a capacity for feeling remorse when this is regarded as appropriate. Moral competence is not only revealed, but also acquired through practice. To clarify the nature of the moral learning process, she refers to Gilbert Ryle's (1972) and Dreyfus and Dreyfus's (1991) ideas on acquiring virtues or ‘ethical expertise’. Hermann rejects, as they do, the view that learning rules is sufficient, that becoming moral is merely a matter of understanding rationally that some things are right and others wrong, and that we become moral by reasoning. Moral teaching and learning involve a significant amount of training, such as conditioning, i.e., pure drill in addition to arguments, teaching and learning by example, being corrected, being praised and blamed, learning in concrete situations, and gradual improvement through practice.

We have noticed that capacities should be distinguished from competences. Competences are the result of a process of education and training. Absence of the required capacities makes every process of education and training impossible. Moral capacities – more precisely the capacities that are a precondition for moral development – are mental capacities. Brain damage and mental illnesses may affect these capacities. Persons with mental disorders often also show moral deficiencies in moral reasoning or in moral behaviour, or in both. Researches into the areas of the brain that are affected by mental disorders have revealed a lot about the neurological basis of moral capacities. Lacking the relevant capacities, persons with brain damage will never develop fully competent moral agency, which is the condition for moral responsibility. But exactly how do mental disorders affect moral responsibility? This question is central to Gerben Meynen's chapter. He aims to develop a conceptual framework that can help explicate and straightforwardly communicate our intuitions on mental disorder and diminished responsibility. Four factors are identified which, according to Meynen, contribute to the explanation of why mental disorders excuse. The first concerns the cluster of 'free will' – or agency-related phenomena (like having alternative possibilities, or being the genuine source of the action); the second factor concerns extreme urges; the third factor concerns false beliefs; the fourth factor is moral sensitivity. Referring to one or more of these factors should, says Meynen, enable us to explain the various instances in which we either partially or completely excuse a person because of a mental disorder.

Referring to a number of studies, Darcia Narvaez states that moral capacities seem to be diminishing, at least in the U.S.A. In her chapter, she offers an explanation. According to her, the source may lie in the deficient developmental environments of children which sometimes are only apparent in adolescence or adulthood. It all starts with early life experience. Our brains and bodies evolved to expect responsive care, of which children nowadays don't get enough. Current child-rearing practices are insufficiently responsive to the emotional needs of children, which may lead to insecure attachment. Personalities that emerge from insecure attachment are less empathic, more aggressive, and have difficulty getting along with peers. The deficient developmental environments also have a negative impact on morality, because early experience shapes moral functioning later.

1.5 Part IV: Religion and (Im)Morality

In many publications of educationalists and philosophers of education autonomy is presented as the aim of moral education. Autonomous persons act on values and principles they can endorse. Moreover, they are internally motivated and don't act out of fear of sanctions such as social disapproval and legal punishment. They are considered to be motivationally self-sufficient. This Enlightenment ideal of moral autonomy still meets mistrust in some parts of the world, for example in the U.S.A., where a belief in God as divine Lawgiver and Judge is regarded as a condition for being moral. That explains why Americans distrust atheists. Atheists cannot be counted on to respect morality. Many authors, however, have argued that morality is logically independent from religion. In the first contribution to *Part IV*, Stephen Maitzen wants to take a further step. He argues that the existence of theism's God who is omnipotent and possesses perfect knowledge and perfect goodness, logically precludes the existence of certain basic moral obligations on our part, such as the obligation to prevent or relieve terrible suffering by a child when we easily can. We have that obligation, according to Maitzen, only if no perfect being is allowing the child's suffering to occur, and hence only if no perfect being exists. No recognizable morality remains, argues Maitzen, if we lack even the basic obligation to relieve a child's suffering. One might respond by saying that Maitzen's argument presupposes a classical theistic conception of God that is left behind by modern theology and doesn't connect with the faith of modern believers. However, Maitzen argues that modern conceptions of God which lack one or more of the classic perfections, provide no better implications for morality.

While (some) religious believers think that atheists are not morally trustworthy, (some) atheists point at the intimate connection between religion and violence. For them, religion is the main cause of violence. Religion is not only the cause of violent conflicts, it also justifies violence. In the Islamic world suicide terrorists are seen as martyrs, at least by radical believers. Suicide terrorism is, just as any form of terrorism, condemned by moral philosophers such as Michael Walzer, Igor Primoratz and Marcel Hénaff as immoral or even evil. They might be right in their condemnation, says Anton van Harskamp in his contribution, but they have too narrow a view of suicide terrorism. They tend to overlook the interplay of terrorism with cultural, economic, political and individual violations by non-state actors, by state actors and by structural violence. Moreover, they don't pay attention to the religious dimension of the actions of the suicide bombers. For suicide terrorists, their act is a religious one. Van Harskamp refers to Kierkegaard's idea of the teleological suspension of the ethical. The basic idea in this suspension of ethics is that God is the only source of morality. If God orders Abraham to sacrifice his son, such sacrifice is what morality demands from him even if it contradicts common morality. If Allah wants you to kill infidels by killing yourself, executing His will is your duty, a duty that goes far beyond common morality.

Maitzen rejects theism because it is logically incompatible with the most evident moral obligations. After World War II many people gave up their belief in God for historical reasons. The God they believed in would have prevented the Holocaust. That the Holocaust could occur proves that God doesn't exist. While in many conflicts both parties claim to have God on their side, in the conflict between the Jews and the Nazis God seemed to be absent on both sides. The Nazis didn't need God's help or His approval, while the Jews who called for His help didn't get it. For most of us the atrocities of the Nazis are as incomprehensible as suicide terrorism. Hannah Arendt argued in *Eichmann in Jerusalem* (1963) that Nazis such as Adolph Eichmann were ordinary people, just like you and me. Her view gets support from influential social psychologists such as Stanley Milgram (1974) and Philip Zimbardo (2007) who state that their studies show that what makes people immoral goes beyond all situational and systemic forces. In her chapter, Bettine Siertsema discusses American-French author Jonathan Littell's controversial novel *The Kindly Ones* (2010). In this novel Littell tells us in a detached way about Max Aue, a fictional figure who was involved in the mass murders in Eastern Europe, such as Babi Yar, the defeat of the German army at Stalingrad and the deportation of the Hungarian Jews and the death marches from Auschwitz. Aue's attitude is ambivalent: he shows both unease and lack of guilt in describing his actions. His justifications show similarities with those of historical perpetrators like Rudolf Höss and Adolf Eichmann. Aue wants to convince the reader that he is 'just like you', that the reader would have acted in exactly the same way when in his position. He seems to implicitly refer to social psychologists such as Philip Zimbardo who in his *The Lucifer effect; How good people turn evil* (2007) says that, depending on the situation in which one finds himself, anyone can become an evil person. Siertsema stresses that such an appeal cannot excuse what Aue did. The basic cause of his immorality is not 'the situation', but his belief in the dominance of the collective over the individual. This goes as much for his own individuality, which he to a great extent sacrifices to the idea of the nation or the *Volk*, embodied in the *Führer*, as it does for the individuality of the victims, whom he sees first and foremost as parts of a hostile collective, however incredible their potential danger as such may be.

1.6 Part V: Morality Beyond Naturalism

Naturalism is clearly the dominant approach in most accounts of the origin of morality, which is also illustrated by the contributions to this volume. The dominant naturalistic tradition in moral philosophy, says David Edward Rose in the chapter that opens *Part V*, assumes that the basic capacities and characteristics of the individual exist prior to his socialization and that morality arises in the necessity of cooperation. Naturalistic moral philosophers either assume that peace and social stability require that aggressive instincts are directed from external objects to the individual himself or that humans have a natural tendency to cooperation. Rose offers us a third alternative which he finds in Giambattista Vico's moral sociology. Vico, says Rose, referring to an article by Ernesto Grassi (1976), conceives of morality as an expression of the transfer of our innate instincts and passions into a human system of relations designed by our imagination and forming the roots and foundations of our ethical motivations. Society is not an organic development – not a product of evolution – but an act of will. It requires a leap of belief from the world as it appears, to the arcane world of reasons and values. Reason is made possible by the symbolic, poetic expression of human understanding. Morality is artificial, not natural.

Even naturalists don't think that we are able to explain moral and immoral behaviour by solely referring to the psychological competences of individuals. We need beliefs and values to motivate and orientate our moral actions, and we need rules, roles, rituals and institutions to provide stability to our moral life. Adam Seligman points at the importance of rituals. In his chapter he says that, embedded in a world of particularities and change, of probabilities and contingencies, the world of real life decisions is inherently uncertain. Accommodating this ambiguity, he argues, is a necessary, if perhaps not sufficient perquisite for morally informed acts. Its necessity can be most easily accessed through the very idea of empathy; that imaginative blurring of boundaries that makes social life possible. Ritual is for him a crucial resource in building a capacity for ambiguity, for leaving a space for the other and so for empathy. 'Rituals create a subjunctive space, a shared "could be" that constructs individuals in relation to others. Ritual, in its formal, iterated and enacted moments, presents a unique human resource for dealing with ambiguity and the multifocal nature of all relationships – with beings human and divine. Ritual defines and binds entities, times and spaces. By creating such borders it also links entities, times and spaces to what lies beyond their immediate field. As such, it presents a coherent and embracing way to live in a plural and hence also deeply ambiguous universe, one where order's rules can never really be known, but still must be acted upon.'

Every collection of papers is at risk to show only a loose connection between the contributions. This introduction made an attempt to show that there is coherence between the parts of the book and the individual papers. Together they offer a wide, be it still incomplete, perspective on the various aspects of what makes humans moral.

References

Arendt, H. 1963. *Eichmann in Jerusalem. A report on the banality of evil*. London: Faber and Faber.

Dreyfus, H.L., and S.E. Dreyfus. 1991. Towards a phenomenology of ethical expertise. *Human Studies* 14: 229–250.

Dwyer, S. 2008. How not to argue that morality isn't innate: Comments on Prinz. In *Moral psychology*, vol. 2, ed. W. Sinnott-Armstrong, 407–418. Boston: MIT Press.

Frankfurt, H.G. 1999. On caring. In *Necessity, volition, and love*, H.G. Frankfurt, 155–180. Cambridge: Cambridge University Press.

Gauthier, D. 1986. *Morals by agreement*. Oxford: Oxford University Press.

Grassi, E. 1976. Marxism, humanism, and the problem of imagination in Vico's works. In *Giambattista Vico's science of humanity*, ed. G. Tagliacozzo and D. Verene (trans: Azodi, A.), 275–294. Baltimore: John Hopkins University Press.

Haidt, J., and C. Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus: On Human Nature* 133: 55–66.

Hauser, M.D. 2006. *Moral minds: How nature designed our universal sense of right and wrong*. New York: Ecco Press.

Hegel, G.W.F. 2008. *Outlines of the philosophy of right*. Trans. T.M. Knox, ed., rev. with introduction by Stephen Houlgate. Oxford: Oxford University Press.

Littell, J. 2010. *The kindly ones* (translated from the French by Charlotte Mandell, *Les Bienveillantes*, 2008). London: Harper Perennial.

Mikhail, J. 2008. The poverty of the moral stimulus. In *Moral psychology*, vol. 1, ed. W. Sinnott-Armstrong, 353–359. Boston: MIT Press.

Milgram, S. 1974. *Obedience to authority; An experimental view*. New York: Harper & Row.

Morgan, C.L. 1903. *An introduction to comparative psychology*, 2nd ed. London: W. Scott.

Prinz, J.J. 2007. *The emotional construction of morals*. Oxford: University Press.

Rawls, J. 1971. *A theory of justice*. Harvard: Harvard University Press.

Ryle, G. 1972. Can virtue be taught? In *Education and the development of reason*, ed. R.F. Dearden, P.H. Hirst, and R.S. Peters, 323–332. London/Boston: Routledge and Kegan Paul.

Sen, A. 1977. Rational fools; A critique of the behavioral foundations of economic theory. *Philosophy and Public Affairs* 6: 17–34.

Sterelny, K. 2010. Moral nativism; A skeptical response. *Mind and Language* 25: 279–297.

Zimbardo, P. 2007. *The Lucifer effect; How good people turn evil*. London: Rider.

Part I
Morality, Evolution and Rationality

Chapter 2
Rationality and Deceit: Why Rational Egoism Cannot Make Us Moral

Alejandro Rosas

2.1 Human Cooperation and Evolutionary Altruism

Cooperation is a pervasive phenomenon in the biological world. Evolutionary biologists hold it responsible for the existence of hierarchical levels of biological organization. Genomes and multi-cellular organisms behave as individuals and are treated by scientists as such, but they evolved out of independent, lower level biological units, in a process e.g. from single cells to groups of such cells, before those groups evolved into multi-cellular organisms as individuals in their own right. Biologists believe with good evidence that cooperation between lower level units drove this process (Buss 1987; Maynard Smith and Szathmáry 1997; Michod 2007). Through cooperation, individuals at the lower level obtained benefits that were not otherwise available. In this process, either cheating was not a problem, or natural selection had to solve it. It is not exaggerated to say that evolutionary biologists place nowadays as much emphasis on the role of cooperation in evolution as traditionally was placed on competition.

The ubiquity of cooperation in the biological world suggests that cooperation among humans has a biological basis. The claim is that we have been designed by natural selection to cooperate throughout history and pre-history at the large scale peculiar to humans. In this paper I shall not present or develop arguments explicitly defending natural selection as the designer of humans as co-operators. Those interested in the arguments for an evolutionary explanation of human cooperation can find them in the first four chapters of Joyce (2006). (I assume, as Joyce does, that cooperation is the main theme of moral norms.) Most evolutionary biologists assume the legitimacy of the evolutionary approach to morality. Their efforts in this subject have been mainly devoted to solve a particular evolutionary problem.

A. Rosas (✉)
Philosophy Department, Universidad Nacional de Colombia, Bogotá, Colombia
e-mail: arosasl@unal.edu.co

B. Musschenga and A. van Harskamp (eds.), *What Makes Us Moral? On the capacities and conditions for being moral*, Library of Ethics and Applied Philosophy 31,
DOI 10.1007/978-94-007-6343-2_2, © Springer Science+Business Media Dordrecht 2013

Morality seems to involve a paradoxical lifetime sacrifice in fitness to the benefit of others, including those that are the least moral in the group. Why doesn't selection eliminate moral traits, if it selects the fittest organisms? Biologists have at least three theories available to solve this paradox: kin selection, reciprocal altruism and group selection. Group selection enjoys popularity with some philosophers and in particular with a group of social scientists, who advocate strong reciprocity to explain human cooperation, a strategy that requires group selection to evolve (Sober and Wilson 1999; Gintis et al. 2003).[1] Darwin was perhaps the first scientist to advance a group-selection hypothesis. He was painfully aware of the paradox involved in viewing human morality as an adaptation and speculated that groups of morally motivated humans proved more adaptive than groups of selfish individuals in ancestral intertribal warfare (Darwin 1981, chap. 5).

For those endorsing an explanation based on group selection, human cooperation is biologically altruistic. This means, as stated above, that the cooperative agent suffers a paradoxical lifetime sacrifice in fitness to the benefit of other individuals. The sacrifice is paradoxical because, *prima facie*, a trait that sacrifices fitness should not evolve. Since there are currently rival proposals about how the sacrifice is to be conceptualized and measured, the concept of evolutionary altruism is bedevilled with uncertainties that are getting more complicated as the debate develops. For example, most evolutionary theorists consider reciprocal altruism as a biologically selfish (fitness enhancing) trait (Lehmann and Keller 2006; West et al. 2007): it entails a short-term sacrifice in fitness, but in the long term the trait enhances the fitness of its carrier. Nonetheless, reciprocal altruism is potentially vulnerable to exploitation, which was probably the reason why Trivers used the label 'altruism'. In any case, if a helping trait is to evolve, it must have higher fitness than rival traits: labelling a helping trait 'altruistic' in the sense of involving a fitness sacrifice apparently contradicts this obvious fact. Biologists invoke differences between direct and indirect fitness effects (West et al. 2007), or between within-group and between-group fitness (Sober and Wilson 1999) to legitimate the view that some traits really sacrifice the fitness of their carriers and, nevertheless, are selected. Except for by-product mutualisms, which clearly involve no sacrifice (Sachs et al. 2004), cooperative traits seem both to enhance and to sacrifice the fitness of carriers, depending on the point of view.

The question of evolutionary altruism is a tricky one. I gladly endorse a conceptual reform based on the concept of positive assortment, as suggested in some recent literature (Fletcher and Doebeli 2009; Rosas 2010; Bowles and Gintis 2011, table 4.1, p. 75). But luckily, I do not need to discuss this here. For the purposes of this paper, I am happy to follow Joyce (2006, pp. 16, 38) in adopting an uncommitted view on the question of evolutionary altruism. Natural selection favours helping or

[1] For a recent sophisticated defence of strong reciprocity and group selection see Bowles and Gintis (2011), especially chapters 6 and 7. For a review see Rosas (2012).

cooperative traits and, for the purposes of this paper, I leave open whether these traits enhance the fitness of their carriers and evolve for this reason, or sacrifice it and evolve for benefits they receive in an indirect, though reliable, manner.

2.2 Social Preferences Versus Selfish Cooperation

The question addressed in this paper is, namely, whether social preferences, specifically altruistic ones, are among the proximate mechanisms explaining cooperative behaviour and morality in humans; and indeed, whether they are indispensable. Social preferences, as in general other conative mental states (motives, desires or emotions), are altruistic if they satisfy two conditions: (1) their content is the welfare of another person or persons, and (2) wanting another's welfare is not instrumental to the agent's own welfare, such that the agent would stop wanting it if she could promote her welfare by other means. An altruistic preference is therefore a non-instrumental positive interest in the welfare of some other person or persons.

When saying that social preferences are necessary both for morality and for cooperative/moral behaviour, I do not mean to conflate morality with behaviour that is merely in accordance with moral demands and was not prompted by motivations that we would deem moral. In particular, I do not want to belittle the importance of judgments with the predicate 'is wrong' as necessary elements in motivations and dispositions that we consider moral. Joyce (2006) has argued this point cogently. However, these considerations are compatible with the view that social preferences, though not sufficient, are necessary for moral judgment. We would not be able to make moral judgments without them. The judgment 'X is wrong' arises, on this view, only if the subject experiences a conflict between the opposing requirements of selfishness on one hand, and a non-instrumental (altruistic) inclination or desire that others do well on the other. This conflict emerges particularly in situations known as social dilemmas. In these situations, it is wrong to follow the morally selfish temptation of profiting at the expense of others' contributions.

Positing social preferences as necessary conditions for cooperation challenges the views of classical economists and game theorists. Their model of the human agent, *homo economicus*, is a non-tuistic agent, i.e., an agent who lacks social preferences. Nonetheless, classical economists and game theorists expect cooperative behaviour from *homo economicus* in repeated games, because cooperation is what rational egoism demands. Recently, this view of cooperation based on an 'invisible hand' coordinating the desires of selfish agents has been challenged by social scientists pointing to the 'dark side of self-interest' (Bowles and Gintis 2011, pp. 5–6). They disapprove of *homo economicus* as a complete explanatory model and demand the inclusion of social preferences, in particular altruistic ones, in the model of a human agent.

Game theorists and classical economists believe that humans cooperate on the basis of egoism because they think selfish rational agents can reason through to the conclusion that cooperation is in their best interest, particularly in iterated prisoner's dilemma games (iPD). Rational egoists run through a normative argument to the effect that they ought to cooperate in iPD games to enhance their own material benefits. On this ground, rational egoism appears as a sufficient explanation for cooperation. One version of the normative argument is the folk theorem in game theory, proving the existence of (many) Nash equilibriums with cooperative strategies in iPD games. Therefore, a rational agent ought to choose one or another cooperative strategy. Another version of the argument is that of Gauthier who tries to prove that even in one-shot PD games cooperation is rational (the best response) under conditions of transparency or translucency (Gauthier 1986).

For those who believe in some role for social preferences in human cooperation, the question arises: what exactly is their role in human cooperation? If rational egoists, (rational agents lacking social preferences) ought to cooperate in their best interest and if they know this, why do they have social preferences and do they (we) really need them? If we need them, shouldn't we simply mistrust the normative argument based on rational egoism? It will be useful to provide here the basic tenets of three different answers to these questions.

Many authors grant that actual cooperative behaviour is more likely based on moral emotions than on a rational argument. Nonetheless, the normative argument need not be superfluous. It offers a justification of the rationality of social and moral emotions. Gauthier, for example, acknowledges that moral behaviour is usually driven by moral emotions (instantiating social preferences like fairness), but the normative question is still worth pursuing, for it examines whether the behaviour prompted by those emotions is rationally justified or should be discarded as irrational (Gauthier 1986, p. 338). His theory thus offers a 'rational reconstruction', and gives reasons for being moral that are not the reasons humans actually follow. But supposing some humans are really able to act on the reasons provided by the normative argument, it does not follow that social emotions or preferences are superfluous. Minds can be designed with redundant mechanisms for purposes that are important for survival, as cooperation in this case.

However, alternatively, it may be the case that humans, being imperfectly rational, are not able to act on the reasons provided by the normative argument unless social emotions come to our aid. Imperfect rationality is the basis for another explanation of the role played by social preferences. They are not simply backup mechanisms: we have them because we need them. They help us comply with what is rational. In this explanation, cooperation is rational for agents that lack social preferences; and perfectly rational agents would be able comply with this rational demand even in the absence of social preferences. But humans, being imperfectly rational, cannot comply because, e.g., of excessive temporal discounting. Social preferences are then required to remedy our imperfect rationality and bring our behaviour back in line with what is rational, namely cooperation. Frank (1988) has pursued this sort of argument, and Joyce (2006, chap. 4) follows this same line.

Finally, a still different explanation says that cooperation is not the rational move for rational agents that lack social preferences. Rationality will not lead rational agents to cooperate with each other if they lack social preferences, because they foresee that their payoffs will be higher if they can coerce or deceive others. In the following I shall argue for this third view in an evolutionary setting. Cooperation would collapse in a human population in the absence of a sufficient proportion of agents with social preferences. And since cooperation is important for survival, we are lucky that natural selection managed to graft social preferences onto our rational mind.

2.3 Selfishness and Deceit

I have sketched three possible evolutionary explanations for why we have social preferences as evolved proximate mechanisms for cooperation. I assume that we do have social preferences. Since perhaps not everybody is willing to grant this, I mention that evidence in favour has been accumulating lately in the field of experimental economics (Bowles and Gintis 2011, chap. 3). With these two assumptions, I now concentrate on arguing for the third view. The third view says that rationality does not recommend cooperation as the best move if rational agents lack social preferences, i.e., if they have no interests in the interests of others or are mutually unconcerned (the interests of others do not figure in their utility function). The reason lies in our imperfect mind-reading abilities. Having imperfect knowledge of the practical intentions of other agents, deception and several forms of coercion are rational for agents without social preferences. Mary Gibson (1977) and Jean Hampton (1991) have made this sort of argument before. I here place it in an evolutionary context.

Assuming natural selection designed the human mind with what it takes to make cooperation possible, there is a wide consensus identifying the basic selection pressure for the emergence of cooperative traits in humans: it is a general structure of interaction represented in the iterated prisoner's dilemma (iPD). In a simple one-shot PD, cooperation is rational and adaptive only if opponents are conditional co-operators that predict reasonably well the intention of their co-players. In repeated games, where players A and B will interact many times, player A can punish through defection at $t + l$ a defection by B at period t. In this case cooperation is also the best response. Punishment is crucial for the stability of cooperation in iPDs. It encourages would-be defectors to cooperate in order to profit from the opportunities of mutual gain (Trivers 1971; Axelrod and Hamilton 1981).

However, repeated interaction is not sufficient and some extra conditions are required. For example, players must have equal power. A dominance relationship prevents the emergence or stability of cooperation, if cooperation is conceived as instantiating a principle of fair distribution of profits (after subtracting costs). An unfair distribution is not cooperation, even if both players gain something beyond their investments. But suppose interaction is free from coercion; cooperation could still fail to evolve. If the iPD is finite, as it always is, there is a temptation to lure others into cooperation only to cheat grandly on the last round. An experienced cheater plans the last round beforehand and then disappears. In this case, cooperation preceding the last cheating move is selfishly instrumental to the substantial profits obtained in that move. This strategy poses a problem for cooperation. It corresponds in spirit to Hobbes's Fool and Hume's Sensible Knave. Those characters use cooperation instrumentally and selfishly to maintain a deceptive reputation for honesty and to create opportunities for exploitation. They endorse morality and the cultivation of a good reputation not for their own sake, but instrumentally: they behave fairly or honestly only when not doing so would damage their reputation. Notwithstanding honesty being a good general policy, it is subject to many exceptions, so that the (egoistically) wisest man is he who observes the general rule and free rides on all the exceptions (Hume 1902, §232). The instrumental cultivation of a reputation for being moral is essentially deceptive (Sayre-McCord 1991). Deceivers will do whatever it takes to appear as moral individuals in the eyes of others, but whenever they can cheat undetected, they will.

The combination of selfishness and deceit is perhaps the crucial objection against the view that cooperation is rational for agents without social preferences. Despite Hume's official doctrine that justice arises from rational self-interest, his words about the Sensible Knave seem to confirm that rationality is not what he lacks, but rather a primitive motive for justice: 'If his heart rebel not against such pernicious maxims, if he feel no reluctance to the thoughts of villainy or baseness, he has indeed lost a considerable motive to virtue' (Hume 1902, §233). In contrast, Hobbes overtly argued that the Fool is in fact irrational (Hobbes 1651). But Hobbes's argument works only if you assume that the Fool cannot deceive others. If he can, he is rational to profit at their expense. Therefore, morality cannot rest merely on selfish motives and rational choice. Morality requires a primitive disposition to take persons as equals, a quality that cannot be constructed out of the rational choice of non-tuistic agents (Hampton 1991). This disposition blocks the use of deception as a rational option, because deceiving others contradicts a sense of fairness. Deceit is a rational option for agents who lack social preferences, but only for them. In this sense, the Hobbesian, rational choice derivation of morality is mistaken. Morality cannot be derived from rationality if rational agents lack social preferences. Human cooperation rests necessarily on an irreducible and non-instrumental motive for fairness.

2.4 A Theory of Morality as Disguised Selfishness

The view that human cooperation is based on rational egoism appeared originally among philosophers, namely among the Greek Sophists as depicted by Plato in *The Republic*; and then in the *Leviathan* of Thomas Hobbes. It has been taken up by some evolutionary theorists of social behaviour, of whom I shall mention Richard Alexander (1987), who plays an important role in the argument to follow. It is important to confront and refute this view, if social preferences are to be advanced as evolved mechanisms necessary for cooperation, and not simply, if at all, as backup mechanisms to rational choice. Joyce (2006, p. 17) criticizes Alexander's view, but misrepresents it as a naïve conceptual confusion. Regarding Alexander's claim that selfishness is the motive for cooperation in humans, Joyce says: ‘But such attitudes, posing as hard-nosed realism, erroneously conflate distinct explanatory levels (see Tinbergen 1963). In particular, they commit the basic blunder of confusing the cause of a mental state with its content.’ (Joyce 2006, p. 17).

Joyce claims, as Tinbergen (1963) could have put it, that Alexander conflates the evolutionary with the proximate cause: if the social trait evolves because it brings fitness benefits to genes or individuals (evolutionary cause), then the mechanism driving it, in case it is an intention, *must be* the intention to produce a benefit to the individual or its genes, a selfish intention (proximate cause). But in fact, Alexander is not guilty of such naïve confusion. He is perfectly aware that selfish genes can produce psychologically altruistic motivations, for example between close kin (Sober and Wilson 1999, chap. 10, have spelled out an argument that shows how ‘selfish’ genes code for psychological altruism in parental care, namely because motivational altruism causes parental care more reliably than egoism, and parental care is crucial for reproductive success in humans). Alexander knows that selfish genes *do not have to be* expressed in selfish motivations in every case. He does think that they *are* so expressed in the relations between non-kin, but he thinks this on the basis of an argument. His reasons for believing that selfishness governs interactions between non-kin in large societies tally with those that convinced Hobbes, or the Smith of *The Wealth of Nations*. We cannot hold Smith or Hobbes guilty of a conceptual confusion involving proximate and evolutionary causes. In fact, Alexander’s theory is probably the strongest objection to the view that genuine social preferences are necessary for human cooperation. His theory is designed to explain away the philosophical claim that moral systems rest on social preferences or non-instrumental desires for the welfare of others. According to him, cooperation is possible among selfish agents making widespread use of deception and self-deception in their mutual interactions. This theory is perhaps his main reason for believing that our motivations are mainly selfish.

Alexander’s view relates neatly to the two characters already introduced: the Fool and the Sensible Knave. The core of these characters is to endorse morality and honesty only instrumentally. This was, for Hume at least, a distortion of the place that morality really occupies in the human mind, although he was prepared to accept the Knave as a rational character. In contrast to Hume, Hobbes believed that morality is instrumentally rational for selfish agents and that it is not rational to take advantage of others by deceiving or coercing them (Hobbes 1651). The end of the last section casts a doubt on Hobbes’s claim that the Fool is irrational. In any case, the claim about the Fool’s irrationality contrasts with Plato’s Sophists, who claimed precisely that justice and morality are instrumentally rational, but only as a lesser good, i.e., only in those cases when the better options of coercion or deception are not available. Deception and coercion provide rational individuals with their highest profit. This is, I think, also the best way to understand the views on morality put forward by Alexander. He demands that our views on morality adjust to the fact that humans have been shaped by natural selection to further their own individual reproductive interests (1987, pp. 34ff.), amidst the conflicts that inevitably arise with the interests of other members of society. Cooperation with others is only instrumental to self-benefits. This is evident, according to him, in the fact that everyone tries to shape cooperative interactions so as to profit more than their partners (pp. 102f, 109f.). In the same spirit, since moral systems are systems of indirect reciprocity and reputation is crucial, humans instrumentally cultivate a reputation as cooperative individuals (pp. 114, 191f.). Except, perhaps, for genuine psychological altruism between close kin, humans are psychological egoists focused only on their own benefit and cooperating only because coercion is not an option. In any case, they are always trying to influence and manipulate others so as to receive from them more than would be fair. This entails, of course, that humans will cheat if they can cheat undetected. In order for our deceitful natures to work effectively, natural selection has shaped us into self-deceivers as well (Alexander 1987, p. 123). We deceive ourselves into believing that we are good-natured, because if we did not, we would easily betray our deceitfulness and deceit would be ineffective. There is no genuine altruism; it is only a masquerade that we all play out so convincingly, that we have come to believe it ourselves. The philosophical idea that justice must be valued and is valued for its own sake (at least by some) is, according to Alexander, the product of self-deception.

Alexander's theory is an attempt to derive our deepest social-psychological nature from the ground-breaking theories on the evolution of social behaviour. These theories say that natural selection favours those genes that do their utmost to benefit their carriers. Alexander believes that this process has shaped our mind to direct all our behaviour, consciously or not, to the reproductive benefit of the agent. As stated above, he admits that humans can have non-instrumental desires for the welfare of close kin. But regarding non-kin, given the conflicts of interest that necessarily arise between them, all that humans need is the *appearance* of genuine altruism, although individuals may occasionally incur in large sacrifices for others, either by mistake or as victims of manipulation (pp. 104, 114, 191f.).

An interesting feature of his view is that he has a theory, inspired by Trivers (1971, 1985), for why our common-sense moral experience hides our basic selfishness. We deceive ourselves into believing that we are non-instrumentally interested in the welfare of others, so that we can better deceive them into the same belief (cf. the project for an evolutionary science of self-deception in Trivers 2011). The best way for unavoidable egoists to reap the benefits of cooperation is to build a reputation as fair players, because humans, both now and in our evolutionary past, value fairness over selfishness in social partners and prefer to interact with agents disposed to be fair. But if humans evolved to be motivationally selfish by biological design, there are no fair partners to choose from. Had this simple fact not been effectively concealed, large-scale human cooperation would never have evolved. The only chance for cooperation to evolve and prosper depended on the ability to conceal selfishness. Since conscious deceivers too often betray themselves involuntarily, deceiving others required deceiving oneself as well (Alexander 1987, p. 123). Thus, natural selection favoured the evolution of self-deception. A fundamental element in this self-concealment is that we denigrate selfishness and praise altruism in order to deceive ourselves and others into believing that we are, in fact, non-instrumentally (altruistically) interested in their welfare (p. 125). If we feel uneasy at a theory that assigns this role to deceit and self-deceit in human morality, this feeling is just another symptom of self-deception. The theory thus explains the fact that altruism is high ranked in our system of values, while denying that humans are, or could be, genuine altruists. Alexander defuses in this way the testimony of everyday moral intuitions against the troubling claim that we are all egoists pursuing only our own benefit and willing to deceive others when it pays off.

If Alexander's view of our social nature is true, human cooperation rests on fragile foundations and the often quoted phrase authored by Michael Ghiselin 'Scratch an altruist and watch a hypocrite bleed' would exactly capture the frail nature of human morality. As Plato's Sophists claimed, humans cooperate only when the prospects of cheating and getting away with it are faint, while silently lurking for the opportunities where those prospects increase and a ruthless pursuit of individual advantage promises to pay. Obviously, this is no flattering picture of who we are. The world therein depicted is not one where most of us feel at home. Alexander would surely reply that we feel troubled by it because self-deception makes us think we are better. But I think he misses one reason why it should be troubling: he seems to think that cooperation is guaranteed as a stable expression of our deceptive and self-deceptive selves, but he is probably mistaken. In so far as it is only an expression of self-deceit, human cooperation is a façade that hides a manipulative agenda and a struggle for power. If his theory is true, no matter how real cooperation may look like, the struggle for power beneath it is the real master that determines our fate. If this is the world we live in, pessimists are fully entitled to have gloomy views about our future survival as a species. This is no argument against the theory itself, but it is one against the claim that the theory allows us to hope for a better future, or a future at all.

Fortunately, this apparently well-constructed theory has a fatal logical flaw, a contradiction in the way it conceives the evolutionary scenario. The starting point for evolution of self-deception consists in agents approaching others not only with selfish motivations, but also with a disapproval of selfishness that threatens to make cooperation impossible. If we disapprove of selfishness and know that everybody is selfish, how would cooperation even get started? The answer seems to be: only by hiding selfishness through deception and self-deception. Notice that agents disapprove of selfishness and approve of altruism *before the evolution of self-deception*. Self-deception is explained as having evolved under a selective pressure for genuine fairness in partners. But this character trait, given how natural selection works, cannot exist. So it seems that we valued genuine fairness before the evolution of self-deception. However, Alexander also makes self-deception responsible for the fact that we value fairness, because it is part of the strategy of concealing selfishness from consciousness. In sum, we value fairness as the product of self-deception, but at the same time self-deception evolved because we valued fairness and needed to cooperate for survival. But you cannot have it both ways.[2] The theory may work as an explanation of self-deception, but as such it requires the previous and

[2] This criticism of Alexander's theory was first argued in Rosas (2004).

independent existence of our commonsense values. Therefore, we cannot explain the latter as an output of the self-deceptive mechanism. We need an alternative explanation that does not undermine our everyday praise of moral agents as a form of self-deception.

2.5 Cooperation in a World of Selfish Agents

Evolutionary models explain why a trait exists by observing its fate in a population where the trait in question competes with rival traits. Simulating the dynamics of a population where agents of different strategies (given by competing traits) interact is a way of discovering the fate of the trait. A payoff matrix gives the utilities for all possible outcomes of the interactions between strategies and the strategies reproduce in direct proportion to obtained utilities. In this paper, I describe the evolutionary dynamics of social traits in a population as a thought experiment about cooperative or exploitative interactions, without the simulation tools. First we picture a population of selfish agents alone under the assumption of perfect mind-reading abilities. Then we relax this condition; and finally we introduce agents with social preferences in competition with agents that are selfish and lack those preferences. The result of the thought experiment will tell us whether morality and cooperation can subsist or not, supported only on selfish motivations and rationality (where natural selection replaces rationality in these models). If the extinction of cooperation follows in a population of rational egoists without social preferences, this provides evidence for the view that a motivation for treating others with fairness (a basic social preference) is a necessary requirement for morality.

The first two thought experiments are designed to illuminate the relative contributions of deceit and selfishness in a theory where morality is instrumental and eventually disappears. They show that human deceitfulness is far more important than selfishness in producing this outcome. In order to see this, imagine first a hypothetical world consisting purely of selfish agents, where deceit is impossible because intentions are transparent to everybody. In this case, intentions to defect will instantaneously be known by others. In interactions with a PD structure, perfect mind reading induces defection in all those that interact with latent defectors. Thus, if everybody is rational and is intent only on their benefit, it is better to form the intention to cooperate and carry it through. This brings the benefit of mutual reward, whereas second thoughts about not complying would be read by partners and would induce them to defect, resulting in the payoff for mutual defection. Since in a PD mutual defection is worse for both players than mutual cooperation, transparent rational egoists would always choose mutual cooperation. This is, in essence, Gauthier's argument to derive moral constraint from instrumental rationality (Gauthier 1986), only that he argues with the assumption of translucency instead of transparency. I shall show below that the argument does not work with translucency. But we can nevertheless acknowledge that cooperation would be the natural outcome in a world of egoists with transparent intentions, under the assumption that everybody chooses what is best for them (everyone is rational). Selfishness can produce a perfect imitation of a moral world. Altruism would not exist, but neither would exploitation of others. Moreover, altruism would not be highly valued, nor would selfishness be disapproved of. Our common-sense system of values would not exist. The thought experiment shows us that our evaluative attitudes must result from the fact that we are not transparent, but translucent and sometimes opaque, and cannot always read correctly the intentions of others in cooperative interactions. Next, I show that in a world populated *only* by selfish agents who are not transparent, but rather translucent or opaque, cooperation is bound to disappear.

Translucency cannot support cooperation in a world of rational egoists. The crucial insight here is the link between translucency and the possibility of deceit. Imagine now that in the world of selfish, rational and transparent agents depicted in the previous paragraph, all agents change suddenly and become translucent. In contrast to transparency, translucency implies that agents are not infallible about others' intentions before action. There is a probability of misinterpreting their intentions or dispositions, although they have a better than random chance of correctly identifying them (this is Gauthier's definition of translucency). We can conceive this as a brute fact about the natural, involuntary signs of mental states through the body, without any deliberate conscious manipulation. Given this brute fact about the bodily expression of mental states, selfish agents can be expected to exploit translucency, 'engineering' misinterpretation in a specific direction: manipulate signs such that others believe to detect a cooperative intention where there is none. Translucency opens a door to the strategy of deceit, a development that Gauthier does not take properly into account (Sayre-McCord 1991). As agents gradually become better in deception, translucency is replaced with opacity or worse, because some agents succeed in putting a misleading appearance most of the time. Notice, however, that deceit does not make others believe in altruistic motives, for these do not exist in a world of selfish agents. Deceitful agents fake the intention to cooperate. This alters the probability of correctly identifying others' intentions, such that at some point it falls below random. At that point, it will no longer be true that cooperation with those you think will cooperate is the best move. Therefore, cooperation will likely disappear in a world of translucent agents that adopt deceit as a strategy. As soon as cooperation ceases, deceit loses its point. Deceit therefore, is the cause of the decline of cooperation. A disposition to denigrate both deceitfulness and selfishness could emerge along the process.

2.6 Fallible Mind Reading Makes Our Value System Emerge

In a world where agents are selfish and translucent, the evolutionary dynamics takes the population to a world where everybody uses deceit and fakes their character. In such a world full of deceit, the bare intention to cooperate is of no value. It is merely the strategic move of an agent trying to lure others into interaction only to eventually exploit them. Those intentions would exist only as long as others can be deceived, as a transitory remnant of a world where agents had been transparent. To be able to explain the emergence of our praise of fairness, we have to change the initial composition of the population. The population cannot consist solely of selfish agents. Imagine instead that agents with a non-instrumental disposition for fairness are present in an important proportion. If these agents exist and are known to exist, *intentions* to cooperate can be valuable, not as such, but as expressions of the correct *motivations and character*. Since everybody judges a genuine disposition for fairness to be better than its absence, everybody tries to choose partners with that disposition. Consequently, deceitful agents try to fake being fair. Therefore, genuine as opposed to fake dispositions for fairness turn out to be valuable. In a world without transparency, where selfish agents can hide their deceitful plot, interacting with genuinely fair agents is the only guarantee of successful cooperation. And since cooperation is a reliable path to many goods, humans highly value a disposition for fairness. Selfish-deceitful agents survive as parasites of a system of values that praises fairness.

This evolutionary scenario requires the existence of fairness as a character trait simultaneously with the possibility of discriminating it among a population of selfish and deceitful agents. This is not a problem if selection favours a genetic linkage of the character disposition and its recognition in a population. Biologists picture natural selection favouring this linkage in the context of interaction between kin (Hamilton 1964), and then extending it to interaction with non-relatives. Axelrod and Hamilton (1981) speculated that kin selection could evolve into reciprocal altruism if altruism is conditionally expressed towards kin and kin identification focuses on cooperative behaviour. Given that cooperative behaviour can be used as a deceptive lure, recognition must go deeper than mere behaviour, i.e., behaviour must be taken only as *one* channel to character in a multichannel recognition system. If agents disposed to fairness are able to recognize their kind, then the benefits of cooperation fall only or mainly on agents that have a disposition for fairness. The recognition of fairness in others becomes a condition for the expression of cooperative behaviour. Those recognition abilities are a necessary condition for the evolution and preservation of cooperation. If these abilities are absent, genuine fairness is an easy prey to selfish characters. Their presence gives agents a critical adaptive advantage. A positive feedback loop emerges between both traits (fairness and its recognition in others).

Fairness exerts self-constraint: the agent avoids the use of deception as a means to force others into the role of exploited suckers. It dictates a conditional, but reliable cooperative attitude. The recognition of fairness of character is of crucial importance for co-operators and for selfish agents alike, because it is the only reliable avenue to the benefits of cooperation. The high value attached to this character reflects the difference in fitness resulting from interacting with agents having it, compared to interacting with agents lacking it. The adaptive value of detecting and choosing fair partners leads to a genetic predisposition to develop a preference for fairness. This is of course, just a hypothesis, but it is important to have one where the origin of our valuing altruism does not present it as self-deceptive. Both moral and selfish agents value a moral character because it is a secure path to benefits. In contrast, selfishness becomes the target of aversive feelings because it is the vehicle for nasty, deceitful strategies of social interaction when mind reading is fallible. When deception is used in a social environment, choosing partners with a disposition for fairness is the best insurance against being deceived. It is therefore highly valued.

This model assumes that our mind-reading abilities, though imperfect, are accurate enough to support a strong psychological and social selection in favour of dispositions for fairness. This looks like translucency again, which we rejected above as an avenue for the evolution of cooperation. But recall that we rejected translucency in a world where all agents were selfish, where all have a strong reason to adopt deceit as a strategy as soon as any one adopts it. In this case, translucency can no longer guarantee that intentions will be correctly interpreted most of the time. The point here is that translucency cannot help the evolution of morality unless a substantial proportion of agents already have social preferences and thereby resist any temptation to manipulate translucency. Translucency cannot help if every agent adopts deception as their strategy, as rational egoists predictably would. Translucency will really help in the argument only when a substantial proportion of agents have a primitive disposition for fairness and will not adopt deceit as a strategy: their character revolts against it. They give honest signals of their character and thus create a chance for other agents to tap into objective differences between genuine and fake displays of dispositions for fairness. This raises the percentage of correct identifications. In this case, translucency is effectively equivalent to the idea that agents will have a greater than random chance of positively identifying moral or selfish dispositions when they are really there. The natural consequence is that a social-psychological selection for genuine fairness of character can take place in human evolution. In this respect, this informal model preserves one of the insights that make Trivers's defence of reciprocal altruism relevant for morality: in cooperative enterprises partners are judged and chosen in virtue of their motivational dispositions, which are the only reliable signs of a consistently cooperative attitude (Trivers 1971, pp. 50–51; see also Rosas 2007; Nesse 2007). In this way, the benefits of cooperation are circumscribed to those that are non-instrumentally interested in the welfare of others. I have here developed Trivers's idea by completing our picture of the selection pressures involved in the evolution of the preference for fair players. Being imperfect mind readers, we cannot place our trust on the rational egoism of our partners, but only on the fact, where it is a fact, that they have genuine dispositions for fairness.

References

Alexander, R. 1987. *The biology of moral systems*. New York: Aldine de Gruyter.

Axelrod, R., and W.D. Hamilton. 1981. The evolution of cooperation. *Science* 211: 1390–1396.

Bowles, S., and H. Gintis. 2011. *A cooperative species; Human reciprocity and its evolution*. Princeton/Oxford: Princeton University Press.

Buss, L. 1987. *The evolution of individuality*. Princeton: Princeton University Press.

Darwin, C. 1981 [1871]. *The descent of man and selection in relation to sex*. Princeton: Princeton University Press.

Fletcher, J.A., and M. Doebeli. 2009. A simple and general explanation for the evolution of altruism. *Proceedings of the Royal Society B: Biological Sciences* 276: 13–19.

Frank, R. 1988. *Passions within reason.* New York: W.W. Norton.

Gauthier, D. 1986. *Morals by agreement*. Oxford: Oxford University Press.

Gibson, M. 1977. Rationality. *Philosophy & Public Affairs* 6: 193–225.

Gintis, H., S. Bowles, R. Boyd, and E. Fehr. 2003. Explaining altruistic behavior in humans. *Evolution and Human Behavior* 24: 153–172.

Hamilton, W.D. 1964. The genetical evolution of social behavior I and II. *Journal of Theoretical Biology* 7(1964): 1–52.

Hampton, J. 1991. Two faces of contractarian thought. In *Contractarianism and rational choice*, ed. P. Vallentyne, 31–55. New York: Cambridge University Press.

Hobbes, T. 1651. *Leviathan*, ed. C.B. Macpherson. London: Penguin.

Hume, D. 1777, 1902. An enquiry concerning the principles of morals. In *Enquiries concerning the human understanding and concerning the principles of morals*, ed. L. Selby-Bigge, 169–285. Oxford: Oxford University Press.

Joyce, R. 2006. *The evolution of morality*. Cambridge, MA/London: MIT Press.

Lehmann, L., and L. Keller. 2006. The evolution of cooperation and altruism – A general framework and a classification of models. *Journal of Evolutionary Biology* 19: 1365–1376.

Maynard Smith, J., and E. Szathmáry. 1997. *The major transitions in evolution*. New York: Oxford University Press.

Michod, R.E. 2007. Evolution of individuality during the transition from unicellular to multicellular life. *Proceedings of the National Academy of Sciences* 104(S1): 8613–8618.

Nesse, R.M. 2007. Runaway social selection for displays of partner value and altruism. *Biological Theory* 2(2): 143–155.

Rosas, A. 2004. Mindreading, deception & the evolution of Kantian moral agents. *Journal for the Theory of Social Behaviour* 34(2): 127–139.

Rosas, A. 2007. Beyond the sociobiological dilemma. Social emotions and the evolution of morality. *Zygon* 42(3): 685–699.

Rosas, A. 2010. Beyond inclusive fitness? On a simple and general explanation for the evolution of altruism. *Philosophy and Theory in Biology* 2: e104.

Rosas, A. 2012. Disentangling social preferences from group selection. *Biological Theory*. doi:10.1007/s13752-012-0013-y.

Sachs, J.L., U.G. Mueller, T.P. Wilcox, and J.J. Bull. 2004. The evolution of cooperation. *The Quarterly Review of Biology* 79: 135–160.

Sayre-McCord, G. 1991. Deception and reasons to be moral. In *Contractarianism and rational choice*, ed. P. Vallentyne, 181–195. New York: Cambridge University Press.

Sober, E., and D.S. Wilson. 1999. *Unto others: The evolution and psychology of unselfish behavior*. Harvard: Harvard University Press.

Tinbergen, N. 1963. On the aims and methods of ethology. *Zeitschrift für Tierpsychologie* 20: 410–433.

Trivers, R. 1971. The evolution of reciprocal altruism. *The Quarterly Review of Biology* 46: 35–57.

Trivers, R. 1985. *Social evolution*. Menlo Park: The Benjamin/Cummings Publishing Company.

Trivers, R. 2011. *The folly of fools. The logic of deceit and self-deception in human life*. New York: Basic Books.

West, S.A., A.S. Griffin, and A. Gardner. 2007. Social semantics: Altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology* 20: 415–432.

Chapter 3
Two Problems of Cooperation

Katharine Browne

3.1 Introduction

Nature is rife with examples of cooperative behaviours. From honeybee societies and altruistic vampire bats to the enormously complex civilizations formed by human beings, cooperation lies at the foundation of all social interactions.

Finding examples of cooperation is not difficult. Explaining how it is possible, however, is less straightforward. The prevalence of cooperation in nature is a bit of a mystery, in light of evolutionary theory. The reason is that cooperation (as it is sometimes characterized) involves a fitness cost to its actor, a characteristic that is difficult to make sense of given that nature selects against fitness-decreasing traits. There is an analogous problem in normative moral theory. Cooperative behaviour (as it will be understood here) requires that individuals constrain self-interested pursuits. But 'rationality' (as it is most commonly understood) requires that agents act as selfish utility-maximizers. It thus seems that cooperative behaviour is at odds with rational behaviour.

This paper has two central aims. The first is to outline these two contextually different problems of cooperation, and provide solutions to each. My primary concern in the biological context will be with describing the possibility of *human* cooperation. What is distinctive about human cooperation is its scale and scope. And these characteristics make it particularly difficult to reconcile with the usual evolutionary explanatory mechanisms. Given that human cooperation extends beyond kin and occurs on a very large scale, the descriptive account will explore what kinds of mechanisms might support cooperation, and whether there is an evolutionary story that can accommodate these. I will defend an explanatory account of the emergence of cooperative behaviours that appeals to cultural group selection.

K. Browne (✉)
Centre for the Study of Mind in Nature (CSMN), University of Oslo, Oslo, Norway
e-mail: katharine.browne@csmn.uio.no

B. Musschenga and A. van Harskamp (eds.), *What Makes Us Moral? On the capacities and conditions for being moral*, Library of Ethics and Applied Philosophy 31,
DOI 10.1007/978-94-007-6343-2_3, © Springer Science+Business Media Dordrecht 2013

In the normative context, I will defend the rather unpopular view of the rationality of adopting the ‘cooperative’ disposition to constrained maximization, as articulated by David Gauthier. This defence requires two moves. The first is to show that the disposition to constrained maximization will yield a greater utility than an alternative disposition. The second is to make the case that the rationality of a disposition entails the rationality of the actions that the disposition recommends. I argue that Gauthier succeeds in both. And if so, we will have a reconciliation between cooperation and rationality.

The second aim of this paper will be to outline how the descriptive and normative projects are connected. One way in which the two problems of cooperation are related is through the social contract tradition. Social contract theory traces moral and political obligations to a contract. When individuals need to band together in cooperative ways, they agree to a set of principles that will govern their social interactions. These principles set the terms for cooperation. Within the contract tradition there is a descriptive branch and a normative branch. Descriptive approaches describe the origin of the social contract and seek to explain how cooperation occurs. The first problem of cooperation identified above falls to this branch. Normative approaches to the contract, by contrast, aim to justify the terms of the contract. Our second problem of cooperation falls to this branch. I will argue that there is a convergence between the descriptive and normative strands, specifically between what evolution produces and reason recommends. I will show that the cultural group selection explanation of the emergence of cooperation provides an explanation of the emergence of dispositions that resemble those that Gauthier defends as rational. This is significant for two reasons. First, it results from a combination of the descriptive and normative questions in a way not found in current literature. It is not all that rare to see some appeal made to evolutionary processes in the philosophical literature, but little attention has been paid thus far to the import of the cultural evolutionary story to moral theory in general and normative contract theory in particular.

Second, we will see that there is a unique structure to the outcome of the cultural analysis of human sociality that blends particularly well with Gauthier’s defence of constrained maximization. The prevalence of cooperation, I will argue, depends on the presence of prosocial dispositions that call for a constraint on the pursuit of self-interest. Rationality calls for the formation of the disposition to constrained maximization, which likewise requires similar constraints on the pursuit of self-interest. These results set the stage for a further examination into the connection between reason, evolution and morality. I will end by pointing to some possible directions in which we can flesh out these equivalences more fully.

3.2 What Is Cooperation?

‘Cooperation’ has distinct meanings in biological and moral contexts and varying meaning within each context. As such, it will first be important to clarify what we mean by the term. In ordinary language cooperation refers to any coordinated, mutually beneficial behaviour. But it is commonly used to mean something different in evolutionary biology and moral theory.

In evolutionary biology, the terms ‘cooperation’ and ‘altruism’ are generally used interchangeably, and understood to refer to behaviour that is costly to the individual and beneficial to others. Peter Richerson and Robert Boyd say that they ‘use the word *cooperation* to mean costly behavior performed by one individual that increases the payoff of others. This usage is typical in game theory, and common, but by no means universal, in evolutionary biology’ (Boyd and Richerson 2006, p. 454). This equates cooperation and altruism, and Elliott Sober and David Sloan Wilson explicitly endorse that equivalence. They say, ‘prevalent among game theorists is their use of the word *cooperation* rather than *altruism* . . . the word *cooperation* is used by evolutionary game theorists, presumably because it is easier to think of cooperation as a form of self-interest. The behavior is the same but it is labeled differently’ (Sober and Wilson 1998, p. 84). In the moral context, David Gauthier (1979) uses ‘cooperation’ to describe the behaviour required by a particular subset of morality, namely that of distributive justice. He distinguishes two parts of morality: ‘distributive justice’ and ‘acquisitive justice’. The first ‘constrains the modes of cooperation’; the second ‘constrains the baseline from which cooperation proceeds’. Our concern in this paper will be with the former part of morality having to do with the emergence and maintenance of cooperation and the question why rational individuals should cooperate.

Thus, in very general terms, ‘cooperation’ can be understood as a type of behaviour that involves some kind of constraint on individual interest, where ‘individual interest’ can be understood in terms of biological fitness or rational self-interest. I will use the terms ‘cooperation,’ ‘altruism,’ and ‘morality’ (i.e., the subset of morality Gauthier identifies) interchangeably. In doing this I blur finer distinctions between them. Nonetheless, this allows me to conveniently talk about the central problems of evolution and rationality outlined above, viz., ‘How is it possible for organisms to act in such a way as to lower their fitness?’ and ‘How it is possible for rational beings to act contrary to their self-interest?’

3.3 The Descriptive Problem

The first aim of this paper is to address the two problems of cooperation outlined above. This involves answering two questions. The first is why cooperative behaviour is so prevalent in nature. This is largely a descriptive or explanatory question. The second is whether rationality dictates that individuals ought to cooperate. This is primarily a justificatory question. With respect to the descriptive question, explanations of cooperation can be one of two types. The first is to explain the behaviour in terms of its proximate mechanisms. Explanations of this sort will usually appeal to the underlying psychological mechanisms responsible for that behaviour. The second explanation appeals to ultimate causes. Explanations of this sort will generally make reference to the ultimate evolutionary mechanisms that produce such behaviour.

We might leave a tip in a foreign restaurant because to do otherwise would leave one feeling guilty. Or we might help a stranger in need because their plight elicits in us pangs of empathy. These psychological affectations are the proximate mechanisms of said behaviour. The ultimate explanation, on the other hand, will appeal to the evolutionary benefits that having such dispositions has for the one who possesses them. We might say, for example, that being disposed to feeling guilty for transgressions of norms pertaining to, say, tipping in foreign restaurants, leaves one more biologically fit than another who is not so disposed.

Thus, explaining the emergence of cooperation requires addressing what psychological mechanisms underlie that behaviour, and what evolutionary processes produce those mechanisms. But providing these explanations is not straightforward. The central difficulty is to articulate the mechanisms required to support cooperative behaviour in a way that is amenable to what is known about how evolutionary processes work. Given both the general structure of cooperative behaviour – viz., that it imposes a cost on its actor and benefits the recipient – and of evolutionary processes that presumably require any costs associated with a particular behaviour to be recouped if that behaviour is to persist, there is a tension between the explanandum and explanans. Since cooperation involves an individual's sacrifice of his or her reproductive fitness in order to enhance that of another individual, and since natural selection works against fitness-decreasing characteristics, there is a difficulty in reconciling the two.

Two standard mechanisms invoked to explain cases of cooperative behaviour are kin selection and reciprocal altruism. Parents act in ways that appear to directly reduce their own fitness and promote the fitness of their offspring. Kin selection provides an explanation why. Since offspring contain on average one half of the genes carried by either parent, caring for offspring, which is itself disadvantageous to the individual, is a behaviour that promotes the survival of one's genes. We can invoke kin selection to explain other behaviours that appear detrimental to the one performing them. Some birds display warning calls, which renders the individual making the warning call more likely to be picked out by a predator than the birds who are warned. Again, we have here an instance of an apparently fitness-decreasing behaviour, leaving one to ask how that could have evolved. The answer is this. Birds who display warning calls will live in groups of closely related kin. While the individual who warns others may be more likely to fall prey to a predator, his call promises to save many of his kin who share his genes. Thus, this seemingly fitness-decreasing behaviour is rendered compatible with natural selection, since this behaviour is actually one that promotes the survival of one's genes.

Reciprocal altruism permits the same kind of explanation in cases where individuals are not related. Nit-picking in birds, grooming among chimpanzees, and sharing of blood among vampire bats are all examples of apparently altruistic behaviours. On closer inspection, we see that the apparent fitness costs imposed by these acts are recouped through reciprocity and can thus be made sense of within an evolutionary framework. But while it is fairly widely accepted that altruistic behaviour in non-human animals can be explained through the evolutionary mechanisms of kin selection and reciprocal altruism, it is not obvious that these mechanisms are sufficient to explain the large-scale cooperation among non-related individuals common in human societies. Certainly in some cases kin selection and reciprocity come into play. But when individuals are unrelated, we can rule out kin selection. And as group size increases, it becomes more difficult to keep track of one's past cooperative partners, and thus reciprocal altruism no longer seems a plausible candidate to explain cooperation on large scales.

There are a number of more sophisticated models of reciprocity that ultimately hinge on 'enlightened self-interest' to explain the emergence of cooperation, but these too all fall short. Indirect reciprocity is one such mechanism, which relies on reputation to determine one's partners. Other proposed mechanisms, like costly signaling, involve the display of fitness-decreasing traits that signals prestige or strength, and consequently enhance individual reproductive advantage.[1] Thus, the cost of the signal is recouped by it granting greater access to mates. But mechanisms like these that appeal to individual advantage are sometimes difficult to square with observed behaviour. We will often do things such as leave a tip in a foreign restaurant, warn drivers that they left their car lights on, return found items, draw attention to being undercharged, and so on, where the preservation of reputation (or any kind of 'individual advantage') is not a plausible explanation. And as Richerson et al. ask: 'If a mechanism like indirect reciprocity works, why have not many social species used it to extend their range of cooperation?' (Richerson et al. 2003, p. 379). Individuals seem to perform actions that appear to be genuinely fitness-decreasing and we must now ask whether this behaviour can be rendered compatible with the standard tenets of evolutionary processes.

Sober and Wilson (1998) think that they can be, and appeal to group selection to explain how. According to them, individual-level explanations of the emergence of altruism require that the cost of altruism be recouped elsewhere and, thus, fail to provide an explanation of *genuine* altruism. That is, parental care, or helping one's neighbour, are really instances of selfish behaviour: these are instances of behaviour that benefit the individuals themselves or their genes. Sober and Wilson contend that altruism – *genuine* altruism, i.e., behaviour done for others that lowers the individual's fitness without genetic or other recouping – can be explained in terms of group selection. According to the group selection hypothesis, some behaviours or traits evolved, not because they were advantageous to particular individuals, but because members of groups containing those traits did better than members of groups that did not. While individual selection will favour the evolution of selfishness within groups, and thus altruists will be less fit relative to non-altruists within a single group, matters are different at the level of groups. Altruists will do, on average, worse than selfish individuals within the same group. But, so the argument goes, members of groups of altruists will do better than will members of selfish groups. And if we grant that selection can occur at the level of groups, then

[1] An example of this is the peacock's tail: the ornate tail is certainly a hindrance but it does serve as a signal that the peacock is strong and healthy enough to survive and preserve its nice tail and, thus, is likely a good selection for a mate.

we can explain the existence of altruism in terms of it: altruism evolved because members of altruistic groups did better than members of non-altruistic groups. Group selection thus promises an explanation of the emergence of behaviours or traits that genuinely reduce the fitness of an individual within a particular group, so long as those behaviours or traits increase the fitness of the members of the group that contains them relative to members of groups that do not.

But how plausible is this? Let us say that for every recipient of an altruistic act, an individual gains 2 Darwinian fitness points. Every altruist loses 1 fitness point for every act of altruism she performs. When two altruists meet, each gains 2 but loses 1. When an altruist meets a selfish individual, she loses 1 and confers a benefit of 2 onto the selfish individual. When two selfish individuals meet, neither gains and each receives a payoff of 0. Of two populations, one containing all altruists, the other containing all selfish individuals, the population of altruists will be fitter than the selfish group, since altruists each receive 1 point when they encounter one another, and selfish individuals receive nothing. If we translate these fitness points to number of offspring, we can see that the proportion of altruists in the global population (viz., the population resulting from the combination of the two groups) will increase.

However, in a population where both altruists and selfish individuals are present, altruists will be at a substantial selective disadvantage. In groups containing both altruists and selfish individuals, even if altruists make members of the group that contains them more fit than members of the group that does not, within the group, altruists will be less fit than non-altruists. Selfish individuals will have higher levels of relative fitness than will altruists of the same group, and will thus have more offspring than altruists. Selfish individuals will prosper, and the number of altruists within the population will decline.

Let us suppose that a population consists of two individuals, Anne and Sam. Anne is an altruist; Sam is a selfish type. As an altruist, Anne will behave in ways that reduce her own fitness and in ways that promote the fitness of those around her. Consequently, Anne will have only one offspring (let us suppose she does so asexually); while the selfish recipients of her altruistic acts will have two. Sam, as a selfish type, will not engage in any fitness-detrimental behaviour but will benefit from the altruistic actions of others. Sam will have two offspring. In the next generation, Anne's offspring, as an altruist like her mother, will have only one offspring. Sam's two offspring, on the other hand, will each have two offspring. Thus, after one generation the population will consist of three altruists (Anne, her offspring, and her offspring's offspring), and seven selfish types (Sam, Sam's two offspring, and two offspring each of Sam's two offspring). In the next generation, the number of altruists relative to selfish types will decline even further. And so on. Thus is appears that so long as selfish types are present in a population, altruists should tend towards extinction. Populations ought to contain very few altruists relative to selfish individuals. But this is not the case, and invites the question why not.

Sober and Wilson appeal to Simpson's Paradox to explain how the evolution of altruism in mixed populations like the above is possible. Simpson's Paradox 'refers to the phenomenon whereby an event C increases the probability of E

in a given population *p* and, at the same time, decreases the probability of *E* in every subpopulation of *p*' (Pearl 2000, p. 1). Sober and Wilson illustrate Simpson's Paradox with an example of a discrimination inquiry at the University of California Berkeley.[2] Based on the smaller overall percentage of women who were admitted than that of men, it was suggested that the University's admission policies were discriminatory. Upon further investigation, however, it was discovered that every department admitted an equal proportion of women to men. And yet fewer women, overall, were admitted than men.

This seemingly paradoxical result can be explained as follows. It was discovered that a greater number of women tended to apply to departments that had lower acceptance rates than those to which men tended to apply. Let us say that department A accepts only 25 % of its applicants, while department B accepts 75 %. Suppose that department A receives the following distribution of applicants: 80 women and 20 men. Department A accepts 25 % of the women applicants and 25 % of the men applicants: 20 women and 5 men. Suppose now that 20 women and 80 men apply to department B. Since department B accepts 75 % of their applicants, the result is that 15 women and 60 men are accepted. A combined total of 100 men and 100 women applied to the two departments. Only 35 women in total were admitted, while 65 men were admitted. And yet, no department was discriminatory in their acceptance rates, since each department accepted an equal proportion of men to women.

Sober and Wilson employ Simpson's Paradox to reveal that it doesn't necessarily follow that, based on the fact that altruists will be less fit in a subgroup than selfish individuals, they will necessarily be less fit overall.[3] While selfish individuals may do better within groups than altruists, if two groups are combined, the reverse effect can be obtained. Thus, 'what is true, by definition, is that altruists are less fit than selfish individuals *in the same group* . . . however, nothing follows from this as to whether altruists have lower fitness when one averages *across all groups*' (Sober and Wilson 2000, pp. 190–191).

We thus get a sketch of the kind of contribution group selection can make to explaining the emergence of traits that appear to perform to the detriment of their possessors and to the advantage of the group in which they occur. Individual selection may favour the evolution of selfishness within groups (and thus altruists will be less fit relative to non-altruists within a single group), but members of altruistic groups will do better than members of selfish groups.

[2] This example has its origin in Cartwright (1979), pp. 419–437.

[3] There is some controversy over how to interpret Simpson's Paradox and also whether Sober and Wilson employ it appropriately for their purposes. A fuller discussion of these issues lies beyond the scope of my aim here. According to Nancy Cartwright (1979) and Judea Pearl (2000), it is the fact that there is a correlation between being a woman and applying to departments with lower acceptance rates that is responsible for the statistical anomaly in the Berkeley case. Pearl says: 'In the case of Simpson's paradox, we have a clash (i) between the assumption that causal relationships are governed by laws of probability calculus and (ii) the set of implicit assumptions that drive our causal intuitions' (Pearl 2000, p. 7). It's not clear that Sober and Wilson recognize this. On this, see further: Peter Gildenhuys (2003), p. 36.

The concept of group selection is not new and was invoked by Darwin himself to explain human moral behaviour:

> It must not be forgotten that although a high standard of morality gives but a slight or no advantage to each individual man and his children over the other men of the same tribe, yet that an increase in the number of well-endowed men and advancement in the standard of morality will certainly give an immense advantage to one tribe over another. (Darwin 1871, p. 166)

But while the promise to explain the emergence of genuinely altruistic behaviours is attractive, the plausibility of the theory is highly contentious. Perhaps the most pressing problem facing group selection is the unlikelihood of the conditions required to make it work. Sober and Wilson maintain that the evolution of altruism is an outcome of a conflict between two competing processes: individual selection within groups and group selection between groups. For altruism to evolve by group selection, altruists must be distributed in populations in such a way that groups containing altruists will increase at a more rapid rate than groups that contain selfish individuals and, thus, the global population of altruists will grow in spite of their being disadvantaged at the individual level. Maynard Smith (1964) produced the Haystack model, which modelled these conditions. They are as follows. (1) Groups must be isolated and sufficiently varied; otherwise the effects of group selection will be negated. (2) Groups must divide and intermix at just the right time to permit differential reproduction of altruism and selfishness and also to prevent individual selection from driving altruism to extinction. (3) Groups must then re-isolate themselves, and the process must then be repeated. Maynard Smith was sceptical that these conditions would actually obtain in nature, and thus that group selection played much of a role in the evolution of altruistic tendencies. More specifically, what is missing from the biological account is the plausibility that genetic differences will be distributed in such a way that there is sufficient variation between groups and sufficient similarity within groups.

The evolution of altruism by group selection thus requires that altruists be aggregated within and between groups in a way that is unlikely to occur at the biological level in human groups. As Richerson and Boyd say, ‘The trouble with a straightforward group selection hypothesis is our mating system. We do not build up the concentrations of intrademic relatedness like social insects, and few demic boundaries are without considerable intermarriage’ (Richerson and Boyd 1998, p. 81). And further, ‘Even very small amounts of migration are sufficient to reduce the genetic variation between groups to such a low level that group selection is not important’ (Richerson and Boyd 2006, p. 463). If so, to the extent to which variation within and similarity between biological groups is needed for altruism to evolve, biological group selection is unlikely to have played a significant role in the evolution of altruism in humans.

But while group selection at the biological level is unlikely a strong evolutionary force due to the implausibility that genetic differences will be distributed in such a way that there is sufficient variation between groups and sufficient similarity within groups, this does not hold in the case of cultural variants. Richerson and Boyd contend that cultural variants in the form of social norms will help to generate suitable variation between groups and homogeneity within groups to permit group selection at the cultural level. Some groups will operate with more successful norms than others and, thus, will out-compete groups operating with less successful norms, and as those groups flourish so will their norms.

According to Richerson and Boyd, ‘Culture is information capable of affecting individuals’ behaviour that they acquire from other members of their species through teaching, imitation, and other forms of social transmission’ (Richerson and Boyd 2005, p. 5). This can range over particular beliefs (e.g., belief in God), skills (e.g., tool use, technological innovations), behaviours (e.g., washing food prior to eating it, removing one’s hat when indoors), or social norms (e.g., pay taxes, obey water restrictions, take only your fair share of resources, etc.). Richerson and Boyd argue that cultural adaptation has resulted in significant behavioural differences between groups. Culture permits individuals to adapt quickly to environmental changes. Consequently, individuals are very locally adapted to a wide range of environments, which has resulted in significant variations in behaviours between human groups. And the conformist tendencies of humans, together with intolerance of differences, tend to keep groups uniform. These differences between groups, together with competition between groups, set the stage for group selection.

Group selection results in the selection of group-beneficial traits. Groups that operate more advantageous norms will out-compete those groups that operate less advantageous norms. Thus when groups with different cultures with differential fitnesses come into conflict, and one wins out over the other, the culture of the winning group grows, and the losing group either is extinguished or absorbed. On their own, these norms might permit societies operating with them to function satisfactorily. But when they are operating within the larger structure in the world, they are being outcompeted by other social structures. And Richerson and Boyd, following Darwin, think that cooperative groups will tend to out-compete non-cooperative groups, and thus a cooperative culture will take root and grow.

Thus, cultural evolution generates differences between groups. These differences make for an environment conducive to group selection. More cooperative groups do better than non-cooperative groups. And since, according to Richerson and Boyd, genes and culture co-evolve (Richerson and Boyd 2005, pp. 191–236), and since a cooperative culture will make groups more successful than groups that do not employ cooperative norms, individuals who have dispositions towards such cooperative norms will do better than individuals without them. An environment is thus created that is conducive to prosocial dispositions. These include an ability to internalize and conform to norms and the capacity for feelings of guilt and shame, and are dispositions that increase the chance that norms are followed. The emergence of prosocial dispositions thus feeds back into the support and maintenance of cooperation.

We now have a promising solution to our first puzzle: group selection on cultural variants provides us with a plausible evolutionary account of the existence of cooperation in human beings. This, however, is not the only possible solution. Its rival is the so-called ‘Big Mistake Hypothesis.’ On this view, the social dispositions that we have are left over from earlier times when human groups were composed largely of closely related kin. On this view, our psychological dispositions towards cooperative behaviour are once-adaptive responses to an environment in which we no longer find ourselves, and evolution hasn't had a chance to adapt to modern social group composition. The widespread cooperation we find is thus a 'Big Mistake'.

The Big Mistake Hypothesis is a coherent story and does appear to explain widespread cooperation that extends beyond immediate kin ties and reciprocal relationships. Confirming or denying the validity of this hypothesis requires in part a fuller investigation into the history of human sociality to uncover at what point early human social structures began to diverge from kin-based clan and at what point maladaptive traits (like prosocial dispositions, according to the Big Mistake Hypothesis) emerged (Richerson and Boyd 2005, pp. 188–189). Evaluating the success of this account lies beyond the scope of my aim here and is unsettled in the literature. But even if the Big Mistake Hypothesis were true, that fact alone would not preclude that culture has played a significant role in our evolutionary history and shaped at least part of human social behaviour.

3.4 The Normative Problem

I now move on to the second problem of cooperation, namely reconciling cooperation and rationality. David Gauthier has claimed that, 'the reconciliation of morality and rationality is the central problem of modern moral philosophy' (Gauthier 1990, p. 150). Given that rationality requires the pursuit of one's self-interest, and morality (or, in our case, cooperation) constrains the pursuit of individual interests, it seems that moral behaviour is irrational.

The Prisoner's Dilemma provides a formalization of the problem of cooperation. It involves two accomplices who are caught for committing a crime, interrogated separately, and offered a deal. If one player incriminates the other, or 'defects', while the second remains silent, or 'cooperates', he will be given a sentence of 1 year, while the other player will get four. If both remain silent, both will be sentenced to 2 years, but if both defect, both will receive 3 years. The following matrix represents this game.

		Prisoner 2	
		Cooperate	Defect
Prisoner 1	Cooperate	2 years, 2 years	4 years, 1 year
	Defect	1 year, 4 years	3 years, 3 years

The first number of each pair represents Prisoner 1's possible outcomes; the second number Prisoner 2's. In this particular case, no matter what the other player does, defecting is the utility-maximizing response. However, given that each player is rational, both will employ this equilibrium strategy, which will lead to a situation that is less preferred to the one where both cooperate. Rationality thus sometimes leads players to a suboptimal outcome.

Those who think cooperation can be reconciled with rationality will point to the fact that mutual cooperation will yield a higher utility than will mutual defection. But proponents of a strictly maximizing conception of rationality will contend that in the Prisoner's Dilemma to cooperate is a dominated strategy (that is, no matter what one's opponent does, defection is always the best reply in terms of utility maximization) and is thus positively irrational.

Hobbes's Foole takes this line, and asks why one cannot violate the rules of morality in cases where doing so is advantageous.

> The Foole hath sayd in his heart, there is no such thing as Justice; and sometimes also with his tongue; seriously alleaging , that every mans conservation, and contentment, being committed to his own care, there could be no reason, why every man might not do what he thought conduced thereunto: and therefore also to make, or not make; keep, or not keep Covenants, was not against Reason, when it conduced to ones benefit. (Hobbes 1651, ch. 15, par. 4, 74)

The Foole's objection points to a structural problem in the Prisoner's Dilemma: it will always be to one's benefit to violate one's agreement. And insofar as one is rational to the extent that one pursues one's benefit, the rational course of action will always be to defect.

In reply to the Foole, Hobbes argued that violations of morality are liable to be detected and punished, and that the consequences of being caught – that is, being excluded from civil society – are so grave that defection is never a good gamble. This reply in effect changes the payoff structure of the Prisoner's Dilemma such that unilateral defection carries with it grave consequences rather than rich rewards. But in order for this reply to be successful, the highly unlikely state of affairs must obtain where, for every single potential defection, the risks and possible costs associated with defection outweigh any possible gains. Hobbes's ultimate solution to the problem of non-compliance is a political one: to have a sovereign with sufficient power of surveillance and authority to punish so as to make non-compliance counterproductive. This, however, can be costly and inefficient, and it would be desirable if compliance could be achieved by more efficient non-coercive means.

Gauthier presents us with such a means. He locates the rationality of compliance with agreements in the adoption of the disposition he refers to as 'constrained maximization'. Constrained maximizers conditionally dispose themselves to cooperation. This disposition to cooperate distinguishes the constrained maximizer from what Gauthier refers to as a 'straightforward maximizer', who 'seeks to maximize his utility given the strategies of those with whom he interacts' (Gauthier 1986, p. 167). Straightforward maximizers (like the Foole) are rational utility-maximizers; they will defect when it is advantageous for them to do so. The constrained maximizer, by contrast, will cooperate when he expects others to cooperate and defect only when he anticipates that others will do the same.

The underlying presupposition of constrained maximization is the rationality of constraining utility-maximization in order to gain a mutually optimal strategy. The constrained maximizer disposes himself, essentially, to forgo token opportunities to make big gains through defection in order to obtain the benefits of mutual cooperation. Gauthier must then show that doing so is rational. This requires showing that the gains through mutual cooperation outweigh gains through defection, a conclusion that is not obvious given the structure of interactions between the two types of maximizers.

In a society composed of a mix of straightforward and constrained maximizers, individuals will defect in all but two cases. First, two constrained maximizers interact and are able to identify each other as constrained maximizers. In such a case, both will cooperate. Second, a constrained maximizer mistakes a straightforward maximizer for a constrained maximizer. In such a case, the constrained maximizer will adhere to the agreement, and the straightforward maximizer will defect.

In Prisoner's Dilemmas the most utility-maximizing strategy is unilateral defection. If the straightforward maximizer is able to fool the constrained maximizer into thinking that he is a constrained maximizer, then he will be able to gain at the expense of the constrained maximizer. This will yield the best outcome for the straightforward maximizer and the worst for the constrained maximizer. The second best outcome for both is mutual adherence, and the third best for both is mutual defection.

When neither party cooperates, both the straightforward and the constrained maximizer will receive the third best payoff. In an encounter between a straightforward maximizer and a constrained maximizer, if the constrained maximizer adheres to the agreement while the straightforward maximizer defects, the straightforward maximizer will receive a payoff greater than he would had he adhered. The straightforward maximizer is thus able to reap advantages unavailable to the constrained maximizer through unilateral defection. Although the constrained maximizer has available to him opportunities for gain through mutual adherence, mutual adherence yields a utility less than does unilateral defection.

In order to avoid the impending conclusion that rational individuals ought always to defect, Gauthier must rely on the assumption that straightforward maximizers cannot pass as constrained maximizers. For, given that straightforward maximizers will do better than constrained maximizers as long as they are given the same opportunities as constrained maximizers, and whether one is able to partake in agreements with others will depend on whether it appears that one can be trusted, so long as one is able to maintain the illusion of being a constrained maximizer, one will be able to partake in agreements with others while reaping the benefits of defection. Straightforward maximizers thus will do better than constrained maximizers, which will prevent one from claiming that constrained maximization is rational.

Gauthier recognizes how crucial the detection of dispositions and intentions is. If people were (to use Gauthier's terminology) *transparent* (i.e., their characters were always accurately detectable by others), then constrained maximization would be the rational strategy. For in that case constrained maximizers would be able to identify straightforward maximizers, and consequently exclude them from agreements. If people were, on the other hand, what Gauthier describes as *opaque* (i.e., their characters remained hidden to others), then straightforward maximizers would do better. For straightforward maximizers would then be able to continue to make agreements with others and gain through successful exploitation. Gauthier argues that neither of these is realistic, and claims that people are *translucent*, according to which 'persons are neither transparent nor opaque, so that their disposition to co-operate or not may be ascertained by others, not with certainty, but as more than mere guesswork' (Gauthier 1986, p. 174). In other words, if our characters are translucent, it is assumed that one has a better chance of correctly identifying another's character than one would by randomly guessing.

If so, and if one who is suspected to be a straightforward maximizer will be excluded from cooperative agreements, then the straightforward maximizer will not reap all projected benefits of both cooperation and defection. The straightforward maximizer will forgo many benefits of cooperation and reap only those gains that he can through successful exploitation. As his untrustworthy character becomes more widely known, opportunities for exploitation will diminish. And since constrained maximization affords one the possibility of gaining through mutually-beneficial co-operative interactions –opportunities unavailable to the straightforward maximizer – it becomes plausible to suggest that constrained maximization will yield a higher utility to those who so dispose themselves than will straightforward maximization. Gauthier thus concludes that rational persons will become constrained maximizers (Gauthier 1986, p. 128). And if Gauthier's defence of constrained maximization is successful, then he will have provided an account that reconciles cooperative behaviour with rationality.

The success of Gauthier's reconciliation project depends primarily on the answers to two questions. The first is whether constrained maximization is rational. The second is whether acting on that disposition is rational.

In order to establish the rationality of constrained maximization, Gauthier must argue that those who develop that disposition will do better than those who do not. This rests primarily on the plausibility of translucency. This seems to be largely an empirical matter. There is evidence on both sides. Gauthier could point to physiological reactions that accompany deceit – rapid heartbeat, flushing, aversion of eye contact, and so on, to support the view of the detectability of dispositions. He could also say that we can determine the dispositions of others in more impersonal ways by examining the history of their behaviour. McClennen writes:

> As it turns out, the iterated game framework provides a setting in which the epistemological problem of assurance can be resolved. If interaction is sufficiently on-going, then for many kinds of encounters, a given individual can have the requisite assurance regarding the disposition of other participants. The history of past encounters between participants typically will provide the needed information. It is plausible to suppose, moreover, that for many such contexts, the requisite information will be securable from anecdotal sources – that is, it will be unnecessary to resort to formal mechanisms for the compiling and transmission of this information. At the 'street level', each typically will be able to

> consult personal experience and informally shared information with friends and family members to determine whether or not the level of voluntary cooperation in more impersonal, 'public' settings has been great enough to warrant voluntary compliance on one's own part. (McClennen 2001, p. 203)

Critics will say otherwise, and may point to instances where individuals successfully lie or cheat their fellow men. The issue is yet to be settled, but as it stands, there is no real knockdown argument against translucency, and I thus conclude that Gauthier's argument for constrained maximization remains undefeated on the ground of the implausibility of translucency. Furthermore, there is reason to believe that the evolutionary story of the origins of human cooperation can help to bolster the view that humans have evolved as translucent creatures. If the story I have endorsed about group selection on cooperative cultural variants and individual selection of prosocial dispositions is defensible, then it would be plausible to suggest that among those prosocial dispositions would have also evolved an ability to detect the character of others and signal one's own trustworthy character to others.

It has been suggested that even if it can be established that those who develop a disposition to constrained maximization do better than those who do not, that does not entail that it is also rational to act on that disposition (Kavka 1987; Parfit 2001). In other words, the success of Gauthier's project rests not only on whether it is rational to acquire the disposition to constrained maximization, it must also be established that it is rational to act on that disposition. Critics of Gauthier might grant that it is rational to form a disposition to constrained maximization but deny that it is rational to carry through with that disposition. Consider Gauthier's example of harvesting one's crops, borrowed from Hume (Gauthier 1994, p. 692; Hume 1988, Book iii, Part ii, Section iv, pp. 520–521). Persons A and B have crops to harvest. They can do so alone or they can agree to help one another. Helping one another will involve one person helping the other at T1, and the other returning the help at T2. In order for person A to receive assistance from person B in harvesting her own crops, she will have to assure B that, after B helps her at T1, she will provide B with assistance at T2. Suppose that B helps A at T1. Is it now, at T2, rational for A to help B?

Critics of Gauthier will contend that it is not. Their argument might go like this. A's act of assuring B constitutes A forming a disposition to help B. Assuming translucency, A's forming of the disposition causes B to help A. A then gains from the disposition – a gain she would not have obtained had she not formed the disposition. It is thus rational for A to form the disposition to help B. But now comes T2. Is it now rational for A to actually do what she disposed herself to do? Helping B at T2 imposes a cost on A. At T2 A has already received B's help and so (let us suppose) has nothing to further gain from helping B at T2. Thus, A has no reason to help B at T2. How can we say that helping B is now rational?

To relate this example to constrained maximization, while it might be rational to dispose oneself to constrained maximization at T1, insofar as doing so secures a cooperative gain at T2, it is not clear that at T2 carrying through with one's part of the agreement is rational. At T2, having already secured the cooperative outcome, the individual disposed to constrained maximization has nothing further to gain from that disposition. At T2, it is utility-maximizing for an individual to act as a straightforward maximizer. Thus it would seem rational to, at T2, abandon actions recommended by constrained maximization.

Gauthier resists this conclusion and argues that the rationality of forming the disposition to constrained maximization does indeed entail the rationality of actions recommended by that disposition. According to him, there is a crucial relationship between an agent's assurance that she will carry through with her agreement, her intention to do so, and her success in securing the cooperative outcome. Specifically, Gauthier contends that in order for A to receive the cooperative benefit from B at T2, A must provide B at T1 with a sincere assurance that she will carry through with the agreement. In order to provide a sincere assurance at T1, A must have the intention that at T2 she will carry through with the agreement. In order to have the intention to cooperate at T2, A must believe that it is rational for her to do so. According to Gauthier, A cannot, without inconsistency, provide B with sincere assurance that she will cooperate at T2 if she knows that at T2 it will no longer be to her benefit (and will thus be irrational) to cooperate.

But if A evaluates the rationality of her actions at the level of action, she must concede that it is not rational to cooperate at T2, since at T2 cooperating is not the utility-maximizing option. If A evaluates her action this way, then she cannot make a sincere assurance at T1 to cooperate at T2, since she knows that at T2 cooperating is not the utility-maximizing (and thus rational) option. Without the sincere assurance that A will cooperate, B will not either. Both A and B will end up with the non-cooperative outcome, and each will harvest her crops alone.

To avoid this outcome, Gauthier recommends the adoption of a different perspective from which to evaluate the rationality of action. His concern shifts to the rationality of dispositions rather than actions. According to Gauthier, 'intentional structures create problems for the orthodox account of deliberation, which insists that rational actions are those that directly promote the agent's aim, taking as illustrative the aim that one's life go as well as possible.' He continues, 'If my aim is that my life go as well as possible, then I should not take all of my reasons for acting directly from that aim, considering only which action will have best consequences for my life. For if I always deliberate in this way, then my life will not go best for me' (Gauthier 1994, p. 692).

Gauthier's revised account of rational deliberation permits A to cooperate at T2 so long as making the assurance at T1 to cooperate at T2 yields a greater utility than doing otherwise. And, 'since the direct link between rational deliberation and particular outcomes has been severed, an action may be rational even though at the time of performance it is not, and is not believed to be, part of a life that goes best for the agent' (Gauthier 1994, p. 171). Thus, it is rational for A to cooperate at T2.

Thus, on the above account of rational deliberation, good reasons are those that lead to optimality. This reconceptualization of reason permits Gauthier to say that an action recommended by constrained maximization is rational, even if there is an alternative act that will yield a higher utility so long as the disposition to constrained maximization yields greater utility than does the disposition recommending the alternative action (in our case, the disposition to straightforward maximization).

Actions that, if taken as individual tokens are non-utility-maximizing are rational, so long as they are recommended by a disposition that it is rational to have. This account resolves the problem of compliance and ensures a superior outcome to one obtainable from a maximizing conception of rationality.

One might ask whether Gauthier is entitled to make this kind of reconceptualization of rationality. McClennen (2001, pp. 189–208) thinks so. On his view, while the cooperative outcome is a dominated and non-equilibrium point in the Prisoner's Dilemma, that fact alone should not exclude that outcome as a resolution to the problem. According to him, convergence can in fact occur at loci other than equilibrium points, and he thinks that Pareto-Optimality is one such locus.[4] He considers games of pure coordination, i.e. games where there is no conflict of interest between players and the goal is merely to coordinate their strategies. Of these games, he says:

> The appropriate concern for rational players... is to coordinate strategies so that the outcome will satisfy the Pareto condition. It is true, of course, that outcomes satisfying the Pareto condition satisfy the equilibrium condition. But from the perspective of the strategic problem that players face in a game of pure coordination, this additional property is purely accidental. That is, the equilibrium concept adds nothing that illuminates the nature of the deliberation that persons face in such game. In this context, it does no work. (McClennen 2001, p. 196)

On McClennen's view, since in coordination games it is not clear that the equilibrium concept itself plays much of a role in the determination of strategic action, and since instead much of the work is (and, according to him, ought to be from the point of view of rationality) done by Pareto-considerations, there is reason to suppose that the same might go for cooperation problems. This opens the door to an account like Gauthier's where reason is reconceptualized and actions are rational when recommended by rational dispositions, even if those actions are not individually utility-maximizing (and are, in fact, out of equilibrium in one-shot Prisoner's Dilemmas). We thus arrive at a reconciliation between cooperative dispositions, the actions they recommend, and rationality.

3.5 Connecting the Descriptive and the Normative

Thus far we have taken up two questions. The first is the descriptive question of how to explain the emergence of cooperation. The second is the normative question of why one should cooperate. These two are structurally similar. The constraints imposed by natural selection in the descriptive case and by rationality

[4] Gauthier defines this as follows: 'An (expected) outcome is optimal (or, more fully, Pareto-optimal) if and only if there is no possible outcome affording some person a greater utility and no person a lesser utility. Alternatively, an outcome is optimal if and only if any other outcome affording some person a greater utility also affords some other person a lesser utility' (Gauthier 1986, p. 76).

in the normative case make explaining or justifying cooperation difficult. How is it possible that cooperation evolved, given the workings of natural selection? How can we justify cooperation, given the self-serving conception of rationality? Regarding the former, I have argued that cultural group selection provides us with a plausible explanation of the emergence of the widespread cooperative behaviours among human beings. In the normative context, I have argued that David Gauthier's argument for the rationality of adopting the disposition of constrained maximization is a defensible route to reconciling cooperation with rationality.

In this section I will examine the relationship between the descriptive and normative projects. I will argue that the descriptive and normative projects are not only dependent on one another, but converge on the same outcome. This convergence comes from two independent lines of enquiry. Cultural group selection permits us to explain the emergence of behaviours that are genuinely fitness-decreasing at the individual level but that are beneficial to groups of individuals that display these behaviours. Explaining the emergence of cooperation requires us to shift our perspective from the individual to the level of groups. Individually, cooperation is more costly than selfishness. Collectively, cooperation pays.

A similar shift in perspective is required to justify the rationality of cooperation. Gauthier's project is to show that cooperative behaviour is rational, in spite of it being disadvantageous at particular instances. This parallels the explanation of cooperation in nature, which, as I have argued, evolves in spite of it being disadvantageous at the individual fitness level. Non-cooperation is utility-maximizing at the level of outcomes: if rationality is evaluated as a best-reply to one's partner's actions, non-cooperation will always be rational. But by shifting evaluation of rationality from outcomes to strategies, the cooperative – and superior – outcome can be achieved. This perspective permits us to rationally justify the constraints that morality requires. Thus, just as it is the case that, contrary to what we might expect, evolution supports cooperation, so too is it the case that, contrary to appearances, rationality supports cooperation. There is an advantage to cooperative behaviour in a particular context. That is, when others cooperate and when cooperation permits out-competition of other groups in the evolutionary context, and when it permits a mutually preferred outcome to universal defection in the moral context, then cooperating is better than not cooperating.

If I am right that the correct explanation of the existence of cooperation in nature appeals to group selection on cultural variants, then we also are able to arrive at an evolutionary story of the emergence of certain prosocial dispositions, viz., ones that dispose us to comply with norms and agreements. Group selection helps to shape an environment that favours prosocial psychological mechanisms. As Richerson and Boyd put it, 'if generally cooperative behavior is favored in most social environments, selection may favor genetically transmitted social instincts that predispose people to cooperate and identify within larger social groupings' (Richerson and Boyd 2005, p. 215). These emotions, in turn, help to reinforce cooperative behaviour. As Bowles and Gintis suggest:

> Some prosocial emotions, including shame, guilt, empathy, and sensitivity to social sanction, induce agents to undertake constructive social interactions; others, such as the

> desire to punish norm violators, reduce free riding when the prosocial emotions fail to induce sufficiently cooperative behavior in some fraction of members of the social group. (Bowles and Gintis 2003, p. 433)

Prosocial dispositions are ones that, if evaluated at individual instances, are not fitness maximizing. And insofar as this is true, they are also structurally similar to those actions required by constrained maximization. And if Gauthier is right that the actions recommended by constrained maximization are rationally defensible, then we will have shown that the dispositions emerging from the evolutionary story that I endorse can also be rationally defended. Cultural evolution yields prosocial dispositions, which constrain our self-interested pursuits. In a similar manner, rationality requires that we form dispositions towards constrained maximization. Thus we have a structural coincidence between what evolution produces and reason dictates.

The structural similarity in the outcomes of both projects provides a mutual support for each. The prosocial emotions that emerge from the cultural story also help to fill in the contingent facts upon which the rationality of constrained maximization depends. The cultural story provides evidence for the existence of a population structure where constrained maximization is rational. And Gauthier's analysis of the rationality of actions recommended by rational dispositions suggests the possibility of a rational justification for the dispositions that evolution produces.

3.6 Implications of the Convergence

I will end by pointing to some further implications of the convergence that I have illustrated above. The prosocial dispositions that emerge from the cultural evolutionary story suggest that mere considerations of self-interest do not exhaust the list of reasons that agents will employ. Gene-cultural co-evolutionary theory gives an account of a genetic component of cooperative tendencies, and is able to tell a story of how it might be the case that certain genetically regulated dispositions (e.g., empathy, conscientiousness, etc.) might have evolved in the first place. Put roughly, the theory states that these dispositions emerged as adaptations to an environment where cooperative behaviours were advantageous. Thus, genetic dispositions to behave in 'appropriate ways' were selected in this new, culturally created 'cooperative' environment.

The presence of these prosocial dispositions points to the possibility of a reorientation of rationality and a corresponding naturalistic normative moral theory. Developing that lies beyond the scope of my aim here. However, the promise is this: an evolutionarily-informed reconceptualization of rationality would stretch the standard conception of self-interest to include those preferences endowed to us by nature, such as sharing, group membership, fairness, and so on. On this suggestion, included among those things that we want to maximize will be fairness, sociality, etc.

Such a revision and broadening of what counts as an interest in light of evolutionary theory promises to resolve some of the tensions of a Gauthier-type analysis with respect to including the vulnerable in the sphere of moral concern. The notion of cooperation for mutual advantage that lies at the heart of contractarian moral theory entails that those who are unable to contribute anything to the cooperative outcome – such as the mentally underdeveloped, infants, animals, and so on – are thereby excluded from the sphere of moral concern. This failure to extend moral consideration to non-contracting parties is a pressing problem for the contractarian. Gauthier handles this problem by restricting the scope of morality with which he is concerned. But by introducing a wider conception of interests, which can include things such as care for others, inclusiveness, and so on, we can accommodate those to whom we think concern ought to be granted without having to either narrow the scope of morality or abandon rationality. An evolutionary perspective supports a broadening of interests and the resulting conception of rationality, I contend, provides a more suitable basis for normative moral theory.

There are also some significant practical applications of an evolutionary account of cooperation, in particular with respect to institutional design. Institutional design has typically been modelled on the *homo economicus* view of human beings, and heavily governed by rules, monitoring, incentives and disincentives. However, the cultural evolutionary analysis of human cooperation supports a human motivation structure according to which we are not only driven by self-interest, but can also be altruistic. If so, the incentive systems that go with the *homo economicus* view can be replaced or supplemented with ones that stress other values such as fairness, autonomy, achievement, and teamwork. Such systems provide a more efficient and effective means of generating and sustaining cooperation, as evidenced by their success in other fields such as the automotive industry and software development. These less authoritarian frameworks promise to be generalizable to the development of policies and institutional design on larger scales, and to produce better outcomes for all.[5]

Acknowledgments Much gratitude is owed to my dissertation committee at the University of Toronto: Paul Thompson, Wayne Sumner, and Denis Walsh. I am very thankful for their comments on the dissertation, from which this paper is derived, and on earlier versions of the paper itself.

I am also grateful to my father and fellow philosopher, Alister Browne, who has patiently read and commented on everything I have ever written, and provided much needed support along the way.

References

Benkler, Y. 2011. *The penguin and the Leviathan*. New York: Crown Business.

Bowles, S., and H. Gintis. 2003. Origins of human cooperation. In *Genetic and cultural evolution of cooperation*, ed. P. Hammerstein, 429–443. Boston: Massachusetts Institute of Technology Press.

[5] For a development of this line, see Benkler (2011).

Boyd, R., and P.J. Richerson. 2006. Culture and the evolution of the human social instincts. In *Roots of human sociality*, ed. N.J. Enfield and C. Levinson, 453–477. Oxford: Berg Publishers.

Cartwright, N. 1979. Causal laws and effective strategies. *Nous* 14(4): 419–437.

Darwin, C. 1871. *The descent of man and selection in relation to sex*. London: Murray.

Gauthier, D. 1979. Bargaining our way into morality: A do-it-yourself primer. *Philosophic Exchange* 2(5): 14–27.

Gauthier, D. 1986. *Morals by agreement*. Oxford: Oxford University Press.

Gauthier, D. 1990. Justice and natural endowment: Toward a critique of Rawls's ideological framework. In *Moral dealings: Contract, ethics, and reason*, ed. D. Gauthier, 150–170. Ithaca: Cornell University Press.

Gauthier, D. 1994. Assure and threaten. *Ethics* 104(4): 690–721.

Gildenhuys, P. 2003. The evolution of altruism: The Sober/Wilson model. *Philosophy of Science* 70: 27–48.

Hobbes, T. 1651. *Leviathan*. London: Murray.

Hume, D. 1988. *A treatise of human nature*, ed. L.A. Selby-Bigge. Oxford: Clarendon.

Kavka, G. 1987. Review: Morals by agreement. *Mind* 96(381): 117–121.

Maynard Smith, J. 1964. Group selection and kin selection. *Nature* 201: 1145–1146.

McClennen, E. 2001. The strategy of cooperation. In *Practical rationality and preference*, ed. A. Ripstein and C. Morris, 189–208. Cambridge: Cambridge University Press.

Parfit, D. 2001. Bombs and coconuts, or rational irrationality. In *Practical rationality and preference*, ed. A. Ripstein and C. Morris, 81–97. Cambridge: Cambridge University Press.

Pearl, J. 2000. *Simpson's paradox: An anatomy*. Cognitive Systems Laboratory, Computer Science Department, UCLA. Retrieved from: http://escholarship.org/uc/item/8gh5613r. Accessed 13 Oct 2011.

Richerson, P.J., and R. Boyd. 1998. The evolution of human ultra-sociality. In *Ideology, warfare, and indoctrinability*, eds. I. Eibl-Eibisfeldt and F. Salter, 71–95. New York: Berghan Books.

Richerson, P.J., and R. Boyd. 2005. *Not by genes alone*. Chicago: University of Chicago Press.

Richerson, P.J., R. Boyd, and J. Henrich. 2003. Cultural evolution of human cooperation. In *Genetic and cultural evolution of cooperation*, ed. P. Hammerstein, 357–388. Boston: Massachusetts Institute of Technology Press.

Sober, E., and D.S. Wilson. 2000. Summary of Unto others: The evolution and psychology of unselfish behavior. In *Evolutionary origins of morality*, ed. L.D. Katz, 185–206. Bowling Green: Imprint Academic.

Sober, E., and D.S. Wilson. 1998. *Unto others: The evolution and psychology of unselfish behavior*. Cambridge, MA: Harvard University Press.

Chapter 4
The Importance of Commitment for Morality: How Harry Frankfurt's Concept of Care Contributes to Rational Choice Theory

Catherine Herfeld and Katrien Schaubroeck

4.1 Introduction

Economics and morality has become an odd couple, which is surprising given the roots of economic theory in moral philosophy. Nowadays almost every economist works and thinks within the constraints of the Rational Choice Theory (hereafter RCT), but the application of RCT to moral agency has always had some uncomfortable aspect to it. It is true that since the 1990s behavioural economists have increasingly conducted experimental investigations into the existence and properties of various different motives such as altruism, social norms, concerns about justice and reciprocity that can lead to pro-social behaviour of the individual (e.g. Bicchieri 2006; Gächter and Thöni 2007). However, it seems that their conceptualization within the framework of RCT leads either to a distorted picture of moral agency (as most arguments reduce moral agency to behaviour motivated by self-interest); or to an exclusion of moral actions from the domain to which the theory can be applied, which is a deplorable result given how intertwined agency and morality are. Economist and philosopher Amartya Sen has repeatedly drawn the economists' attention to this blind spot in RCT and has introduced the idea of commitment as a necessary condition for many forms of rational behaviour, including moral agency. We take up his suggestion that our ability to commit ourselves is one crucial aspect in thinking about what makes us capable of moral behaviour, and see whether and if so, how, economists could make use of this idea to conceptualize what characterizes moral agency, or – in the words of this book – what makes us moral.

C. Herfeld (✉)
Center for the History of Political Economy, Duke University, Durham, USA
e-mail: c.herfeld@gmx.de

K. Schaubroeck
Utrecht University, Utrecht, The Netherlands
e-mail: k.schaubroeck@uu.nl

B. Musschenga and A. van Harskamp (eds.), *What Makes Us Moral? On the capacities and conditions for being moral*, Library of Ethics and Applied Philosophy 31,
DOI 10.1007/978-94-007-6343-2_4, © Springer Science+Business Media Dordrecht 2013

The paper is structured as follows: first, we give a short answer to the question of why moral behaviour might be of interest for economists. Second, we introduce RCT. We then present Sen's attempt to make room for moral behaviour in modern economics, on the one hand by outlining the concept of what he calls 'commitment', and on the other hand by introducing his idea of a 'meta-ranking'. We scrutinize the challenges allegedly presented by these two concepts for economic theory against the backdrop of what RCT is and the purpose it is supposed to serve. We agree with Sen that if RCT aims to account for moral behaviour, it needs an enriched conceptual framework of human motivation. But we reject Sen's proposal of a Kantian inspired concept of commitment because it fundamentally contradicts the conceptual framework of economics. Instead, we want to revisit and develop Sen's idea of a meta-ranking as a strategy to include moral behaviour into economics. We argue that Harry Frankfurt's concept of care could be used to bolster Sen's tentative notion of a ranking over preference rankings, thereby offering a plausible conceptualization of moral behaviour that is compatible with the fundamental principles of RCT.

4.2 The Puzzling Distance Between Morality and Economics

The first question that might arise when a philosopher and an economist approach each other with the aim of exchanging their views about morality is why an economist should seriously be interested in moral agency and in what makes people behave morally. Economics shall obviously be concerned with the behaviour of and the interaction between agents *in the marketplace*, where other-regarding concerns and moral principles play, if any, a secondary role. One might even think that moral behaviour contradicts the selfish striving for personal goals and monetary gains sought after in a competitive economy. In his profound works on the foundations of RCT and welfare economics, Amartya Sen has repeatedly addressed the tensions between economics and ethics. His work suggests an answer to the question why modern economics should be concerned with moral behaviour.

In the first passages of his book 'On ethics and economics', published in 1987, Sen makes the following observation: 'While [J.S. Mill's] view of economics is quite widely held, and not without reason, given the way modern economics has evolved, there is nevertheless something quite extraordinary in the fact that economics has in fact evolved in this way, characterizing human motivation in such spectacularly narrow terms' (Sen 1987, p. 1). Sen goes on to provide two reasons why this is so extraordinary: first, economics is supposed to be concerned with real people, and the behaviour of real people is shaped – *inter alia* – by ethical considerations of 'how one should live' (Sen 1987, p. 2). According to Sen, moral values constitute an essential aspect of what motivates people, which is why they are reflected in behaviour. However, economics maintains a conception of human motivation in which self-interest is the only motivation. It thereby fails to give morality its deserved place in the picture. Secondly, Sen is surprised by the self-evident 'non-ethical' character of modern economics, given its historical evolution as an offshoot from the Scottish Enlightenment and especially the moral philosophy and ethics of Adam Smith. Sen outlines the contrast between what he calls 'the ethics-related origin' of economics (going back to Aristotle's *Politica*) and the 'engineering-based origin' of economics (developed mostly by engineers such as Léon Walras and Vilfredo Pareto). The latter tradition of economics is the one which leads to a very narrow understanding of what has become known nowadays as 'Rational Choice Theory', and which focuses on questions of logistics and allocation and on the optimal realization of specific ends. In contrast, the ethics-related tradition was concerned more with questions related to the appropriateness of ultimate ends, the good life, and to normative questions related to politics and the social order. It allowed for addressing questions of welfare and the judgment of social achievement (Sen 1987, p. 6).

Sen concedes that one could find justifying grounds for either view. However, he maintains that the 'nature of modern economics has been substantially impoverished by the distance that has grown between economics and ethics' (1987, p. 7). He argues that economics, as it has emerged, could be made more productive by paying greater attention to the ethical considerations that shape human behaviour and judgment, which would also imply the availability of an acceptable framework that accounts for the moral dimension of human agency. Moral behaviour is not treated very extensively in the economic literature, but it is important because it might result in behavioural patterns that are different from ones resulting from individual preference satisfaction. For example, the behaviour of a morally motivated person is more stable than that of a rationally calculating agent, because it is not contingent upon circumstances or changes in tastes. The economist finds himself in an explanatory dilemma, so Sen argues, as RCT does not allow for descriptively realistic explanations of such behaviour.

The importance of Sen's observations and worries is amplified in the face of the interesting but perplexing fact that RCT has become the common framework in all (American) social sciences and provides the main template for accounting for human behaviour, from political decision-making and military strategies to marriage and legal decisions, voting behaviour, conflict analysis, altruism, norm-conformity and cooperation.[1] In light of this profusion of the theory, also labelled as 'economics imperialism' (Mäki 2009a), RCT has even been praised by eminent social scientists as a universal theory of human behaviour that captures the core characteristics of human decision-making processes. So how can it be that while being criticized for its impoverished view of human motivation and deprived from the moral dimension of agency, a theory gains an established reputation as a universal paradigm in all the social sciences?

[1] For example Gary S. Becker in economics, James S. Coleman in sociology and Jon Elster in political science.

4.3 Rational Choice Theory and Its Limitations

As its name suggests, the basic idea of RCT is that human beings choose rationally. The concept of rationality is interpreted as instrumental rationality, i.e. the agent always chooses one course of action over another when he expects it to have the best consequences for him, given his beliefs and desires. An action is motivated by a specific desire and the belief that the performance of an action is the best means to satisfy this desire, given external constraints and the situation the agent finds himself in. Thus, the ultimate end is the satisfaction of one's own desires. On the formulation of instrumental rationality as utility-maximization that is common in economics, this basic idea of rational choice has been conceptualized as a form of optimizing behaviour that results from choosing the course of action that promises the greatest benefit, so-called utility, to the agent.[2] The agent chooses from a finite set of options by calculating the costs and benefits of the outcome of each course of behaviour against another. Thereby he takes into account his desires and the external constraints he confronts. Being rational, an agent chooses the option that he expects to provide him with the maximal utility.[3] Formally, this is the option that, according to his calculation leads to the maximum subjective expected utility. Thus, when the agent is rational he maximizes his individual utility.[4]

When rational choice theorists explain an action, they focus on observable behaviour rather than on underlying rationales. This is because according to the standard economic reasoning the underlying motivation for performing a given action cannot be observed. Rational choice theorists distinguish motivated action from mere behaviour on the basis of whether a certain act is performed *for a reason*. Reasons are modelled as consisting of mental states, i.e. a desire and a belief that a certain action serves as a means to satisfy this desire (Elster 1994; Rosenberg 1995). For the social scientist, the motive for an action is thus identified with the reason,

[2] The term 'utility' is mostly taken to represent the individual's personal welfare or well-being, which amounts to understanding the striving for utility as motivated by self-interest (Sen 2002, p. 27).

[3] In this paper we do not distinguish explicitly between certain, risky and uncertain prospects.

[4] This is a delicate point that often gets confused and thereby leads to misunderstandings and false interpretations of the concept of rational choice. The agent does not aim at maximizing utility in the first place. He is motivated to act because he wants to satisfy his preferences. And *because* the agent prefers one option x_1 to another x_2, the utility of x_1 is higher than the utility of x_2, not vice versa. And this is why, in choosing x_1, the agent also gets the highest utility, namely U_1. So utility-maximization is not the reason for his action but a consequence of his individual preference ordering. As preferences cannot be observed, it is assumed that the agent takes the option he prefers most which allows to infer some information about the nature of agent's preferences from the observed choice. This idea provides the basis for the well-known view of revealed preferences (Samuelson 1938; Sen 1977).

i.e. a desire and a belief of the agent. The relation between the action and the reason is assumed to be causal as in Donald Davidson's theory of action (Davidson 1963).[5]

In the standard vocabulary of mainstream economics, the desires of the person are translated into preferences. The beliefs a person holds are taken to be the information available to the agent. Formally, preferences have to fulfil the axioms of transitivity and completeness to avoid cyclic preferences and to ensure a full ordering of all options. Beliefs are represented in terms of a probability distribution in situations of limited information about the consequences of each possible action. The agent is assumed to order all feasible options on a scale according to his preferences, whereby each option promises to endow the agent with a certain amount of utility. The ordering that reflects the agent's preferences becomes formally represented by a utility function, which is then maximized. The idea behind this formal description is that, as the agent aims to best satisfy his preferences, he chooses the option that provides him with the highest (expected) utility and acts accordingly. RCT assumes that optimal satisfaction of his individual preferences is always a person's main end. This is why RCT has been interpreted as implying that the agent chooses only what lies in his own interest. We will work with this view of RCT, but we should note that there is an abundance of literature on the topic offering a wide variety of interpretations and uses (see for instance Bunge 1996).

On the basis of this understanding of RCT, making use of RCT to explain moral behaviour would require modelling it as instrumentally rational. Even though RCT does not exclude the possibility that people consider other persons' ends as important to them, it depicts moral behaviour as ultimately driven by a cost-benefit calculation, in which the morally 'right' action is the one that maximizes utility (in a broad sense). This fundamentally contradicts what common sense (and many moral philosophers for that matter) would consider as the constituents of moral agency. An individual often considers behaviour to be morally right on the grounds of a commitment to the values that such behaviour embodies; it reflects what an agent considers to be right or wrong rather than what he prefers or what would be best for himself.

Attempts have been made to model moral behaviour within the RCT-framework, either as being the outcome of a (stable and uncompromised) preference for doing the morally right thing[6]; or as the result of moral norms that constrain the set of options available to the individual, so that the morally right thing to do becomes the option that provides the agent with the maximum utility within a very limited set of alternatives (see e.g. White 2006a, b; Kahneman et al. 1986). In both cases, the moral agent is ultimately still represented as maximizing utility and satisfying

[5] In the RCT-literature no difference is made between choosing a course of action or choosing an option and the actual performance of an action. It is assumed either that the choice of the agent based on his preferences necessarily leads to the performance of an action or that the choice is the actual action. This assumption is not obvious but cannot be discussed here.

[6] Kahneman et al. for example argue that standards of fairness do in fact influence people's behaviour and suggest the introduction of a 'preference for fairness in the objective function' (1986, p. 115).

his individual preferences.[7] Sen appears to argue that these models still fail to capture a normative dimension of agency as it is embodied in moral agency, and that therefore the models do not provide a descriptively realistic or approximately realistic representation of the phenomenon in question.[8]

Sen has been very influential in his exposal of this flaw in RCT. On his view, actions that are chosen on the basis of a moral judgment (the action or the end is right) rather than on the basis of preference-satisfaction are hard to accommodate for RCT, because moral behaviour is to be understood as 'committed behaviour', which means that it is action resulting from counter-preferential choice. In the next section we will analyse Sen's concept of commitment and argue that it is impossible for RCT to accept the Kantian interpretation of practical rationality that underlies Sen's concept of commitment, even if it would possibly enhance the descriptive realisticness of moral behaviour in terms of the structure of its actual motives. In Sect. 4.5 we pursue a more promising route in his alternative conceptualization of moral behaviour in terms of a choice that implies a meta-ranking, i.e. a ranking over preference-rankings. In Sects. 4.6 and 4.7 we argue that enriching the concept of meta-ranking with Harry Frankfurt's concept of care would enable the rational choice theorist to fit moral behaviour into the economic picture in a way that is compatible with the most fundamental concepts of economics.

4.4 Sen's Concept of Commitment and Beyond

In his article 'Rational fools: A critique of the behavioral foundations of economic theory', published in 1977, Sen argues against the reduction of different kinds of motives for action under the maximization of self-interest. Moreover, Sen introduces

[7] There is a considerable amount of literature that suggests conceptualizing moral behaviour as rule-following behaviour in economics (e.g. Vanberg 2008; see also Mantzavinos 2001, pp. 106 ff. for a review). While this approach does in fact preserve the normative character or moral behaviour, it does abandon the RCT-framework and suggests the alternative of a theory of rule-following behaviour, which however has not been taken into account by mainstream economists.

[8] How much and what kind of 'realism' can be expected from economic theories is a vexed question. For our purposes it suffices to define unrealistic behavioural assumptions as a matter of descriptive inaccuracy with regard to the psychological 'make-up' of agents. Part of the complication is due to the fact that explanations can always be given on different levels (e.g. micro-level, macro-level). Depending on the level of explanation, RCT can have different degrees of 'being realistic'; this directly relates to its purpose in the model and the level upon which the *explanandum* phenomenon occurs (Mäki 2009b). Furthermore, an explanation can have several degrees of generality or specificity (Mantzavinos 2001, p. 4). How detailed the explanation of individual agency has to be in explanations of phenomena that are of interest to the economist solely depends, again, upon the *explanandum* phenomenon. If we aim at an explanation of the phenomenon 'mutual exchange' occurring on the social rather than on the individual level, it is questionable how descriptively accurate the theory of individual choice upon which it relies has to be. In contrast, if we aim at the explanation of a singular instance of an individual's behaviour (which occurs at the micro-level), then we might want the theory of rational choice to be as descriptively accurate as possible.

commitment as the motive for a kind of behaviour that is not motivated by personal welfare-concerns. According to Sen, committed action even ‘drives a wedge between personal choice and personal welfare, and much of traditional economic theory relies on the identity of the two’ (1977, p. 329). In his ‘Goals, commitment and identity’, published in 1985, Sen goes one step further and argues that the committed person does not even choose on the basis of his own goals. To explain his idea, Sen draws on a distinction between sympathy and commitment as two possible motives for other-regarding behaviour. Although sympathy is a ‘pro-social’ motive, it does not require a departure from individual utility-maximization. In the case of an action motivated by sympathy, personal welfare is directly affected by the position of others (Sen 2005, pp. 6–7). For Sen an action motivated by sympathy is ultimately self-interested, i.e. ‘the concern for others directly affects one’s own welfare’ (Sen 1977, p. 326). Succinctly he puts it: ‘If the knowledge of torture of others makes you sick, it is a case of sympathy; if it does not make you feel personally worse off, but you think it is wrong and you are ready to do something to stop it, it is a case of commitment’ (Sen 1977, p. 326).

Formulated in these terms Sen’s distinction is reminiscent of Kant’s philanthropist example outlined in his *Groundwork*. Kant introduced the friend of mankind, who finds pleasure in spreading joy around him and who can take delight in the satisfaction of others. Kant identifies this delight as the satisfaction of an inclination; an action motivated by this inclination, even if admirable, has no moral worth (Kant [1785] 1997). Moral worth is accorded only to actions motivated by duty, that is, actions performed independently of personal desires and their effect on personal well-being. In Sen’s vocabulary, an action motivated by commitment thus requires a ‘counter-preferential choice’ (Sen 1977, p. 328). This does not mean that the agent is not allowed to have any kind of additional inclinations that make the action attractive for him. It means that under at least one counterfactual condition, the utility expected by the act chosen would be unaffected. Hence, the agent would unconditionally perform the action, even if it negatively affects him or if an alternative action would make him better off. Thus, taking commitment as a possible motive for action, the identification of personal welfare as resulting from individual choice no longer holds (Sen 1977, 1985, 2005).[9] What moves the agent to action in the case of commitment is that he considers the action as the right thing to do. The group of actions done out of commitment is wider than moral actions; for instance, a healthy lifestyle or ecological awareness can serve as examples. In these cases a person considers it right to do sports or cultivate his own vegetables because he commits himself to the non-moral values of living a healthy life or protecting the environment.

[9] This identity mainly depends on the underlying understanding of the concept of a preference or the nature of reasons. As claimed before, personal welfare and individual utility are (technically) equated in mainstream economics. This assumption provides the main basis for conclusions about what we understand as a preference and what the nature of reasons is (Sen 1977, 1985).

Although Sen's paper was written more than 30 years ago, the two philosophers Fabienne Peter and Hans Bernhard Schmid consider it to be as relevant and important now as it was then and they devoted a collection of papers to it in 2007. However, without denying the enormous influence and significance it has had, we wonder to what extent Sen's critique is still legitimate. Peter and Schmid (2007) identify three main lines of argument that constitute Sen's critique from their point of view: first, he establishes the (empirical) claim that commitments play a crucial role in human behaviour; second, he demonstrates that RCT is not able to accommodate committed action ('at least in the interpretation favored by most economists'); and third, he argues in favour of developing a theory of rationality in action that is able to accommodate commitments (Peter and Schmid 2007, p. 5). Indeed, taking Sen's critique seriously does raise doubts concerning the accuracy of RCT as a descriptive framework to account for moral behaviour in the social sciences. It reactivates the debate about the usefulness of unrealistic assumptions in models, as well as the interpretation of rationality in economics and the status and purpose of RCT on a more general level. However, how challenging Sen's critique is, crucially depends upon whether we accept the presuppositions he makes about the character and demands of moral behaviour and moral reasoning and about the purpose and potentials of economics as a social scientific enterprise.

As has been pointed out by many philosophers (Pauer-Studer 2007; Anderson 2001) and Sen himself, in characterizing moral behaviour as counter-preferential choice, Sen works from within the Kantian picture of moral agency. This starting point motivates him to describe phenomena in a way that confirms the Kantian understanding of morality, but it also makes the problems he ascribes to RCT 'home-made'. The force of his objections depends on whether one believes that moral behaviour indeed is best captured in Kantian terms. Sen does not give us many arguments for why that would be so.

The second set of assumptions underlying Sen's criticism concerns the nature of economics. Roughly put, we think that Sen presupposes a level of realistic explanations of individual behaviour that might not be necessary for the economic endeavour. When Sen wrote his paper 30 years ago, the notion of a human being as a selfish egoist was still dominant in the economics profession. Sen explicitly directed his critique against the behaviourist assumption in economics at the time according to which rational choice should be identified with self-interested choice. Furthermore, he argued against the revealed preference approach that demanded from a rational choice only the minimal requirement of internal consistency of choice while relying on the picture of the utility-maximizing individual that ultimately cares only about his welfare.[10] Sen's work has made a contribution to a body of literature in economics that provoked research that seeks to specify

[10] It becomes clear that Sen identifies egoism and internal consistency as the two main constituents of rational choice theory. As he writes: 'this approach of definitional egoism sometimes goes under the name of rational choice, and it involves nothing other than internal consistency' (Sen 1977, p. 323).

the limitations of RCT as an empirical framework. Especially in the last two decades, economists have taken the criticisms seriously and modified the concepts of rational choice theory in various ways. For instance, although they assume that the agent's ultimate motive is maximizing his utility, they opened the utility-function for material and immaterial gains. They extended their concept of preference beyond the interpretation of self-interest and thereby allowed for altruism, social concerns, emotions, social norms, and reciprocity (e.g. Bicchieri 2006).

However, despite the refinements added over the last 30 years, it remains true that all motives covered by economic models are ultimately reduced to personal preference satisfaction and conceptually presuppose maximization. And these concepts, when taken literally, contradict fundamentally our common sense understanding of the normative dimension of agency, i.e. the capacity of a person to undertake judgements about right or wrong independently from immediate desires or inclinations, to base or adjust one's actions on these judgements and thus to exercise a capacity of self-determination and self-control. In Sen's concept of commitment, these are the constituting characteristics of rationality and moral agency. However, even if these were the agreed upon constituents of moral behaviour, it is worth questioning whether economics has as its primary aim the explanation of individual behaviour, or an all-encompassing theory of the formation of rational behaviour (in all its forms). It seems to us that unlike their interpretation in philosophy, the concepts of rationality and RCT are used as *tools* in economic practice and their interpretation primarily serves this purpose. When we take economics to be concerned about aggregate behaviour with the purpose of primarily understanding market phenomena, then it comes as no surprise that, when used to generate empirically testable hypotheses about individual behaviour, RCT fails. However, this failure is not a flaw if that is neither the final purpose of economics, nor what most economic practice is about. Economics is an empirical enterprise, explaining prices and their changes, reactions of consumer and producer groups, extracting the structure of particular markets and predicting effects of policy changes. The rationality-principle is the primitive underlying this enterprise from which economic analysis departs. And RCT as the principle's formal equivalent enables to capture the most difficult phenomenon of social reality, which is human choice and the resulting behaviour, while abstracting from the complexity and the individual subtleties that make human beings human and unique. The theory of individual behaviour used in economic explanations should, we think, indeed be approximately realistic, by which we mean behaviourally realistic. However, we do not agree that they necessarily have to be psychologically realistic.[11]

Summing up, we find the force of Sen's commitment-based objection against RCT questionable for two reasons: he presupposes rather than argues for a Kantian

[11] A theory that contains an as-if clause is '*behaviorally realistic* if it allows for describing human behavior in a realistic way' and it is '*psychologically (or intentionally) realistic* if the mental processes it evokes can be truthfully attributed to the agents'(Lehtinnen and Kuorikoski 2007, p. 124).

conception of (i) morality and (ii) rationality. With regard to (i) we think it is important to note that rather than giving neutral descriptions of the data, Sen already describes moral behaviour in such a way that it is impossible to explain for a non-Kantian. With regard to (ii) we argued that Sen's objection against RCT is biased in its assumptions about the nature of economics. Therefore we do not find the commitment-based objection against RCT fundamentally threatening and useful to engage with (incorporating these ideas would result in a theory that is nothing like RCT anymore). Yet there is another interesting part in Sen's article 'Rational Fools' that we would like to draw attention to, arguing that this aspect may indeed contribute to the development of a version of RCT that is open to and capable of accounting for moral behaviour.

4.5 Sen's Concept of Meta-rankings

It is remarkable, though to our knowledge not often remarked, that Sen's attempt to alleviate the economist's trouble with explaining moral behaviour consists of two proposals that are quite different in nature. In 'Rational fools' he first builds an argument round the notion of commitment, but later moves on to an argument based on the notion of meta-ranking. We think that these two concepts pull in opposite directions and fit in different philosophical traditions. The emphasis on commitment as counter-preferential choice has a Kantian flavour, whereas the idea of a ranking of one's own preference-rankings does not exit the Humean framework. Sen believes that both concepts could help to characterize moral behaviour. As we argued in the previous section, we think the first route is inaccessible for RCT. In this section we want to examine Sen's second proposal.

The idea of meta-ranking was first introduced by Sen in a 1974 paper 'Choice, orderings and morality' and further developed in 'Rational fools'. In both these papers, Sen objects to traditional RCT insofar as it has 'too little structure' and lacks the resources to conceptualize moral reasoning. On the traditional economic theory of utility, 'a person is given *one* preference ordering, and as and when the need arises this is supposed to reflect his interests, represent his welfare, summarize his idea of what should be done, and describe his actual choices and behaviour. Can one preference ordering do all these things?' (1977, pp. 335–336). Sen concludes that it cannot. At least for capturing moral judgments a single ordering does not suffice. To capture moral reasoning, the structure needs to be elaborated, and we need to consider 'rankings of preference rankings'. A given morality, Sen argues, reveals itself when agents are asked to rank the rankings of their actions. We take his idea to be that a given morality shows itself not in the actual choices that an agent makes, but in comparing different rankings of an action-set X, whereby the actual choices make up only one possible ranking of this action-set. Another ranking could rank the actions according to personal welfare, or any other criterion. The way in which these rankings are ordered tells us much more about a given morality than could be derived from merely focusing on one particular ranking of the action-set X. Observing people's actual behaviour only reveals one ranking. The technique of meta-ranking also gives a role to introspection and communication. In the case of an individual agent, for instance (but one could apply the technique to groups, aggregates as well), we take the importance of a meta-ranking to be that the moral values are also located and revealed in the meta-ranking that comes out of this agent's introspection, not only in the actual choices that he makes.

As Sen says himself, the technique of meta-ranking can be used for various purposes. It can provide the format for expressing what preferences one would have preferred to have (for example: I wish I did not like red meat so much), or for expressing ambivalence towards one's actual choices (for example in the context of addiction: I wish I did not crave for cigarettes, but given my current desires I am better off with cigarettes). The lesson to take from Sen is that economists (or social scientists) should not identify a person with his current preferences as expressed in his actual choices, but pay greater attention to the possibility of alternative preference rankings that we can associate with a person. Thus Sen intended to create room within RCT for the idea of self-scrutiny as central to our understanding of rationality, which he relates in its turn to autonomy. Creating room for autonomy and scrutiny entails allowing that an agent has 'a voice on the status of her own preferences. It is for her to decide what importance to attach to the preference that she happens to have, rather than some other preference which she might have preferred to have' (2002, p. 620).

The concept of a meta-ranking is connected to many concepts and comes up in several contexts in Sen's work, but always seems to be the translation of one fundamental concern of his, namely that the identity of an agent should not be reduced to his actual preferences but is rather to be found in the higher-order attitudes the agent has with regard to his preferences. The fact that the technique of meta-ranking can be used for various purposes prompts the question of how it can be used in a distinctive way to conceptualize *moral* reasoning. Probably the structure of a meta-ranking should be understood in a particular way for it to reveal a given morality as opposed to, for instance, a person's opinion about the preferences he would prefer to have for non-moral reasons (as in the example of addiction). Moreover, it seems plausible on an independent basis that in order for the technique to reveal something about morality, the rationale on the basis of which the meta-ranking is made should be of a particular kind. Sen does not tell us much about the rationale for making meta-rankings. If the ranking of the action-rankings is done on the basis of just another preference, does this 'enriched' structure differ enough from the traditional single-ordering structure? All that the meta-ranking does is relegate the matter of identifying preferences to a higher-order level. Is the creation of a second order as such that helpful? Or, more importantly, can this manoeuvre do what Sen wants it to do, namely express a given morality?

The need to articulate the rationale for the meta-ranking becomes even more pressing when Sen also wants the meta-ranking to be expressive of an autonomous choice. Why would this preference over preferences reveal any more 'autonomy' than the first-order preference? They both are just preferences after all. Does the difference in level make a difference as such? In order to unravel these shortcomings in Sen as well as to find solace for them, it is fruitful to examine a view in recent analytical philosophy, reminiscent of Sen's meta-ranking idea, namely Harry Frankfurt's hierarchical view of autonomy. The similarity between Sen's concept of a meta-ranking and Frankfurt's idea of second-order desires has been noted by Sen himself as well as many critics. What has not received attention so far is the substantial addition made by the later Frankfurt to his own hierarchical view of the will out of dissatisfaction with the concept of second-order desires. It is worth looking into Frankfurt's concept of 'care' to see whether this could also help in completing Sen's idea of a meta-ranking. In doing so we reveal the potential of the concept 'care' to fulfil the role that Sen assigns to 'commitment'.

4.6 Frankfurt on Autonomy and Rationality; A Matter of Caring (Not Desiring Alone)

What makes Frankfurt such a promising philosopher for economists trying to mitigate the tensions between preferences and commitment is that his anthropology contains analogues of both notions – actually his work could be interpreted as an attempt to reconcile two motivational sources in an agent: the agent's desires and the agent's cares. However, Frankfurt's notion of 'caring', as he understands it, cannot just substitute Sen's notion of 'commitment', because they really are different concepts. But our idea is that 'caring' can do the work that Sen's notion of commitment is supposed to do without creating difficulties for RCT. From his early career Frankfurt was convinced that the concept of 'desire' was not enough to build a theory of practical rationality upon. But it took him many years to formulate the idea of 'caring about something' as what was lacking from the narrow instrumental approaches to rationality. Showing the development in Frankfurt's work is interesting, we believe, because it shows a path that RCT could follow as well.

Frankfurt's original aim was to develop a theory of autonomous or authentic agency. Later on, he widened his scope and also described his theory as an account of rational agency and reasons for action. From the start he was convinced that the key to understanding human agency was agents' capacity to reflect on what moves them. According to Frankfurt, the distinguishing feature of our species is our self-consciousness, which he specifies as our capacity of *taking ourselves seriously*: 'Taking ourselves seriously means that we are not prepared to accept ourselves just as we come. We want our thoughts, our feelings, our choices, and our behavior to make sense' (2006, p. 1). The capacity to take ourselves seriously is the capacity to divide our consciousness and objectify our mental life. What we *spontaneously* feel and desire, in other words, our *first-order* desires, is described by Frankfurt as 'psychic raw material' (2006, p. 6). Animals and small children remain passively indifferent to this material insofar as they wantonly act on their impulses. But in order to achieve the status of an autonomous agent, it is fundamental that we develop higher-order attitudes and responses to these first-order desires. That is the basic idea underlying Frankfurt's hierarchical model of autonomy and the identity of the self.

Most of the time we are identified with the content of our own mind. In Frankfurt's terminology, this means that most of the time we form second-order desires or second-order volitions with regard to our first-order desires: we want our first-order desire to become effective, that is, to motivate us. Identification with our desires usually comes natural, to the extent that we do not notice that it takes place. But sometimes we have desires that we do not identify with. Those desires are less 'our own'. Of course, they take place in us and in a gross literal sense they are 'our desires', but we dissociate ourselves from them and would rather want them to be ineffective. Instead of incorporating them, we reject and thereby externalize them.[12] Frankfurt gives the example of the addict who desperately wants to quit smoking. The addict does not want to give in to his craving for a cigarette and does not desire that his first-order desire for a cigarette becomes effective. By refusing to identify with the desire, he rejects it. Although the rejected desire may persist as an element of a person's mental history, it is no longer to be attributed to the person. It is not expressive of the agent's real self. When it moves the agent to action (as will likely happen in the case of the unwilling addict), despite the fact that he does not want it to move him to action, he acts against his will and does not act autonomously. The hierarchical approach grants a special status to second-order attitudes. They are considered to express what a person really wants, to be expressive of 'the real self'. But why are second-order desires more a person's own than first-order desires? Why are they considered to have authority? After all, second-order desires are nothing more than desires, just like first-order desires. Gary Watson (1975) makes this point when he writes that the notion of orders of desires or volitions does not do the work that Frankfurt wants it to do. It does not tell us why or how a particular want can have, among all of a person's desires, the special property of being peculiarly 'his own' (1975, p. 29). It is indeed unclear why higher-order desires would have any special relation to the real self. Considered in itself, a person's higher-order desire is just another desire. Frankfurt admits that there is a problem: 'How can the [second-order desire] claim to be constitutive of what [an agent] really wants? The mere fact that it is a second-order desire surely gives it no particular authority. And it will not help to look for a third-order desire that serves to identify the person with this second-order preference. Obviously, the same question would arise concerning the authority of that desire' (Frankfurt 1992, p. 105). Infinite regress looms, and the whole hierarchical approach appears to be doomed.

Frankfurt proposes several solutions. One of them is to provide the process of identification with a rationale that is expressive of the agent's identity, which in its turn is defined by what the agent cares about. The criterion to decide between desires is a criterion that connects to an agent's concerns or 'cares' as Frankfurt says.

[12] Frankfurt (1971) compares external desires to spasms of the body, and urges that a person is no more to be identified with everything that goes on in his mind than he is to be identified with everything that goes on in his body.

Caring about something differs significantly from desiring something, such that it does not merely amount to another higher-order desire, but rather can take up a role in evaluating desires and in according or denying desires reason-giving force. The fact that an agent acts on a desire does not suffice to make this act rational, according to Frankfurt. Desires rationalize action only when they are in line with the things that an agent cares about. Originally, the concept of caring is added to Frankfurt's picture of the autonomous agent, in order to stop the regress set-off by the question 'why does a higher-order attitude reveal my true self?' But in developing the notions of care and love, Frankfurt also takes up the project of answering the question 'why does my second-order volition provide me with a reason for action?' Autonomy and practical rationality are intricately tied up in Frankfurt's theory of agency. Especially the relationship between caring and rational action is of interest with regard to RCT.

Frankfurt does not offer a definition of caring in terms of necessary and sufficient conditions.[13] Throughout his writing he describes some of the salient features of a phenomenon he assumes we are all familiar with: we all deeply care about something, be it our job, our spouse, our family, a football team, human rights, our health, and so on. So a first thing to note is that caring can have various kinds of objects. Secondly, and very importantly, what we care about is bound up with our personal identity, and that is important to understand its authority and power over us. What we care about expresses a *personal* identification with someone or something.[14] If a person cares about something, he identifies with this object in the sense that he makes himself vulnerable to losses and susceptible to benefits depending upon whether his cared-about object is harmed or benefited. When the cared-about object is affected, he himself is affected accordingly. Or, to care about something is to be 'invested' in it with one's own personality (Frankfurt 1982). When a person cares about a football team, he is disappointed when they lose and happy when they win. His happiness depends – in part – on the good or bad fortune of the team.

[13] However, it should be noted that most of the time the notion of 'a desire' is left unanalysed as well. Philosophers and economists adopting the Humean belief–desire theory of motivation mostly work with desire and belief as folk psychological concepts and they define them simply as dispositional states (e.g. Hausman 2005). As Robert Stalnaker puts it: 'To desire that P is to be disposed to act in ways that would tend to bring it about that P in a world in which one's beliefs, whatever they are, were true.' (1984, p. 15). Neil Sinhababu (2009) defends a Humean theory of motivation on the basis of a thicker notion of desire, as a mental state that involves pleasurable thoughts, but it is unclear what the status of his theory is. It is not backed up by empirical data nor by common sense intuitions.

[14] Yet what we care about is not under our volitional control, says Frankfurt. We might try to care about something, or to stop caring about something, but whether we succeed is not under our control. This passive element in a person's constitution is typical of Frankfurt's non-voluntaristic picture of care-based identity, but will not concern us so much in this paper. Note though that this characterization of care as a state one is in regardless of one's own decisions, makes care so different from what Sen means by commitment.

For our argument it is important to understand why cares should not be conflated with desires.[15] What is the difference between caring about something and desiring it? First, caring can only exist over some extended period of time, because 'the notion ... of caring implies a certain consistency or steadiness of behavior' (1982, p. 83). Nothing in the nature of a desire, however, requires that it must endure. Second, the fact that caring about something implies an 'investment' also distinguishes it from mere wanting or desiring something. People desire things they do not really care about. From the fact that I desire some ice cream, it cannot be inferred that I care about ice cream or that ice cream is something I consider to be important to me. What caring about X has in common with desiring X is that it has motivational force. When we care about something or someone, we are moved to promote its well-being. That just is what caring means. The motivational force of caring might be due to the desires that are involved (Frankfurt does not tell us because he does not mean to offer an action theory). But at least it is clear that care involves more than having a desire. Cares are patterns of dispositions (e.g. expectations, desires, emotions) that are directed towards an ideal, person or object whose well-being or promotion one is non-instrumentally concerned about.

How can one tell whether an action was motivated by desire or care? It can be done by looking whether one of the described features obtains. The sustainability of care over desire is an important indicator. To illustrate the difference, Frankfurt gives the example of someone who has tickets for a concert and who desires to attend the concert, but who is asked by a close friend for an important favour (Frankfurt 1999, pp. 159–161). Though it will make it impossible for him to go to the concert, he gladly agrees to do the favour, because he says to himself that going to the concert is not all that important to him, sincerely believing this. But when it turns out that, unexpectedly, he suffers an uncomfortable sense of loss about missing the concert, it becomes clear that he did care about attending it. Even after deciding to help his friend, he still wants to go to the concert and missing it hurts. The desire to attend the concert persists – that is partly what it means to care about something. Frankfurt explains: 'His caring about the concert essentially consists in having and identifying with a higher-order desire ... that this first-order desire [would] not be extinguished or abandoned' (1999, p. 161). Although desiring something does not mean that one cares about it, caring about something does involve desires; more specifically, it involves that one wants certain desires to persist. This future-oriented aspect differentiates caring, as some type of second-order attitude, from the second-order

[15] Caring about something ordinarily involves feelings and beliefs that express and support a person's cares. When a person cares about a football team, he desires to watch the game, he feels happy when they are playing well and he believes that this team is the best. These desires, feelings and beliefs are only of secondary importance when characterizing the phenomenon of care, which essentially consists in volitional identification. Frankfurt writes: 'The heart of the matter, however, is neither affective nor cognitive. It is volitional. That a person cares about ... something has less to do with how things make him feel, or with his opinions about them, than with the more or less stable motivational structures that shape his preferences and that guide and limit his conduct' (1994, p. 129).

volitions, which are desires to satisfy a desire. In order to know whether someone has a second-order volition one only needs a time slice. But to find out whether someone cares about something, one has to look at his behaviour over time and search for patterns. This feature of caring explains why Frankfurt claims that caring provides an agent with a diachronic unity, or with an identity. It can do this because 'caring about something implies a *diachronic* coherence, which integrates the self across time' (2006, p. 19). Suppose we cared about nothing, Frankfurt writes, 'In that case, we would be creatures with no active interest in establishing or sustaining any thematic continuity in our volitional lives' (1999, p. 162). We would still have desires, but we would no longer be engaged in guiding the course of our desires along a continuous path. The objects that a person cares about capture his attention and direct his desires and inclination towards a stable goal.

Caring is thus a higher-order attitude of some kind. It allows an agent to take a stance towards his desires, and to evaluate them. The things that an agent cares about inform this evaluation. They provide a rationale or a criterion to decide between different desires. As a descriptive picture of motivation, this seems not too controversial. Also the connection between caring and being an autonomous agent is quite uncontroversial, provided one accepts a diachronic perspective on what being an autonomous agent means. More contested are the normative claims that Frankfurt makes about caring and practical rationality. What an agent cares about is the source of his practical reasons – such that an agent has a reason to do X only if doing X promotes something he cares about. And rightly, from a moral philosopher's point of view it seems quite appalling that an agent could only be said to have a reason to help someone, for instance, if he cares about the person or about an ideal that summons to help others in need. However, Frankfurt's theory of normative reasons does not need to concern us too much in this paper. For economists are not interested in the sources of normativity. What they need is a picture upon which acting morally makes sense, not a picture that prescribes why everyone ought to act morally. Their aim is to explain the occurrence of moral behaviour, not to explain the normative force of the moral law. In this respect, caring is a useful concept for them. It adds a resource to their theory to the effect that they can better capture the phenomenon of moral behaviour. Or so we claim.

4.7 Care and Morality: Opportunities for RCT

If commitment is the only model for moral behaviour (as Kant would say, and Sen seems to say as well: acting out of sympathy is not 'the real thing') and if commitment is defined as essentially counter-preferential, then there is no way to integrate moral behaviour in RCT. But we think that moral behaviour could be integrated in RCT using another conceptualization, one that is enriched by the concept of care. Caring about something is different from desiring it, and also from desiring to desire it. For one, our cares are bound up with our personal identity in a way that mere desires are not. That I desire an ice cream does not tell you much about who I am. That I care about a healthy lifestyle, and therefore refuse to act on desires for ice cream, does reveal something about the kind of person I am. It seems to us that a person's moral position also reveals something about the identity of this person, and is therefore more likely to be bound up with his cares rather than with his desires.

It is telling that in 'Goals, commitment, and identity', Sen (1985), too, introduces the idea of 'identity' when discussing moral behaviour. He writes about the influence of identity on our goals and choices, and how economists tend to ignore the role of identity. By 'identity' he means the same as what Korsgaard and other philosophers call 'practical identity' or 'self-conception', or in other words, the way we see ourselves. Sen argues that the way we see ourselves can give rise to 'self-imposed restrictions on the pursuit of one's own goals' (Sen 1985, p. 348). He connects this influence of identity with the notion of commitment as counter-preferential choice, but we don't think that that is necessary. With Frankfurt one might think of a person's identity in terms of what he cares about, the things that matter to him, or the things he values. These higher-order attitudes do not go against our preferences, but rather organize and order them in a self-revealing way. Because 'caring' is per definition long-term and bound up with a person's identity which rationalizes certain actions, it makes sense to place a person's motives for moral behaviour in the realm of what he cares about rather than in the realm of pure reason.

Do cares generate a behaviourally realistic picture of moral behaviour? *Prima facie* the following description is recognizable and close to many of our experiences with moral behaviour. Imagine someone who desires to pursue his own pleasure, but this desire is overcome by his caring about fairness, or the environment, or his job as a teacher, which motivates him to do as he promised, or to travel by bike instead of by car, or to prepare his classes carefully. If one only takes desires into account, moral behaviour might come as a surprise. It is true that some people may always desire to do the right thing, but many of us are not that angel-like and yet we also often succeed in doing the right thing. In order to explain the moral behaviour that many people exhibit on a regular basis, we think it is not necessary to assume the interference of pure reason like Kantians suggest. Making room for higher order reflection and choice between various desires, holding them against the light of the concerns and ideals and projects an agent cares about, might suffice. If it is sufficient, this is good news for RCT.

It is good news for RCT because it is easier for economists to incorporate the concept of care into a rational choice framework than the concept of commitment. Cares are not by definition counter-preferential; rather they are 'super-preferential'. This means that they cannot be identical to what Sen calls commitments. But it also means that it is not in principle impossible for the concepts of cares and desires to co-exist within the same framework of rationality, which is crudely speaking a Humean framework.

The advantage of Frankfurt's concept of care over Sen's concept of commitment is that Frankfurt is still working in a broadly Humean framework and furthermore that it provides a rationale different from preference-satisfaction. True, Frankfurt's idea of autonomy and rationality transcends the pure instrumental means-end reasoning and he creates conceptual space for the evaluation of ends (= to rank the rankings), but all this is still compatible with the Humean-inspired theory of rational choice. In this context it is important that we express our disagreement with a certain characterization of the difference between Kantian and Humean rationality that is assumed by many philosophers and has been spelled out by, for instance, Herlinde Pauer-Studer (2007). She holds that acknowledging standards that allow agents to critically assess ends is enough to call a conception of practical rationality 'modestly Kantian'. But we think that also on a Humean account room can be created for the assessment of ends; there is even room for desire-independent standards, as long as these standards are not created by 'pure reason' but for instance by 'what the agent cares about'. On Pauer-Studer's terminology, Frankfurt would be modestly Kantian, but that is a surprising (and we think implausible) outcome. Admittedly, Frankfurt's view of practical rationality and agency bears similarities with the views of Kantian philosophers like Christine Korsgaard, insofar as they both take practical reasons to be dependent on, or are even created by, the way people conceive of themselves. But this element of 'reason-giving practical identities' is an existentialist and rather anti-Kantian aspect in Korsgaard, as has been pointed out for instance by Thomas Nagel (1996; Korsgaard 1996). What is quintessentially Kantian about Korsgaard's view of practical rationality is the idea that the structure of reflexivity itself implies the normativity of the moral law: purely because we are reflexive beings, agents who require reasons for actions, we are subject to the moral law.[16] The internal relation between rationality and morality is the defining feature of Kantian practical reason. Not according to Pauer-Studer, but the fact that on her terms also notoriously anti-Kantian philosophers (including ones working in a broadly Aristotelian framework, or ones expanding a Humean framework such that reflection on desires is not merely an exercise in making them coherent) are Kantian, indicates that Pauer-Studer's moderate interpretation of what is a Kantian view of practical rationality is too moderate. Pauer-Studer's label is too broad, we believe, and we suggest that there might be several ways in which standards to assess ends can be conceived, only one of these conceptions being really Kantian.

The use of labels ('Kantian' or 'Humean') is not always illuminating. And we do not want to get trapped in a discussion about what is 'truly' Humean or 'truly' Kantian. All we need to claim is that Sen's idea that commitment or counter-preferential choice is required to enrich the characterization of 'rational behaviour' is incompatible with the rationality-conception of RCT. In suggesting this additional aspect of practical rationality Sen moves in the Kantian direction, allowing pure reason to create practical reasons, regardless of the agent's motivational set-up. However, RC-theorists work with a Humean conception of rationality and as such defines rational actions as dependent on a person's desires. Frankfurt enriches the concept of 'desires', makes it more complex and allows for its critical assessment

[16] Roughly the view of Kantians like Korsgaard is that from the moment that someone takes anything to be a reason (relying on his practical identities), consistency requires him to value his underlying and enabling identity as an agent or a human being, and by extension 'humanity' in general.

within an agent's personal perspective. The evaluative standards that are constituted in the agent's cares, are nothing like the Kantian evaluation of ends on the basis of their universalizability.

At the beginning of Sect. 4.6 we said that care can do the work that Sen's notion of commitment is supposed to do, but one might wonder whether this is really true. At the face of it, 'care' comes close to 'sympathy' and thus to a motive that is the exact opposite of what Sen has in mind. There are at least two things to be noted about the connection of care to sympathy. First, sympathy is, as the etymology reveals, a feeling. One sees a person suffering and suffers with him. Caring about someone else, or about an ideal, requires more than that. On the care-based account of moral motivation, what makes it rational to disapprove of torture is that torturing violates something that you stand for. Your ideals and values and cares are bound up with your identity, while a feeling of sympathy might remain superficial. Apart from noting this distinction between care and sympathy, we should, secondly, pay attention to Sen's reasons for calling an action out of sympathy not a genuine moral action: would Sen's verdict be the same for action out of care? It might be, and then the relevance of care, no matter how different it is from sympathy, would be minimal in Sen's eyes. But we want to use reflection on the motive of care as a means to lift us above the Kantian opposition between 'acting on a desire and therefore out of self-interest' and 'acting against one's desires'. Reflection on what it means to care about something shows that a person's identity is made up of more than self-interested concerns. There are plenty of resources in the person himself that motivate him to transcend his self-interest in a narrow sense. Sen (1985) acknowledges this last point in his writings about identity, where he describes the phenomenon of an agent identifying with the interests of a community and choosing his actions on the basis of what would serve the community rather than his own goals. Probably Sen thinks that this identification bypasses an agent's preferences, otherwise his characterization of commitment in terms of identification is hard to square with his earlier definition of commitment as counter-preferential choice. But could it be that this insistence is due to a failure to distinguish between a motive and a purpose of an action? Frankfurt (2002) invokes this distinction when he is confronted with the criticism that acting out of care is still self-interested, and cannot serve as a model for truly other-concerned behaviour. The objection goes that on Frankfurt's model one helps another person because it affects one's own welfare and not because one owes it to the other person (Scanlon 2002, p. 182; note the similarity with Sen's criticism of acting out of sympathy). But Frankfurt rightly points out the difference between saying (i) that certain facts, such as the fact that someone is in need, become reason-giving *in virtue of* an ideal that one cares about and (ii) that one helps someone *in order to* serve an ideal. Frankfurt does not defend (ii), nor do we when we suggest that caring could be the motive of moral behaviour.

To sum up, Frankfurt's concept of care adds something new to the vocabulary with which Sen and proponents of RCT describe the moral domain. Caring differs from having a desire, even a stable preference (because care is per definition constitutive of a person's identity), it is not sympathy (because it is not a feeling), and is not commitment in Sen's definition (because it is not counter-preferential).

Rather, care indicates a pattern or structure in an agent's motivational set-up, and gives content to the idea that even if a theoretical framework (like RCT) presupposes that motivation requires a desire and that counter-preferential choice is out of the question, it is still possible to account for how genuine moral behaviour is frequently characterized.

4.8 Concluding Remarks

In this paper we narrowed down the overall question of departure of the present book 'what makes us moral' to the question 'what makes us behave morally'. We identified one important motivational force of moral behaviour, i.e. commitment to values and the norms that embody those values or ideals. Given this starting point, our aim of this paper has been three-fold: first, we observed that RCT might at first sight seem unable to accommodate for moral behaviour. Secondly, we questioned that observation by focusing attention on the various ways in which one has conceptualized moral agency, the Kantian conceptualization being only one of them. Of course what constitutes moral agency is hotly debated among philosophers themselves. We believe that there is at least one plausible picture of moral motivation that is available to RCT but sadly underexplored by economists, which is reflected in Frankfurt's concept of caring. A lot more needs to be said about the notion of 'caring' and how it can serve as a source of moral behaviour, and how it can play a distinctive role in RCT. But to our knowledge, no one in economics or the philosophy of economics has looked into the potential significance of the work of the later Frankfurt on 'care', for RCT (as opposed to his earlier work on the hierarchical account and the concept of second-order desires, which is often referred to in the literature). Thirdly, our most bold claim is that too often philosophers who criticize RCT start from wrong premises. The aims of economists are different from the ones of philosophers. Economists mostly attempt to explain data about aggregates and create models that predict market behaviour and interaction. They are frequently not *primarily* interested in arriving at psychologically correct or phenomenologically plausible descriptions of individual behaviour – but only insofar as more nuanced, more refined psychological conceptualizations may help improve explanations and predictions of aggregate social phenomena. The objection that RCT is 'unrealistic' because it cannot adequately explain individual moral behaviour does not automatically call for its rejection. Economists use RCT in order to understand and predict behaviour and behavioural changes on the social level. And to that extent indeed they are interested in the real world, so in this regard the behavioural structure should approximate reality. But we do not think it necessary for economics to give a psychologically adequate explanation of what goes on in people's heads in order for their behaviour to have moral worth; we only need to know a principle that roughly captures the rationale that underlies people's behaviour – moral or other – in order to account for robust changes on the aggregate level.

Philosophers can help economists to refine the concepts they are already working with, but their attacks on the instrumentalist conception of rationality is too often generated by ignorance about actual economic practice or by profound disagreement about the fundamentals upon which the discipline of economics is based in order to fulfil its primary purposes, a disagreement that requires a strong argumentation. Imposing Kantian rationality on RCT *de facto* contradicts RCT. While drastic changes or paradigm shifts may be warranted from time to time, we wonder whether they are called for in the matter at hand. There seems to be room within the confines of RCT that is not yet fully explored, and that welcomes a notion of care-based moral behaviour. Enriching economic models with Frankfurt's notion of care serves as one illustration of what we consider to be a fruitful site for collaboration between economists and philosophers in the venture of capturing what makes us moral.

Acknowledgments We would like to thank Malte Doehne, the audience at the conference 'What makes us moral?' (Amsterdam, June 2011), and an anonymous referee for their helpful comments and suggestions.

References

Anderson, E. 2001. Unstrapping the straightjacket of 'preference': A comment on Amartya Sen's contributions to philosophy and economics. *Ethics and Philosophy* 17: 21–38.

Bicchieri, C. 2006. *The grammar of society*. Cambridge: Cambridge University Press.

Bunge, M. 1996. *Finding philosophy in social science*. New Haven: Yale University Press.

Davidson, D. 1963. Action, reasons and causes. *The Journal of Philosophy* 30(23): 685–700.

Elster, J. 1994. Rationality, emotions and social norms. *Synthese* 98(1): 21–49.

Frankfurt, H.G. 1971. Freedom of the will and the concept of a person. *Journal of Philosophy* 68: 5–20.

Frankfurt, H.G. 1982. The importance of what we care about. In H.G. Frankfurt (1988) *The importance of what we care about*, 80–94. Cambridge: Cambridge University Press.

Frankfurt, H.G. 1992. The faintest passion. In H.G. Frankfurt (1999) *Necessity*, *volition*, *and love*, 95–107. Cambridge: Cambridge University Press.

Frankfurt, H.G. 1994. Autonomy, necessity, and love. In H.G. Frankfurt (1999) *Necessity*, *volition*, *and love*, 129–141. Cambridge: Cambridge University Press.

Frankfurt, H.G. 1999. On caring. In H.G. Frankfurt (1999) *Necessity*, *volition*, *and love*, 155–180. Cambridge: Cambridge University Press.

Frankfurt, H.G. 2002. Reply to T.M. Scanlon. In *Contours of agency. Essays on themes from Harry Frankfurt*, ed. S. Buss and L. Overton, 184–188. Cambridge, MA: MIT Press.

Frankfurt, H.G. 2006. *Taking ourselves seriously and getting it right*. Stanford: Stanford University Press.

Gächter, S., and C. Thöni. 2007. Rationality and commitment in voluntary cooperation: Insights from experimental economics. In *Rationality and commitment*, ed. F. Peter and H.B. Schmid, 175–208. Oxford: Oxford University Press.

Hausman, D. 2005. Sympathy, commitment and preference. *Economics and Philosophy* 21(1): 33–50.

Kahneman, D., J.L. Knetsch, and R. Thaler. 1986. Fairness and the assumptions of economics. In *Rational choice: The contrast between economics and psychology*, ed. R.M. Hogarth and M.W. Reder, 101–116. Chicago/London: University of Chicago Press.

Kant, I. [1785] 1997. *Groundwork of the metaphysics of morals*. Cambridge: Cambridge University Press.

Korsgaard, C. 1996. *The sources of normativity*. Cambridge: Cambridge University Press.

Lehtinnen, A., and J. Kuorikoski. 2007. Unrealistic assumptions in rational choice theory. *Philosophy of the Social Sciences* 37(2): 115–138.

Mantzavinos, C. 2001. *Individuals, institutions, and markets*. Cambridge: Cambridge University Press.

Mäki, U. 2009a. Economics imperialism: Concepts and constraints. *Philosophy of the Social Sciences* 39(3): 351–380.

Mäki, U. 2009b. Unrealistic assumptions and unnecessary confusions: Rereading and rewriting F53 as a realist statement. In *The methodology of positive economics: Friedman's essay after half a century*, ed. U. Mäki, 90–116. Cambridge: Cambridge University Press.

Nagel, T. 1996. Universality and the reflexive self. In *The sources of normativity*, C. Korsgaard, 200–209. Cambridge: Cambridge University Press.

Pauer-Studer, H. 2007. Instrumental rationality versus practical reason: Desires, ends, and commitments. In *Rationality and commitment*, ed. F. Peter and H.B. Schmid, 73–104. Oxford: Oxford University Press.

Peter, F., and H.B. Schmid. 2007. Rational fools, rational commitments. In *Rationality and commitment*, ed. F. Peter and H.B. Schmid, 3–13. Oxford: Oxford University Press.

Rosenberg, A. 1995. *Philosophy of social science*. Boulder: Westview Press.

Samuelson, P. 1938. A note on the pure theory of consumers' behaviour. *Economica* 5: 61–71.

Scanlon, T. 2002. Reasons and passions. In *Contours of agency. Essays on themes from Harry Frankfurt*, ed. S. Buss and L. Overton, 165–183. Cambridge, MA: MIT Press.

Sen, A. 1974. Choice, orderings and morality. In *Practical reason*, ed. S. Körner, 54–67. Oxford: Blackwell.

Sen, A. 1977. Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy and Public Affairs* 6(4): 317–344.

Sen, A. 1985. Goals, commitment, and identity. *Journal of Law, Economics and Organization* 1(2): 341–355.

Sen, A. 1987. *On ethics and economics*. London: Blackwell.

Sen, A. 2002. *Rationality and freedom*. Cambridge, MA: Belknap.

Sen, A. 2005. Why exactly is commitment important for rationality? *Economics and Philosophy* 21: 5–13.

Sinhababu, N. 2009. The Humean theory of motivation reformulated and defended. *Philosophical Review* 118(4): 465–500.

Stalnaker, R. 1984. *Inquiry*. Cambridge, MA: MIT Press.

Vanberg, V. 2008. Rational choice and microeconomics: On the economics of moral preferences. *American Journal of Economics and Sociology* 67: 605–628.

Watson, G. 1975. Free agency. In G. Watson (2004) *Agency and answerability. Selected essays*, 13–32. Oxford: Clarendon Press.

White, M.D. 2006a. Multiple utilities and weakness of will: A Kantian perspective. *Review of Social Economy* 6(1): 1–20.

White, M.D. 2006b. A Kantian critique of neoclassical law and economics. *Review of Political Economy* 18(2): 235–252.

Chapter 5
Quantified Coherence of Moral Beliefs as Predictive Factor for Moral Agency

Markus Christen and Thomas Ott

5.1 Coherence – From an Intuition to a Quantified Concept

The term 'coherence' is used in several scientific disciplines. Rigid definitions adapted to specific problems are, for example, found in quantum physics (Winter and Steinberg 2008) and signal processing (White and Boashash 1990). In social sciences, psychology and philosophy the term usually describes the logical and/or semantic coherence of propositions representing attitudes, actions, beliefs, judgments and the like (Thagard and Verbeurgt 1998). In particular, the notion of coherence plays an important role in truth theories (Rescher 1973) and in validating ethical arguments based on coherent moral beliefs[1] – i.e. the term has a *normative role* such that the fulfilment of the criterion 'coherence' serves as justification that, e.g., a structured set of beliefs is a true theory, or that an argument legitimates a specific action as moral.

[1] Most contemporary philosophers characterize a belief as a "propositional attitude" (see the Stanford Encyclopedia of Philosophy, http://plato.stanford.edu/entries/belief/, accessed on June 10th 2011). Whether moral beliefs are propositions in the strict sense (i.e. are truth-apt) is disputed by non-cognitivists. However, for our analysis, this controversy is not relevant. We understand beliefs as state or habit of mind in which the belief holder (the moral agent) places trust or confidence, regardless whether this state of mind refers to empirical or normative issues. Furthermore, we assume that the moral agent is (in principle) able to communicate these mental representations that reflect those issues towards others, i.e. they are accessible to empirical research. We will use the neutral term 'moral sentences' to indicate belief representations that are related to moral issues and that are accessible to empirical research, e.g. by using surveys.

M. Christen (✉)
Institute of Biomedical Ethics, University of Zurich, Zurich, Switzerland
e-mail: markus.christen@pantaris.ch

T. Ott
Zurich University of Applied Sciences, Winterthur, Switzerland
e-mail: thomas.ott@zhaw.ch

B. Musschenga and A. van Harskamp (eds.), *What Makes Us Moral? On the capacities and conditions for being moral*, Library of Ethics and Applied Philosophy 31,
DOI 10.1007/978-94-007-6343-2_5, © Springer Science+Business Media Dordrecht 2013

Many scholars in philosophy have argued that the notion of coherence has a decisive role with respect to ethical justifications. In his essay ‘On the Nature of Moral Values’ Quine writes:

> Disagreements on moral matters can arise at home, and even within oneself. When they do, one regrets the methodological infirmity of ethics as compared with science. The empirical foothold of scientific theory is in the predicted observable event; that of a moral code is in the observable moral act. But whereas we can test a prediction against the independent course of observable nature, we can judge the morality of an act only by our moral standards themselves. Science, thanks to its links with observation, retains some title to a correspondence theory of truth; but a coherence theory is evidently the lot of ethics. (Quine 1979, pp. 477–478)

A prominent representative of the coherentist tradition in ethics, John Rawls, introduced the term when explaining ‘reflexive equilibrium’ – probably the most influential notion of coherence within ethics: ‘Its [a conception of justice’s] justification is a matter of the mutual support of many considerations, of everything fitting together into one coherent way’ (Rawls 1971, p. 21).

Although Rawls never defined ‘coherence’ in a precise way, his intuition was taken by many followers who extended his theory of justice (e.g. Daniels 1979). Within ethics, at least four theories of coherence can be distinguished that cannot be reduced to a single definition of the term (Hoffmann 2008).

However, both in truth theories and in ethics, there are well-known problems associated with the concept of coherence. A classical problem refers to the controversy between foundationalism and coherentism with respect to both justifying truth claims and finding reasons for legitimating actions as moral: A set of beliefs may be entirely coherent but still represent a wrong theory, and an argument for justifying a specific act may be coherent, although the act still is morally wrong. This critique of the normative relevance of coherence is independent of the definition of coherence (i.e. its descriptive use). It denies the normative significance of coherence even if one would have found a clear definition of the term. This point, however, is not the topic of our contribution and will not be discussed further.

We intend to overcome a second, prominent critique of coherence, referring to the vagueness of the term. This line of critique claims that the *descriptive* notion of coherence is ill-defined such that it does not make sense to attribute normative significance to coherence: Since we don’t know what coherence really means, why should the fact that a system of beliefs is coherent have normative significance? Among others, Kirkham has formulated this critique as follows:

> The term ‘coherence’ as used by coherence theories has never been very precisely defined. The most we can say by way of a general definition is that a set of two or more beliefs are said to cohere if and only if (1) each member of the set is consistent with any subset of the others and (2) each is implied (inductively if not deductively) by all of the others taken as premises or, according to some coherence theories, each is implied by each of the others individually. (Kirkham 1992, p. 104)

The point is that even this very general definition of coherence proposed here – according to Kirkham ‘the most we can say’ – has its drawbacks. First, the criterion of ‘consistency’ taken alone is too weak, as it also applies to beliefs that have nothing in common with each other. Second, the criterion of (logical) implication is hard to operationalize as soon as the set of beliefs contains more than a few propositions. Furthermore, it is doubtful whether large belief systems – and it's plausible to assume that real word agents have belief systems that include hundreds, if not thousands of beliefs – ever would fulfil the criterion of coherence with respect to inductive or deductive implication. Therefore, if the notion of coherence should remain a theoretical construct that is not only a mere intuition, but allows for fruitful applications in empirical sciences, alternative approaches are required.

Our interest lies therefore in the *practicability* of coherence as an instrument to analyse the behaviour of moral agents. For example, we are interested in whether coherence can be defined in such a way that it helps to understand irrationalities in decision making and action that conflict with rational choice theories (Hastie and Dawes 2009; Gigerenzer and Gaissmaier 2011). It is well known that the behaviour of people is often inconsistent and/or conflicts with beliefs the agents hold – and when rationality is related to the coherence of beliefs an agent has and the coherence between his or her beliefs and actions (often understood as freedom from contradictions), irrationality seems to be common in many situations. However, this interpretation depends on the notion of coherence, i.e. what it really means when stating that entities e_1 and e_2 'cohere' or a system of such entities is 'coherent'. Furthermore – as most experimental approaches to coherence irrationalities only focus on (in)coherence between only two (or very few) specified beliefs/actions in the sense of logical consistency (Jussim 2005) – it would be of interest whether a definition of coherence can be found that allows us to measure the coherence of large belief sets in a practicable way. In this way, we can analyse an important aspect of the question 'What makes us moral?' namely the role of the structure of large belief sets moral agents hold for their actual moral behaviour.

To address this issue, we organize our contribution as follows: In the next section, we briefly introduce psychological notions of coherence and we argue that no satisfying concept that goes beyond a mere intuition of coherence is used in this field. In Sect. 5.3, we present the proposal of Paul Thagard, which is often called the most sophisticated philosophical notion of coherence. Our proposal is introduced in Sect. 5.4, and in Sect. 5.5 we compare our definition with Thagard's concept of coherence. In Sect. 5.6, we outline – based on our understanding of coherence – the possible causal role of coherence for moral agency, whereas potential applications in moral research are exemplified in Sect. 5.7, using the example of Thagard. In the conclusion, we briefly discuss the importance of our descriptive proposal for the normative use of coherence.

5.2 Coherence in Psychology

The use of 'coherence' as a theoretical term in empirical sciences does not necessarily imply that it is precisely defined and quantified. This holds true also for psychology. For example, early notions of coherence emerged in the context of Gestalt psychology, denoting the 'wholeness' of specific perceptions (Silverstein and Uhlhaas 2004). This notion indeed captures an important psychological marker of the experience of coherence as something that we ultimately judge by 'a seat of the pants' feeling, as Putnam would say (Putnam 1982, p. 133) – but it lacks applicability in the sense that coherence of a specific percept can be computed. One may say that coherence is understood as being a Boolean variable, i.e. something is either coherent or not coherent. Later, quantified notions of coherence have been introduced in perception psychology, although the concept is merely used as a placeholder for denoting correlations between elements (Rodwan 1965) or predictability in temporal sequences of elements (Trumbo et al. 1968).

Since the 1950s – although the term 'coherence' was not used – being 'consistent' (with respect to beliefs, beliefs and actions, etc.) became a major topic within psychology. Here, it is not possible to review fully the conceptual and empirical contributions of the various different theories of cognitive consistency (see Abelson et al. 1968; Abelson 1983). However, both in conceptual explanations (e.g. the Balance theory of Heider 1958)[2] and in experimental tests (like the 'forbidden-toy paradigm', a classic dissonance paradigm)[3] consistency was discussed involving in most cases only few beliefs or cognitive elements. Later, in personality psychology, the term 'personality coherence' has been coined to grasp three related phenomena: The first is coherence in social behaviour and experiences. Across different circumstances, people's experiences and actions are often meaningfully interconnected. People respond consistently across some contexts and display distinctive patterns of variation across others. The second phenomenon is the organization among multiple psychological processes. Personality variables do not function as independent mechanisms but as coherent, integrated systems. The third aspect of coherence is phenomenological. People generally achieve a coherent sense of self. They have a stable sense of their attributes and develop a coherent life story (Cervone and Shoda 1999). Again, this notion of coherence captures an important intuition – but it remains unclear how these quite distinct phenomena can be merged into one defined measure of coherence. However, as soon as the set of entities that are the subject of coherence attributions was refined, more precise notions emerged. An example is the central coherence theory which defines autism as an inability to integrate sources of information to establish meaning, where words and sentences are the objects of coherence measurements (Jolliffe and Baron-Cohen 1999; see also Silverstein and Uhlhaas 2004).

This brief overview outlines the difficulties when trying to define coherence in a precise way. Conceptually, there is the problem of 'holism': Virtually all elements of

[2] The classic condition in Balance theory involves the relations among three cognitive elements constituting a triangle pattern that can be in balance or imbalance.

[3] In this paradigm, children were forbidden to play with a very attractive toy and had to resist the temptation to play with it while the experimenter was out of the room – creating a dissonance between the children knowing they very much wanted to play with this toy and knowing they were not playing with it even thought, if they did, the experimenter would be a bit (or strongly) annoyed (Aronson and Carlsmith 1963).

a system whose coherence is assessed may be connected to all others. Therefore, the problem emerges: How to define a subclass of elements whose coherence should be assessed in order to make a coherence computation feasible? This holism problem may be surmountable by heuristics that prune out links to other elements that are below a certain level of density; but such heuristics are difficult to implement in practice and are highly sensitive to context. Furthermore, if coherence should be made useful for empirical sciences (e.g. for explaining behaviour), one must analyse whether and to what extent people are actually sensible for the coherence of their beliefs. For example, in decision making coherence may be most powerful when all relevant elements are explicitly considered at the same time. However, given the limits of human working memory, the number of such elements actually considered is likely to be quite small (Keil 2006). But it might be that the interrelation between potentially accessible concepts (whose number is high) frames which elements become explicit in a decision task – i.e. an underlying 'coherence' of those concepts may be a decisive element in perception and decision making. This view is supported by research on semantic priming demonstrating that the exposure to a concept influences the response to a later presented, semantically similar concept (e.g. Friederici et al. 1999).

Therefore, quantifying coherence not only involves the task of properly defining the term but also the task of showing whether this definition is actually computable for real-world purposes and whether it can be linked to a specified problem that one wants to solve. Certainly, these tasks are interrelated: a feasible and operational definition is required in order to test the importance of coherence. And finally, demonstrating a role for coherence in real-world (moral) agency is an important aspect for discussing its normative significance. For example, the possibility that coherence critics fear, that there may be coherent belief sets that represent a wrong theory, could turn out to be practically impossible given a more precise definition of coherence.

5.3 The Suggestion of Paul Thagard

Probably the most prominent proposal to define coherence in a precise way and to operationalize it for various empirical questions was published by Paul Thagard (Thagard and Verbeurgt 1998; Thagard 2000). We share the same aim with Thagard – i.e. to turn coherence into a scientifically useful concept. Furthermore, we use (to some degree) the same conceptual framework in which coherence is quantified – i.e. a connectionism or network-based approach. Therefore, we briefly present the main points in Thagard's theory. In Sect. 5.5, we will outline the major differences between his and our proposal, which is presented in the next section.

Thagard frames the coherence problem as the maximization of satisfaction of a set of positive and negative constraints that exist between elements $\{e_1, \dots e_n\}$ of a system, whose coherence should be assessed (Thagard and Verbeurgt 1998). The elements are understood as representations (of beliefs, actions, etc.) that 'cohere' (fit together) or 'incohere' (resist fitting together). The relationships between the elements are included either as positive or negative constraints, which can be weighted to denote the strength of the constraint. For denoting the type of a constraint, Thagard made several proposals: deductive, explanatory, analogical and deliberative coherence (Thagard 1998). The coherence problem then consists in dividing the set of all elements of the system in two subsets A (for accepted) and R (for rejected) such that most of the constraints are satisfied. Satisfying a constraint means that when there is a positive constraint between e_i and e_j, they should be in the same set, whereas when there is a negative constraint between e_i and e_j, they should be in different sets. This partition of the system is associated with a number called coherence, which is the sum of the weights of the constraints which are satisfied. When this number is maximized the partition is optimal in the sense that subset A contains those elements that 'best fit together' by excluding the elements that do not fit together with the A-elements. This definition is finally related to five algorithms that could be used to calculate the coherence of a specific set of representations. Among them, Thagard favours the connectionist algorithm. This method maps the elements and their constraints onto a neural network. Each element of the set is related to a node of the network, a positive constraint between two elements results in an excitatory connection (with specified weight) and a negative one in an inhibitory link. After each node is assigned the same initial activation value, the network is updated in parallel until a stable state is reached (i.e. the activations of each node don't change any more). Nodes with activations above a certain threshold are then assigned to the set A.

Thagard exemplified his notion of coherence by a decision-making task in the murder case of Paul Bernardo, who was convicted in 1995 for the prolonged sexual torture and murder of two young women. In Canada, this case led to a discussion whether capital punishment would be appropriate for Bernardo's crimes, although the death penalty was abolished in the country. For Thagard, the case exemplifies that a person may have different, even contradicting beliefs with respect to that question. These beliefs form a constrained network that becomes the object of a coherence analysis. His example network includes different types of positive and negative constraints; he calls them deductive, explanatory, analogical and deliberative coherence. For example, the statements 'Capital punishment is sometimes justified' and 'Paul Bernardo should be executed' are (positively) connected by a deduction, the statements 'Killing a defenceless victim is wrong' and 'Capital punishment is wrong' are (positively) connected by analogy. Statements[4] with respect to the question whether capital punishment is deterrent or not are connected by explanatory relations and the statements 'Paul Bernardo should be executed' and 'Reduce prison

[4] The fact that Thagard includes the statement 'empirical evidence' in his example is somehow misleading, as the phrase actually is a placeholder for the actual evidence (i.e. either supporting 'Capital punishment is not a deterrent' or 'Capital punishment helps to prevent serious crimes'). It would make more sense to connect former sentence with 'Capital Punishment is wrong' and to understand the empirical evidence as a factor that defines the weight of either connection.

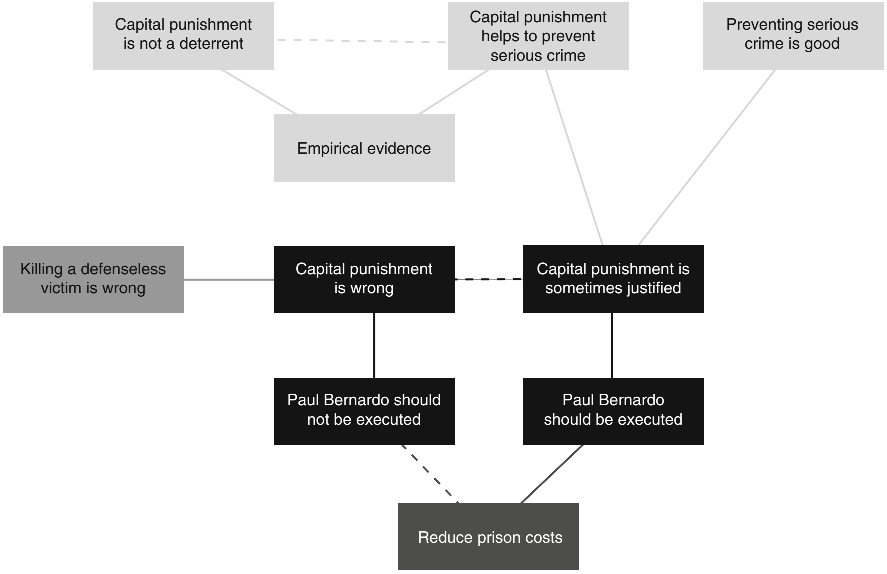
Capital punishment is not a deterrent
Capital punishment helps to prevent serious crime
Preventing serious crime is good
Empirical evidence
Killing a defenseless victim is wrong
Capital punishment is wrong
Capital punishment is sometimes justified
Paul Bernardo should not be executed
Paul Bernardo should be executed
Reduce prison costs

Fig. 5.1 The belief network with respect to the question whether Paul Bernardo should be executed or not, adapted from Thagard (2000, p. 143). Solid lines are positive constraints, dashed lines are negative constraints, and the grey scales indicate explanatory (*light grey*), deductive (*black*), analogical (*medium grey*) and deliberative (*dark grey*) coherence

expenses' are (positively) connected by deliberative coherence, referring to intrinsic goals an agent has for biological or social reasons. Figure 5.1 shows the network (adapted by us to exemplify the different types of constraints). As Thagard stated, this network only shows an extract of a (possibly) much larger network.

The example exemplifies the problem of holding many beliefs that are related in very different ways – probably a more realistic account to understand moral decision making and moral agency compared to pure deductive reasoning. However, he does not 'finish the game', i.e. the example only shows the problem, not the solution. Although he wrote that his computing algorithms were able to provide a solution (i.e. the partition of the belief set that maximizes constraint satisfaction), he did not say how the partition looks like (Thagard 2000, p. 144). This probably results from two problems. First, he needs to quantify the weights of the constraints; a point for which he does not offer a solution: 'I do not have an algorithm for establishing the weights on people's constraints' (Thagard 1998, p. 418). The reason for this may relate to the second problem: the kinds of interrelations between the beliefs he has introduced are hard to operationalize – and sometimes the classification of a specific interrelation between two beliefs as, e.g., either being 'deliberative' or 'explanatory' is ambiguous.

To sum up: Thagard's definition of coherence is precise and includes proposals for actually computing the coherence of a set of representations. He also offers a solution for the problem of circularity (Thagard and Verbeurgt 1998, Section 6) – i.e.

the notion also addresses the critique with respect to the normative significance of coherence. His proposal is therefore able to address central issues for turning coherence into a useful scientific concept.

But the definition also has some drawbacks. From a theoretical point of view, the coherence measure is (in respect of computational complexity) NP-hard, as the number of all possible partitions of a set of size n scales according to 2^n. In other words, the ‘best’ solution cannot be computed when the set contains more than a few elements. Although the connectionist algorithm is able to find a good estimate, there is no guarantee that the global maximum has been reached. Furthermore, there is no guarantee that the neural network reaches a (quasi-)stable state. Thus, these formal analyses are difficult to apply to everyday data sets. From a practical point of view, we note – as mentioned above – that the problem of finding the weights of the connections remains unsolved. In Sect. 5.5, we briefly describe other problems when comparing his definition with our proposal. The practicability of Thagard’s proposal has also been criticized by others (e.g. Keil 2006). Therefore, we will now propose an alternative definition of coherence that takes these theoretical and practical issues into account.

5.4 Our Definition of Coherence

We begin our considerations by outlining the basic idea of coherence. In its most general form, the term ‘coherent’ is a property of a set of entities that are interrelated in a specific way. Therefore, both the entities as well as the kind of interrelations have to be defined – when using the network-terminology the entities refer to the nodes and the interrelations refer to the edges of the network. The entities we are interested in are beliefs (moral sentences) that are related to the evaluation of actions and states of affairs. They form a belief system, i.e. – by following Converse’s classical definition – a ‘configuration of ideas and attitudes in which the elements are bound together by some form of constraint or functional interdependence.’ (Converse 1964, p. 207). These beliefs can take different characteristics. Daniels distinguished three types of beliefs, namely ‘moral judgments’, ‘moral principles’ and ‘relevant background theories’ (Daniels 1979, p. 258) – unquestionably very different kinds of beliefs, as it is not really clear how complex a ‘background theory’ (that probably consists of several sentences) is, compared to a mere judgment.

In the following – as we have an interest in defining coherence such that it can be used for empirical applications – we will refer to beliefs that are stated in one or only a few sentences (or even single concepts) such that they can be understood as distinct moral schemas (Jordan 2009; Lapsley and Narvaez 2004). Furthermore, we include not only purely normative beliefs (e.g.: ‘abortion is wrong’), but also beliefs on (disputed) matters of fact (e.g.: ‘abortion destroys a human being’), as long as they are normatively loaded (i.e. include ‘thick’ terms, Williams 1985) or serve as important reasons for justifying moral beliefs (i.e. belong to the framework of ‘explanatory coherence’ using Thagards’ terminology). The type of question under consideration may lead to new specifications of what counts as a single element of the system whose coherence is measured.[5]

The crucial point in defining coherence is, however, the definition of the relationship between the beliefs. Within truth theories, these relationships are defined in logical terms (implicative or deductive relationships between beliefs) and every belief that represents, for example, a wrong deduction makes the system incoherent and is, consequently, excluded from the system. In this understanding, coherence is a Boolean variable, i.e. a system is either coherent or not coherent[6] (an understanding that is probably close to the interpretation of coherence in Gestalt psychology). In formal sciences, e.g. mathematics, Boolean coherence is useful. For example, an incoherent axiom system involves contradictions that cannot be accepted.

If one allows weighting the relationships between beliefs, as Thagard has proposed, this is equivalent to defining a similarity relation (or a distance function) between them. This means that for each pair of elements e_i and e_j a number (usually between 0 and 1, if the similarity relation can be normalized) is given that reflects the similarity of these two elements.[7] Doing this for all beliefs of the system sets up a matrix that describes the pairwise similarity of all system elements. Depending on the kind of similarity relation, this matrix is symmetric or not symmetric.[8] The choice of the similarity relation depends on the question under consideration, and the values of the similarity matrix are evaluated in empirical investigations. This allows the understanding of coherence as a continuous and multidimensional variable that can be related to qualitatively distinct kinds of coherence of the system. These qualitative different system-states with respect to coherence can then be correlated to different types of behaviour. We consider this to be an adequate understanding of coherence when the concept is applied to empirical questions, as it allows for a more fruitful analysis. In moral psychology, for example, the predictive value of Boolean coherence is probably zero, as it can be assumed that no real-world moral agent has a coherent moral belief system in a strictly logical sense.

The quantitative notion of coherence we have in mind builds on the following two properties which a belief system that is understood as a network of beliefs (i.e.

[5] For example, if one would be interested to assess the coherence of beliefs a whole society holds, this may require defining the agents that hold the beliefs as single entities of the system. In that way one may assess whether there are mutually strong but diverse sub-cultures with respect to morality that may be a danger for the cohesion of a society. However, we will not elaborate on this point in this contribution.

[6] Also in this framework one may introduce a gradual measure of coherence by counting the number of beliefs that are not coherent with the system. However, this misses the point, as the function of a coherence measure in a logical setting is to identify the incoherent beliefs in order to exclude them from the system.

[7] As the similarity of two elements is equivalent to the notion of the distance between two elements, 0 usually denotes maximal similarity (= zero distance), whereas (if the distance function is normalized) 1 denotes no similarity (= maximal distance).

[8] For example, if the matrix represents belief connections as deductions, then the similarity (distance) between element e_i and e_j is 0 (when e_j is deduced from e_i), but the similarity (distance) between e_j and e_i is 1, i.e. the matrix is not symmetric.

entities/nodes with specified interrelations/edges) usually holds: First, we assume the network to be inhomogeneous. It will probably display sub-structures that can be understood as clusters of beliefs with stronger mutual relationships compared to beliefs from other clusters. This allows the introduction of some quantification of the diversity of the system. Second, these structures may display some property of stability that depends on the strength of the mutual relationships of beliefs. In that sense, the coherence of a belief system can be related to its diversity and stability.

Take the example of Thagard as a simple illustration of these two aspects. Obviously, the belief system has a sub-structure formed by at least two sets of beliefs which are either close to the decision 'Paul Bernardo should be executed' or to 'Paul Bernardo should not be executed'. Within each set, further sub-structures may be present that reflect the stability of the cluster. For example, someone may consider the analogy between 'Killing a defenceless victim is wrong' and 'Capital punishment is wrong' as rather weak, which would lower the stability of the 'not execution' cluster.

To operationalize these two dimensions 'coherence diversity' and 'coherence stability', we suggest the adaption of the concept of superparamagnetic clustering (Blatt et al. 1996; Ott et al. 2005) to define coherence. Generally, superparamagnetic clustering is a nonparametric method suitable for detecting and characterizing group structures in data without imposing a prior bias. The algorithm is inspired by a self-organization phenomenon in magnetic spin systems. In physics, this is described as follows: In an inhomogeneous spin system, clusters of correlated (synchronized) spins can emerge, corresponding to groups of spin with strong couplings. Upon an increase in temperature, i.e. an increase in stress on the system, these clusters decay into smaller units in a cascade of (pseudo-)phase transitions. Hence, the physical properties ('coherence') of the spin system are contingent on two factors: stability under stress (captured by the notion of a temperature) and diversity of the clusters. A formal introduction is provided in the Appendix.

A translation of this picture into the world of moral science yields the following correspondence: the spin system is the belief system, the single spins are the beliefs, the spin couplings reflect the similarity of the beliefs, and the clusters stand for the internal structure of the belief system and allow us to define a continuous notion of coherence along the two dimensions of diversity and stability. This allows for the identification of four ideal-typical states of a belief system (Fig. 5.1). Those can be described in terms of their role of the system for powering decisions of the moral agents as follows (see also Sect. 5.6): (1) coherence stability and coherence diversity may be low. Such a state probably does not induce a clear direction towards a decision problem for which the beliefs serve as guiding principles or motivational force. Such a structure may be typical for a decision problem in which the agent has no specific interest in and where the beliefs do not have any decisive role. One may model this type of decision problem as random decision making (with respect to moral beliefs, i.e. moral beliefs do not matter; indecisive zone). (2) If coherence diversity is high and coherence stability is low, the system exhibits a plurality of sub-groups and lacks a strong and stable core. In this situation, the belief system may support several options in the decision process, although no single option is clearly distinguished. The moral agent holds reasons for several options, although he or she does not understand the problem as conflicting (plurality zone). (3) Low coherence diversity and high coherence stability may indicate a belief system with a high degree of unity. Such a system offers a clear direction within a specific decision problem and represents the 'classic' understanding of coherence (i.e. a set of mutually supporting beliefs giving rise to clear reasons for moral action; decision zone). (4) Of particular interest is the combination of high coherence diversity and stability, as several (at least two) strong sub-groups of more or less equal size exist that are inherently stable and mutually incohesive. Such a system may be representative for a dilemmatic decision situation (dilemma zone; see Sects. 5.6 and 5.7 for further explanations).

This understanding of coherence requires two things: a distinction between single elements and a conceptualization of the interrelations between these elements. Both aspects include controversially discussed problems and may, in an empirical setting, require predefinitions. For example, one may have to define whether a specific moral sentence is understood as a single belief or as consisting of more than one belief. Also for defining the type of relationship and for quantifying it (i.e. the distance metrics) more than one option is possible. Whenever a practical problem has to be solved using this approach, one has to take these problems into account and, e.g. perform a sensitivity analysis in order to evaluate to which extent different distance metrics affect the qualitative result.

5.5 Comparison to the Proposal of Thagard

In the following, we briefly outline some distinctions between our and Thagard's definition of coherence. First, we note that Thagard uses the term 'cohere' (or 'coherent') both to describe a system property as well as the interrelation between two elements of the system. This, however, mixes two different levels of the system. We use 'coherent' solely as a descriptor of the system, whereas the pairwise interrelations of the elements of the systems are denoted by the term 'similarity'.

A second, important point refers to the issue of weighting the connections in the network. In the original proposal of Thagard and Verbeurgt (1998) the role of these weights has not been discussed further, but in other contributions (Thagard 1998, 2000), more emphasis has been put on that aspect, although he did not offer an explicit solution of how to generate these weights (see our comment in Sect. 5.3). We suspect that the underestimation of this aspect by Thagard is grounded in the fact that he started his considerations by discussing coherence relations (in our terminology: similarity) of a Boolean type, i.e. deductive or explanatory coherence (Thagard 1998), in which it is a simple yes-no-issue whether entity e_1 and entity e_2 fit together or not. By later distinguishing between types of coherence relations and allowing for weights at the same time creates some questions that are not discussed in detail by Thagard. In deductive coherence, for example, where a general principle e_1 and a particular moral judgment deduced from that principle

e_2 are linked together, the role and assessment of the weight remains unspecified. Basically, one would assume that the (normalized) weight representing the distance should be 0, as e_2 is deduced from e_1. However, Thagard writes that the constraints 'are typically soft rather than hard. A soft constraint produces a tendency to accept two positively constrained elements together, but this constraint can be overruled if overall coherence maximization suggests that one of the elements be accepted and the other rejected' (Thagard 1998, p. 408). In this description, it remains unclear how the fact that one element may be rejected and the other not (which results from the algorithmic procedure that maximizes coherence), is related to the weight of the connection between these two elements. In our proposal, the type of similarity relation defines the weights (expressed by the similarity matrix) of all edges of the network. With respect to the deduction example just mentioned, the weight could reflect the relevance of the deduction between e_1 and e_2 for the agent, i.e. refers to a psychological property that is linked to a specified measurement method. In this way we avoid the problematic entanglement of different coherence relations in the sense of Thagard by introducing a general notion of similarity that is linked to a method to assess the similarity, rather than to distinguish between different types of coherence relations.[9]

A third distinction refers to the algorithmic procedure that Thagard proposes. His goal is to subdivide the set of beliefs in two subsets – one of which would be the 'optimal set' with respect to coherence. This procedure probably reflects the foundation of Thagard's concept of coherence in truth theories – i.e. the aim is to find one single set of elements that cohere in an optimal way. However, by constructing set A consisting of elements with maximal mutual positive constraints, Thagard does not discuss the possibility that W may also consist of elements that share some degree of mutual similarity (although less than those in A) – and it remains unclear whether this fact has any explanatory role when relating the coherence of a belief set an agent holds and the behaviour of the agent. In our concept of coherence, we allow partitioning the set into various subsets (the diversity dimension) and we are able to distinguish between several qualitative states of coherence of the system (see Fig. 5.2). This has advantages with respect to relating the result of a coherence computation to agent behaviour, as we will outline in the next section.

Finally, our approach is also able to deal with the theoretical problem of Thagard's proposal (the issue of convergence of the computation, see Sect. 5.3), although both proposals are similar insofar as they are based on a network

[9] The general notion of similarity also avoids another potential source of confusion in the proposal of Thagard, the distinction between positive and negative constraints. This distinction is introduced as being of a categorical type – but it is actually blurred when allowing weights. When using the connectionist algorithm to compute coherence, negative weights are included in the network (the weight of the inhibitory links) and contribute to the activations of the neurons in the quasi-stable state – i.e. they are considered in a similar way as positive weights. In our proposal, we do not distinguish between positive and negative constraints but we reflect the degree of similarity in a distance function with endpoints 'maximal similarity' and 'no similarity'.

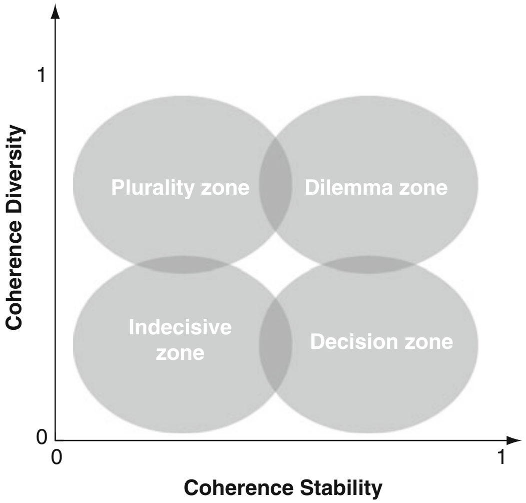

Fig. 5.2 A formal outline of four ideal-typical kinds of coherence along the dimensions 'coherence stability' and 'coherence diversity'

perspective that allows for an interpretation in terms of statistical physics (Thagard and Verbeurgt 1998, pp. 10–11; Ott et al. 2005). In both approaches, coherence is related to the network's ability to synchronize the 'coherent' elements. However, Thagard's approach defines the problem as an optimization problem for a zero-temperature spin glass. As a consequence, convergence to the global optimum, or even to a local optimum, cannot be guaranteed. In our approach we take advantage of a solution to this problem that has been presented in connection with data clustering (Blatt et al. 1996). It proposes a ferromagnetic spin system, rather than a spin glass, in order to avoid the (computationally expensive) problem of spin frustrations. Furthermore, it introduces a statistical description which is not affected by problems of convergence. Consequently, in the ferromagnetic interpretation, incoherence is understood as the inability to synchronize, whereas in the spin glass interpretation, incoherence is related to frustration.

In summary, our proposal has several advantages compared to Thagard's notion of coherence, as we avoid several open problems and questions that his proposal raised. In the following, we will now sketch a practical application of our proposal. This requires in a first step an outline why (descriptive) coherence could be a factor that may explain the behaviour of agents.

5.6 Outlining the (Possible) Causal Role of Coherence

In the following, we will focus on the use of our concept of coherence in moral psychology and moral philosophy. We outline a possible causal role of coherence along the four ideal types of coherence degrees. We repeat that our goal is to make use of this concept under consideration of the fact that real-word moral agents may hold many beliefs, from which only a subset of considerable size is related to a specific decision problem and can be accessed by the agent given limited cognitive resources (e.g. regarding memory). These beliefs can be of different kinds and are (in terms of cognitive psychology) accessible to different degrees (Higgins 1996). The challenge is, therefore, to gain an understanding of the coherence of *large* belief systems, as one can expect a connection between the similarity of beliefs and their accessibility (whether such a connection exists would be a topic of empirical research using our definition). This is why it's important to fulfil the practicability criterion – i.e. the definition of coherence should allow computing, with reasonable effort, the coherence of large systems as well as identifying the mutual similarities between the elements.

In this contribution, we will not enter into a discussion of whether beliefs (that may serve as reasons in specific decision situations) have *any* causal role in actions of moral agents (psychologists like Haidt (2001) raised doubts upon that point) or whether moral beliefs necessarily require motivational force in order to be called *moral* beliefs (the internalism-externalism debate, e.g. Brink 1997). We assume that (1) moral agents have beliefs of various types (regarding both factual and normative issues, whereas it will not be possible in all cases to draw a clear distinction between them), (2) some of these beliefs are recruited in specific decision situations, and (3) there exists at least one type of similarity between these beliefs that is relevant for the specific decision situation. We then claim that the structure of this belief-subset, in terms of coherence, is a decisive factor in understanding the actions of moral agents with respect to the specific decision problem. This claim requires (a) to find correlations between different degrees of coherence and specific behaviour patterns and (b) to show some causal relation between belief coherence and behaviour. If the claim turns out to be true, it would give coherence a predictive value for understanding moral agency. Although this describes basically an empirical project, it is based on a new understanding of the concept of coherence, which makes this claim also interesting from a philosophical perspective.

We acknowledge that our framework has to be related to a model of moral agency that includes the current knowledge of moral science (moral psychology etc.) regarding the processes moral agents use or rely upon when executing moral agency. Such a model will be necessary in order to investigate a possible causal relation between degrees of belief coherence and observed behaviour (e.g. in terms of decisions a moral agent makes). We call this model, that refers to the agent's capacity to process and manage moral problems, 'Moral Intelligence'[10] (Tanner and Christen 2013). Explaining this model in detail goes beyond the scope of this contribution. However, we briefly outline that moral intelligence both includes (psychological) capacities that are structured along a process model of

[10] As far as we know, Lennick and Kiel (2005) were the first who introduced this term in the context of business ethics.

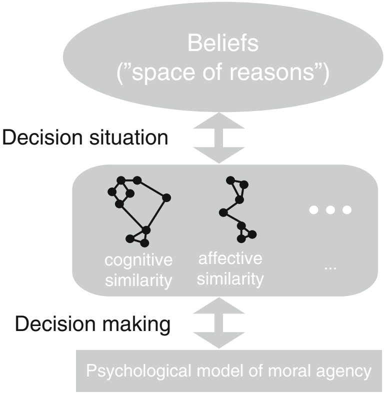

Fig. 5.3 Illustrating the framework for applying a coherence measure in moral psychology: Moral agents possess a large number of moral beliefs accessible to different degrees that are recruited during decision situations. Depending on the kind of similarity between beliefs the belief network (both the single beliefs as well as their interrelations) whose coherence is assessed will differ. Using different similarity measures assessed through empirical research, one may relate the coherence type with the actual decision performed, whereas a psychological model of moral agency serves as an instrument to analyse causalities between coherence type and behaviour

moral reasoning (moral commitment, moral sensitivity, moral problem solving, moral courage; the model can be seen as a variant of the four component model of Rest 1986) and a reference system containing one's (either existing or newly formulated) moral standards, values or convictions which provide the basis for moral evaluation and regulation (moral compass). The elements of the reference system are conceptualized as 'schemas', a standard notion to represent mental representations of beliefs an agent holds within cognitive psychology. In terms of content types, the moral compass is multifaceted. Moral values, moral convictions, ethical principles, religious beliefs, personal goals, self-related beliefs as well as behavioural scripts, etc., form such ingredients. Depending on the agent and the decision problem, it may be possible that only a single element guides decision making in a specific case. An example of this are protected values that refer to non-instrumental values involving strong moral convictions about the impermissibility of trading of specific values in exchange for other goods, in particular monetary benefits (Tanner and Medin 2004). However, in the majority of cases we expect that a set of beliefs is involved in moral decision making – and the question is whether the structure of these beliefs tells us something about how the beliefs influence decision making. To answer this question we believe that the notion of coherence could play a crucial role.

We conceptualize the framework in which the notion of coherence is applied as follows (see Fig. 5.3 for an illustration): A moral agent possesses probably many thousands of beliefs about the world and evaluations of matters of fact. These beliefs – that can be understood as forming a 'space of reasons' using Sellars's (1956) terminology – are accessible to different degrees to the agent and serve as potential reasons in a decision-making process upon moral issues. A specific decision problem recruits a subset of those beliefs that may be activated both through fast and intuitive processes and through deliberation. This subset forms the (potential) reasons the agent has in order to decide upon the various options the decision problem poses – it forms the moral compass of the agent. In real-world decision making we can expect that the number of elements within this set is not fixed and may change during the decision process (e.g. because the agent realizes that a specific problem involves additional aspects). In empirical research settings, however, one may get a better framed set of beliefs, e.g. by defining a survey that includes specific questions and thus activates in this way the beliefs the agent has regarding a specified decision problem.

After having generated a set of beliefs dedicated to a specific decision problem the question emerges as to how to model the interrelations between these beliefs. This problem both includes a qualitative (what type of interrelation?) and a quantitative (which distance metrics?) aspect. In terms of the qualitative aspect, we can assume that all beliefs recruited in the first step share some semantic similarity in respect of the decision problem for which they have been recruited for – the semantic similarity in relation to the decision problem is actually the reason why they form the belief set in question. However, the decision problem consists of (at least) two options that can be taken – and therefore, it will be possible to define a similarity metric of some beliefs in respect of these options. In empirical applications, this can be operationalized e.g. by a survey that asks whether a specific belief supports a specific option. This is one way to create a similarity measure within the belief set.[11] It maps the ‘cognitive structure’ of the belief set in question, i.e. if a specific option is able to recruit many beliefs of the set with strong internal similarity, the moral agent has many reasons for that specific option.

An alternative similarity metric relies on the motivational force of specific reasons, as one may expect that not only quantitative issues (i.e. how many reasons does the agent have to do option X?), but also qualitative issues (do the ‘important’ reasons support option X?) play a role in decision making. One way to model this could be to quantify affective responses in relation to the sympathy or aversion that pairs of specific beliefs may induce. For example, an agent may have strong emotional objections that a belief like ‘doing X is fair towards the employees of the company’ and ‘doing X maximizes profits of the company’ are in the same cluster of supportive beliefs for doing X – and a similarity measure that takes this into account could impose a substructure within a seemingly coherent set of

[11] This point involves a practical challenge as the number of possible interrelations between n elements scales by n^2. Although (depending on the problem) not all possible interrelations may have to be evaluated in empirical research, statistical reasons may require multiple tests of the same pairwise interrelation, i.e. the number of tests that has to be performed can be high. However, a general statement about this issue is not possible, as the number of tests depends on the kind of problems one wants to solve.

beliefs.[12] In empirical applications, the ‘affective distance’ between beliefs may be measured using both survey techniques as well as physiological measurements to assess unconscious responses of the agent (e.g. skin conductance).

Although we do not claim that those two similarity measures (‘cognitive’ and ‘affective’ distance) are the only ones one could use, we suggest that they are plausible candidates for a coherence analysis. Technically, both measures can be combined and weighted individually in order to assess those different aspects of similarity and their weight towards the coherence of a decision-specific belief system. Suggestion of how to operationalize these similarity measures in empirical settings will be discussed the next section.[13]

5.7 Coherence Types of Moral Belief Systems

A theory of moral agency should explain how entities act with reference to right and wrong. Such a generalized theory of moral agency involves the clarification and explanation of various aspects: Agency (e.g. individual and collective agency), the ontogenesis of moral agency (moral development), moral cultural history and the phylogenies of moral agency (the evolution of morality). As the essence of human morality is not only the ability to follow moral norms and principles, but also the ability to question an existing moral framework based on new justifications, one cannot understand moral agency without taking beliefs into account that serve as reasons in the justification process. Not only compliance to moral norms, but also the way we justify the application of specific norms – e.g. in dilemmatic situations – will play a role when moral agency is assessed. Our understanding of coherence intends to give new insights into this justification process by offering a way to empirically assess the belief systems of moral agents. In this section, we briefly sketch the application of our concept of coherence to understand specific aspects in moral agency and exemplify this by discussing the example of Thagard (the Bernardo Case).

A well-studied topic both in moral psychology and moral philosophy are dilemmas, i.e. decision situations that force individuals to make trade-offs between moral

[12] This may be the case irrespective of the cognitive similarity these two beliefs may have in a specific decision context. I.e. although it may be the case that option X is both fair to the employees and maximizes profits, the emotional aversion of seeing a similarity between these two beliefs may encode the past experience of the agent that although these two beliefs are often suggested to support the same claim they turned out to be mutually exclusive.

[13] We remind the reader that the similarity measure alone (i.e. how to quantify a single belief in respect of semantic or affective aspects) is not the only point to consider in practical applications. The distance metrics (i.e. how to quantify the distance between beliefs) may play a role, too. This goes along with the obligation to evaluate several distance metrics in practical applications in order to find a metric that is suited to the problem. Usually, standard distance metrics like the Euclidean or Manhattan distance serve the desired purpose.

values that have a similar status for the agent, or between conflicting consequences of a single value. These ‘tragic trade-offs conflicts’ that, e.g., pit two protected values against each other (such as human life versus another human life) are particularly stressful. Such situations are not only eliciting high levels of negative feelings, they are also perceived as highly difficult to solve, as the decision maker is forced to violate one of the values, if no other new solution can be found that allows upholding both values. In our framework of quantified coherence, we predict that such dilemmatic situations would be described as a combination of high-coherence diversity and stability, as several strong sub-groups exist that are inherently stable and mutually incohesive. This prediction per se is not very surprising – however, the application of different types of similarities (cognitive and affective) may allow distinguishing different types of dilemmatic settings. For example, one setting using cognitive similarity could show a paradigmatic case-4 situation (see Fig. 5.2), i.e. a situation in which two contradicting options are both supported by two belief sets of similar size and stability – whereas the application of an affective similarity measure does not provide this picture but reveals a case-3 situation. This may indicate a dilemmatic situation in terms of fulfilling a certain rationality standard (i.e. the situation is dilemmatic because two contradicting options are both supported by many good reasons) – but not in terms of emotional involvement of the agent (as the ‘important’ reasons are not in mutually exclusive belief clusters). In this way, different kinds of dilemmatic situations could be distinguished and their effect on different decision settings (e.g. personal decision making versus group decision making) could be investigated.

The Bernardo case serves as an example of such a situation, as it illustrates the dilemmatic situation that – confronted with a brutish murder case – an agent may reconsider his initial rejection of capital punishment. Just to exemplify our method, we have applied it to this (small) belief set[14] using three similarity relations: First, we translated the network of Thagard in a similarity matrix. Whenever two beliefs are positively connected, the entry in the matrix is ‘0’ (otherwise ‘1’ for indicating maximal dissimilarity or distance). Second, we performed a simple experiment to evaluate the ‘cognitive’ similarity. In this experiment, one subject weighted the probabilities for each pair of beliefs that these beliefs can be part of a common argument for or against the execution of Paul Bernardo. This probability was translated into a distance (i.e. a distance close to 0 represents a high probability that there is a common argument, whereas a number close to 1 represents low probability). Third, we repeated the experiment to evaluate the ‘affective’ similarity. In this experiment, the subject had to judge his feelings with respect to the concurrent use of two beliefs in an argumentation pro/contra capital punishment (i.e. a distance close to 0 represents a positive feeling with respect to the concurrent use, whereas a number close to 1 represents a negative feeling).

[14] The node ‘empirical evidence’ in Thagard’s initial network has been excluded, as it represents a different kind of belief; see footnote 4 for further explanations.

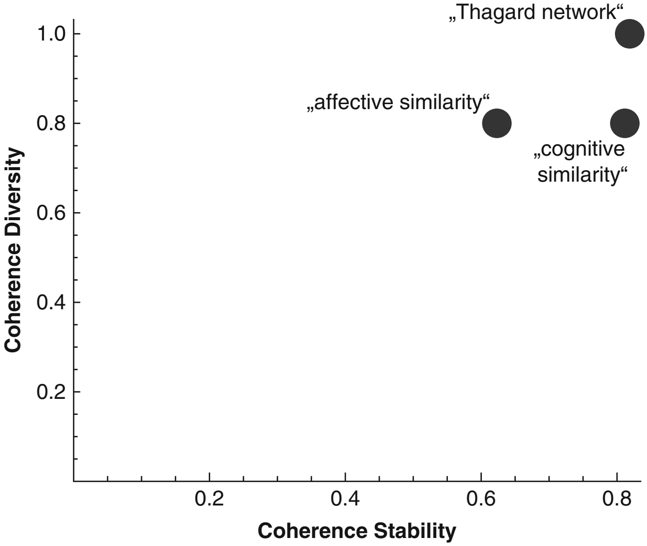

Fig. 5.4 Calculating the coherence of Thagard's belief set for the 'Bernardo murder case' example using our measure with three different similarity measures: a direct mapping of Thagard's network (see Fig. 5.1) into a distance matrix, a 'cognitive similarity' and an 'affective similarity'

The cognitive and the affective similarity of two beliefs can be different. For example, one may put a high probability on the concurrent use of the beliefs 'reduce prison costs' and 'capital punishment is justified' (low cognitive distance), whereas one may have emotional difficulties to jointly use these beliefs, because they seem to indicate that a person is executed to save money (high affective distance). After performing the coherence analysis, we find the following result (Fig. 5.4): Although all three measures put the belief set in the 'dilemma zone' (which is not surprising given the setup of the belief system), there are differences with respect to diversity and stability. In the clustering-based coherence analysis, the 'Thagard-similarity' produced three clusters resulting in a higher diversity – which is not surprising given the fact that the similarity matrix is sparse and does not take into account other potential relations between the beliefs. More interesting are the differences between the cognitive and the affective distance. The cognitive distance represented the dilemma in its classical form: the most stable cluster includes the statements 'Capital punishment helps to prevent serious crime', 'Preventing serious crime is good', 'Capital punishment is sometimes justified', 'reduce prison costs' and, finally, 'execute Paul Bernardo'. Using the affective distance, however, changed the picture. Here, the most stable cluster included the statements 'Capital punishment is not a deterrent', 'Preventing serious crime is good', 'killing a defenceless victim is wrong', 'capital punishment is wrong', and, surprisingly, 'execute Paul Bernardo'. The 'emotional cluster' is less stable than the 'cognitive cluster' – and it obviously represents the conflicting situation the Bernardo Case induced: namely that a person who is generally against capital punishment still thinks that in the case of Bernardo, the person should be executed. Thus, a decision that seems to reflect practical irrationality is reframed as resulting from a decision procedure in which affective, and not cognitive, similarities between the beliefs involved may have played the decisive role. This also has normative consequences, as it is not so clear which of the two kinds of similarities should count as ‘more rational’ given the important role of emotions in establishing what is important for humans (Rolls 2005).

We note the shortcomings of our analysis whose role is purely to exemplify our method: First, the number of the beliefs is small and does probably not include all relevant beliefs a person may have to judge this case. One would have to collect these beliefs in a first round of the experiment in order to make a more valid statement. Second, one would have to perform this experiment with various persons in order to validate the differences that emerge using the cognitive and affective similarity. However, the example shows that allowing for degrees of coherence along two dimensions gives a more detailed picture of the decision situation and allows analysing specific facets of the problem.

5.8 Conclusion

In moral philosophy, coherence is an often-mentioned, but rarely defined concept in order to state whether moral beliefs are justified guides of actions. In our contribution, we presented an alternative understanding of the coherence of belief systems that is quantified (i.e. precisely defined), that allows dealing with large belief sets and that can be adapted to investigate empirical issues about moral agency. We based our argument on the assumption that a moral agent has a large number of (moral) beliefs that are turned into reasons by the agent in specific decision situations. These beliefs are related to each other by various kinds of similarities that presupposes the distance metrics. Our concept allows quantification of the degree of coherence of this set of moral beliefs along two dimensions: the diversity of sub-groups of beliefs and the stability of a set of beliefs. In this way, four ideal types of coherence of belief-sets that are associated with moral decisions are identified, that are predicted to have different implications for the behaviour of the moral agent. We compared our proposal with the definition of Thagard and showed how we are able to overcome some shortcomings of Thagard’s definition.

The question remains, to what extent clarifications with respect to the descriptive use of coherence are of relevance for its normative importance. We believe that there are several answers to this question. The first one refers to the very basic critique that seems to be independent of any definition of coherence – namely the objection that the coherence of a set of belief representing, for example, a theory does not imply that the theory itself is true. However, although it seems to be possible to construct simple sets of coherent beliefs that are obviously false, it remains an open question whether it is really possible to create larger sets referring to real-world problems (or theories about them) that are coherent, but wrong.

A gradual understanding of coherence furthermore may ease the critique: one may indeed find systems of (lower) coherence that are wrong and are outcompeted by systems of higher coherence. The second answer refers to the (psychological well-studied, see Sect. 5.2) fact, that people seem to 'like' coherence – an important factor that motivates the normative significance of coherence – although they are sometimes (or maybe: often) unable to maintain it. Our methodology allows the assessment of this gap in more detail and could find explanations for it, which may support the motivational foundation of the normative use of coherence. Finally, making coherence an observable that can be assigned to qualitative different system states could also serve as a tool for self-understanding and thus could become an instrument that allows training the justificatory use of coherence.

Acknowledgments We thank Clare Estelle Jackson and Mark Alfano for English corrections of the original draft. This work has been supported by the Cogito Foundation (Wollerau, Switzerland, grant number R-143/08) and by the Hasler-Stiftung (Berne, Switzerland).

Appendix: Exposition of the Measure and Operationalization

Our measure of coherence is based on superparamagnetic clustering (SPC, Blatt et al. 1996) and sequential superparamagnetic clustering (SSC, Ott et al. 2005). Based on these algorithms, we have defined a measure of coherence that captures both the stability and the diversity component of coherence (Christen et al. 2009). The stability component of coherence $C_{stability}$ is calculated in the SPC framework. It is evaluated with respect to the disintegration of the largest cluster $\bar{c}$ for increasing temperature T until the system's order completely breaks apart, where T is the parameter that models the stress on the system. This involves the assumption that the largest cluster represents the 'core' of the belief system that disintegrates under stress.

Let $CS(t)$ be the size of the largest cluster for $T = t$. We assume that $CS(0) = n$, where n stands for the total number of data points; i.e., without stress, all beliefs are in the same cluster. Upon an increase in stress, $CS(t)$ decreases until $CS(t) = 1$ for some $t = T_{end}$. The average decay curve serves as a measure of coherence stability.

$$C_{stability} = \frac{1}{T_{end}} \int_{0}^{T_{end}} \frac{CS(t)}{n} dt$$

The measure is normalized to the interval [0,1]. $C_{stability}$ is close to 1 if the largest cluster remains intact for a long time and then disintegrates rapidly for high T, whereas $C_{stability}$ is close to 0 if the largest cluster disintegrates rapidly and only a small core is stable over a longer interval.

In the actual analysis, $CS(t)$ is calculated in $l + 1$ discrete steps $t = 0, \Delta T, 2\Delta T, \ldots, T_{end} = l\Delta T$. For the approximate calculation of the integral, the trapezoidal rule, known from basic calculus, is used.

$$C_{stability} = \sum_{i=0}^{l-1} \frac{CS(i\Delta T) + CS((i+1)\Delta T)}{2nl}$$

Coherence diversity $C_{diversity}$ is calculated using SSC, yielding a binary tree in which the size of each of the k sub-clusters is evaluated. Again, we consider the largest cluster $\bar{c}$ as the 'core' of the system. $C_{diversity}$ is calculated as the sum of the distance of each cluster c_i from the largest cluster in the tree diagram weighted with its size $| c_i |$. The 'tree distance' $\bar{d}_i$ is the number of bifurcation points in the tree between $\bar{c}$ and c_i. Both the maximal tree distance $\bar{d}_{\max}$ and the size of the largest cluster serve as calibration factors, leading to the definition:

$$C_{diversity} = \sum_{i=1}^{k} \frac{\bar{d}_i}{\bar{d}_{\max}} \cdot \frac{|c_i|}{|\bar{c}|}$$

$C_{diversity}$ is not normalized to 1 according to the current definition. Its value is 0 if SSC does not reveal any sub-clusters, and it is close to 0 if only small clusters emerge. However, many large clusters that have a large tree distance from the largest cluster, or fewer clusters with similar size, lead to an increase in $C_{diversity}$. Since $C_{diversity}$ is typically far below the maximally possible value, the normalization was skipped to simplify the calculation.

In this way, the measure consisting of the two components $C_{stability}$ and $C_{diversity}$ is able to capture the intuition of coherence outlined in Fig. 5.2. The concept was tested extensively and approved on the basis of toy data (Christen et al. 2009).

References

Abelson, R.P. 1983. Whatever became of consistency theory? *Personality and Social Psychology Bulletin* 9: 37–54.

Abelson, R.P., E. Aronson, W.J. McGuire, T.M. Newcomb, M. Rosenberg, and P.H. Tannenbaum (eds.). 1968. *Theories of cognitive consistency: A sourcebook*. Chicago: Rand McNally.

Aronson, E., and J.M. Carlsmith. 1963. The effect of the severity of threat on the evaluation of forbidden behavior. *Journal of Abnormal and Social Psychology* 66: 584–588.

Blatt, M., S. Wiseman, and E. Domany. 1996. Superparamagnetic clustering of data. *Physical Review Letters* 76: 3251–3254.

Brink, D. 1997. Moral motivation. *Ethics* 108: 4–32.

Cervone, D., and Y. Shoda. 1999. Beyond traits in the study of personality coherence. *Current Directions in Psychological Science* 8(1): 27–32.

Christen, M., T. Starostina, D. Schwarz, and T. Ott. 2009 A spin-based measure of the coherence of belief systems. In *Proceedings of NDES 2009*, 21–23 June 2009, Rapperswil. http://dx.doi.org/10.5167/uzh-30017.

Converse, P.E. 1964. The nature of belief systems in mass publics. In *Ideology and discontent*, ed. D.E. Apter, 206–261. New York: Free Press.

Daniels, N. 1979. Wide reflective equilibrium and theory acceptance in ethics. *Journal of Philosophy* 76(5): 256–282.

Friederici, A.D., K. Steinhauer, and S. Frisch. 1999. Lexical integration: Sequential effects of syntactic and semantic information. *Memory & Cognition* 27(3): 438–453.

Gigerenzer, G., and W. Gaissmaier. 2011. Heuristic decision making. *Annual Review of Psychology* 62: 451–482.

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

Hastie, R.K., and R.M. Dawes. 2009. *Rational choice in an uncertain world: The psychology of judgment and decision making*, 2nd ed. Thousand Oaks: Sage.

Higgins, E.T. 1996. Knowledge activation: Accessibility, applicability, and salience. In *Social psychology: Handbook of basic principles*, ed. E.T. Higgins and A.W. Kruglanski, 133–168. New York: Guilford Press.

Hoffmann, M. 2008. *Kohärenzbegriffe in der Ethik*. Berlin: Walter de Gruyter.

Jolliffe, T., and S. Baron-Cohen. 1999. A test of central coherence theory: Linguistic processing in high-functioning adults with autism or Asperger syndrome: Is local coherence impaired? *Cognition* 71: 149–185.

Jordan, J. 2009. A social cognition framework for examining moral awareness in managers and academics. *Journal of Business Ethics* 84: 237–258.

Jussim, L. 2005. Accuracy in social perception: Criticisms, controversies, criteria, components, and cognitive processes. *Advances in Experimental Social Psychology* 37: 1–93.

Keil, F.C. 2006. Explanation and understanding. *Annual Review of Psychology* 57: 227–254.

Kirkham, R.L. 1992. *Theories of truth: A critical introduction*. Cambridge, MA: MIT Press.

Lapsley, D.K., and D. Narvaez. 2004. A social-cognitive approach to the moral personality. In *Moral development, self and identity*, ed. D.K. Lapsley and D. Narvaez, 189–212. Mahwah: Erlbaum.

Lennick, D., and F. Kiel. 2005. *Moral intelligence: Enhancing business performance and leadership success*. Upper Saddle River: Wharton School Publishing.

Ott, T., A. Kern, W.-H. Steeb, and R. Stoop. 2005. Sequential clustering: Tracking down the most natural clusters. *Journal of Statistical Mechanics: Theory and Experiment* 2005: P11014.

Putnam, H. 1982. *Reason, truth and history*. Cambridge, MA: Cambridge University Press.

Quine, W.V. 1979. On the nature of moral values. *Critical Inquiry* 5(3): 471–480.

Rawls, J. 1971. *A theory of justice*. Cambridge, MA: Harvard University Press.

Rescher, N. 1973. *The coherence theory of truth*. Oxford: Oxford University Press.

Rest, J.R. 1986. *Moral development: Advances in research and theory*. New York: Praeger.

Rodwan, A.S. 1965. A coherence-criterion in perception. *The American Journal of Psychology* 78(4): 529–544.

Rolls, E.T. 2005. *Emotions explained*. Oxford: Oxford University Press.

Sellars, W. 1956. *Empiricism and the philosophy of mind*. Cambridge, MA: Harvard University Press.

Silverstein, S.M., and P.J. Uhlhaas. 2004. Gestalt psychology: The forgotten paradigm in abnormal psychology. *The American Journal of Psychology* 117(2): 259–277.

Tanner, C., and M. Christen. 2013. Moral intelligence – A framework for understanding moral competences. In *Empirically informed ethics. Morality between facts and norms*, ed. M. Christen, J. Fischer, M. Huppenbauer, C. Tanner, and C. Van Schaik. New York: Library of Ethics and Applied Philosophy/Springer.

Tanner, C., and D.L. Medin. 2004. Protected values: No omission bias and no framing effects. *Psychonomic Bulletin & Review* 11: 185–191.

Thagard, P. 1998. Ethical coherence. *Philosophical Psychology* 11: 405–422.

Thagard, P. 2000. *Coherence in thought and action*. Cambridge, MA: MIT Press.

Thagard, P., and K. Verbeurgt. 1998. Coherence as constraint satisfaction. *Cognitive Science* 22: 1–24.

Trumbo, D., M. Noble, F. Fowler, and J. Porterfield. 1968. Motor performance on temporal tasks as a function of sequence length and coherence. *Journal of Experimental Psychology* 77(3): 397–406.

White, L.B., and B. Boashash. 1990. Cross spectral analysis of non-stationary processes. *IEEE Transactions on Information Theory* 36(4): 830–835.

Williams, B. 1985. *Ethics and the limits of philosophy*. London: Fontana.

Winter, R.G., and A.M Steinberg. 2008. Coherence. AccessScience, McGraw-Hill Companies. Available through: http://www.accessscience.com. Accessed on 28 Sept 2011.

Part II
Morality and the Continuity Between Human and Nonhuman Primates

Chapter 6
Animal Morality and Human Morality

Bert Musschenga

6.1 Introduction

During the Middle Ages and subsequent centuries it was not uncommon in countries in Europe that animals which caused injury or death to humans, or even just a form of public nuisance, were prosecuted, convicted, and punished. These practices continued until the previous century (Girgen 2003).[1] Such behaviour was seen as an infringement of the hierarchical order of nature in which humans were the crown of creation and animals were subordinate to them. While it was natural for predatory animals to kill their prey or even their conspecifics, it was unnatural for them, and therefore morally wrong, to maim or kill humans. Apparently, animals were assumed to know and to comply with the prescriptions and proscriptions of human morality. Animals were seen as moral beings with diminished moral capacity, comparable to mentally handicapped humans, who nonetheless could be held responsible for their behaviour concerning humans. Human morality was the obvious framework for judging animal behaviour.

Nowadays *homo sapiens* is no longer seen as the crown of creation, an absolute sovereign who can dispose of other creatures as it pleases him. It is widely recognized that humans have moral obligations to (other) animals, and not the other way around. Animals are now admitted to the human moral community, not as full and equal members but as patients whose interests should be taken into account by

[1] In *The criminal prosecution and capital punishment of animals* ([1906] 1987) Edward Evans documented more than 191 prosecutions and excommunications of animals between the ninth and twentieth centuries. Research shows that the majority of secular prosecutions were concentrated in the southern and eastern parts of France and in adjacent parts of Germany, Italy, and Switzerland.

B. Musschenga (✉)
Faculty of Philosophy, VU University, De Boelelaan 1105, 1081 HV Amsterdam,
The Netherlands
e-mail: a.w.musschenga@vu.nl

B. Musschenga and A. van Harskamp (eds.), *What Makes Us Moral? On the capacities and conditions for being moral*, Library of Ethics and Applied Philosophy 31,
DOI 10.1007/978-94-007-6343-2_6, © Springer Science+Business Media Dordrecht 2013

humans as agents. Some authors, however, find that animals have a higher moral status than that of moral patients. They have moral rights. The debate on the moral status of animals is an issue within animal ethics which deals with the nature and content of moral obligations of humans towards animals.

It is widely agreed that many social animals have a system for the regulation of their behaviour towards members of their group, something that is functionally equivalent to the role morality has within the human community. I will call that system a 'morality'. Many students of animal behaviour are convinced that (at least some species of) nonhuman animals have a morality, and not just a 'morality'. To corroborate their view, they must be able to point out in what respects a 'morality' differs from a morality, and/or what distinguishes a 'moral' animal from a moral animal. Assuming that it can be proven that at least some animal species, e.g., the bonobos, have a morality, how does the bonobo morality relate to the moralities that are found within human societies? Do human moralities show distinctive characteristics or are all moralities – animal and human – tokens of the same type? These are the questions that I discuss in this paper. At the end of the paper I come back to animal ethics. If some animals are, just as humans, moral animals, what does that imply for their moral status? Can we still hold that all animals should be treated as moral patients? Can we treat moral animals in the same way as 'moral' animals?

Nowadays many people believe that a considerable number of social animal species have a 'morality' or perhaps even a morality. They were convinced by the huge amount of observations and stories collected by students of animal behaviour, and presented to them in popularizing books by authors such as Marc Bekoff, Marc Hauser, and Frans de Waal. I discuss the definition of morality from which criteria for classifying a system for the regulation of social behaviour as a morality must be derived in Sect. 6.2. Section 6.3 categorizes the moral behaviour patterns of animals identified by animal behavioural scientists in four clusters. In Sect. 6.4, I discuss what capacities are needed for moral behaviour. Sections 6.5 and 6.6 consider when behaviour can be said to be rule governed. In Sect. 6.7, I examine whether animals can have moral motives. Section 6.8 goes into the occurrence of social disapproval as a criterion for norm violation. In Sect. 6.9, I analyse the relation between animal morality and human morality, and argue that animal morality regulates behaviour automatically and unconsciously. However, a large part also of human morality is non-reflective and functions in the same manner as animal morality. In contradistinction to animal morality, human morality makes use of both what psychologists call System I processes and System II processes and can be both non-reflective and reflective.[2] Section 6.10 contains some reflections on the moral status of animals belonging to a species that has a morality. Section 6.11 offers some concluding remarks.

[2] There is now considerable agreement on the characteristics that distinguish the two systems. The operations of System I are fast, automatic, effortless, associative, and difficult to control or to modify. The operations of System II are slower, serial, effortful, and deliberately controlled; they are also relatively flexible and potentially rule-governed.

6.2 Definition of Morality

I argued that many social animals have a system for the regulation of social behaviour that is functionally equivalent to what morality does in human society. I didn't provide a description of what the function of morality is. In my view, the best functional definition of *human* morality was given by G.J. Warnock: '... the "general object" of morality, appreciation of which may enable us to *understand* the basis of moral evaluation, is to contribute to betterment – or non-deterioration – of the human predicament, primarily and essentially by seeking to countervail "limited sympathies" and their potentially most damaging effects' (Warnock 1971, p. 26). Becoming a moral person implies according to Warnock: learning to resist and control one's always present self-regarding tendencies. Morality's biggest enemy may be the pure egoist. But pure egoism is as rare as pure altruism. The average person has sympathy and concern, but only for a limited number of people – usually his family and friends. Therefore the proper business of morality is, in Warnock's view, 'to expand our sympathies, or, better, to reduce the liability inherent in their natural tendency to be narrowly restricted' (Warnock 1971, p. 26). Next to self-interest then, favouritism and partiality are in this view the most widespread moral problems.

Warnock speaks of expanding our sympathies. In his definition morality seems to have a universal intent. If universal intent is indeed a characteristic of morality, no system for the regulation of the social behaviour of a non-human species can qualify as a morality. Warnock's definition also doesn't cover the moralities of human societies in which the moral community coincides with the own social group. A similar, but less restrictive functional definition is found in Jessica Flack and Frans de Waal (2000). In their view human morality needs to take human nature into account by either fortifying certain natural tendencies – such as sympathy, reciprocity, loyalty to the group and family, and so on – or by countering other tendencies – such as within-group violence and cheating (p. 23). Flack and De Waal's definition can be broadened to cover animal morality, simply by skipping the adjective 'human' in 'human morality' and by substituting 'human nature' for 'animal nature'. They themselves avoid speaking of animal morality. In their view, non-human primates have a proto-morality – still a 'morality'. Human moral systems, they say, rely on basic mental capacities and social tendencies humans share with other cooperative primates, such as chimpanzees. That is why they regard it as justified to conclude that these other primates have a proto-morality. Humans, however, display unique features such as a greater degree of rule internalization, a greater capacity to adopt the perspective of others, and the unique capacity to debate issues among themselves and transmit them verbally (Flack and De Waal 2000, p. 23).[3]

[3] In *Primates and philosophers* De Waal states: 'The same process [i.e. evolution, BM] may not have specified our moral rules and values, but it has provided us with the psychological makeup, tendencies, and abilities to develop a compass for life's choices that takes the interests of the entire community into account, which is the essence of human morality' (De Waal 2009, p. 58).

The broadest definition of the function of morality is given by Dale Peterson (2011, pp. 51ff.): ‘The function of morality, or the moral organ, is to negotiate the inherent conflict between self and others.’ This definition, he says, includes the possibility that at least mammals have moral systems homologous to ours (p. 58). Marc Bekoff and Jessica Pierce define morality as ‘a suite of interrelated other-regarding behaviours that cultivate and regulate complex interactions within social groups. These behaviours relate to well-being and harm, and norms of right and wrong attach to many of them’ (2009, p. 7). They rightly distinguish between pro-sociality and altruism on the one hand, and morality on the other. To have a morality, they say, a given species must meet certain threshold requirements. These thresholds are: a level of complexity in social organization, including established norms of behaviour to which attach strong emotional and cognitive cues about right and wrong; a certain neural complexity that serves as a foundation for moral emotions and for decision making based on perceptions about the past and the future; relatively advanced cognitive capacities (such as a good memory); and a high level of behavioural flexibility (2009, p. 13). All moralities consist of well-developed systems of other-regarding prohibitions and proscriptions (2009, p. 13). The set of actions that constitute moral behaviours vary among species. So does the degree of moral complexity. Morality can be thought of as nested levels of increasing complexity and specificity. Bekoff and Pierce (2009) do not enumerate the animal species that meet the threshold requirements. What they do say is that animals with a highly developed moral capacity may include chimpanzees, wolves, elephants, and humans (p. 20). This is not an exhaustive list. The distinction between human morality and animal morality is for them quantitative rather than qualitative. Humans appear to have evolved an unusual high level of moral complexity (p. 139).

The definition I suggest combines elements of the definitions by Warnock; Flack and De Waal; Petersen; Bekoff and Pierce:

> Morality cultivates and regulates social life within a group or community by providing rules (norms) which fortify natural tendencies that bind the members together – such as sympathy, (indirect) reciprocity, loyalty to the group and family, and so on – and counter natural tendencies that frustrate and undermine cooperation – such as selfishness, within-group violence and cheating.

This definition leaves the question whether animals can have a morality open. What it does say is that rules constitute the mechanism for the regulation of social behaviour. A species can only be said to have a morality if their supposedly moral behaviour is rule governed. With humans not all rules are moral rules. If we want to find out whether a rule people follow is a moral rule we ask them, for instance, how they justify it and what their motives are for following the rule, how other people react when they violate the rule, and what kind of feelings they themselves have on such occasions. Students of animal behaviour must start with observing the behaviour, body language, and facial expressions of animals and the sounds they produce. If they observe certain regularities in their social behaviour, the next thing to do is to examine whether the regularity is caused by following a rule. Even if they can prove that animals follow a rule, additional evidence is needed to establish that the rule is a moral one and that the animals that follow it generally have moral motives. However, direct proof is impossible. The usual approach of animal behavioural scientists is more indirect. They try to find out whether an animal species possesses the capacities that are needed for acting morally. If the answer is affirmative, they still have to show that an explanation of a given behaviour in moral terms is the best one available.

If we want to find out what the morality is of another society, we usually start by looking for behaviour and practices that are similar to the behaviour and practices that fall within the scope of morality in our society. Students of animal behaviour do the same. If they do not observe patterns of behaviour in an animal species that look like what is regarded as moral behaviour in human society, they probably will not get interested in finding out if that species has a 'morality' or a morality.

6.3 Clusters of Moral Behaviour

Bekoff and Pierce propose to structure the patterns of the 'moral' behaviour of animals into three rough categories which they call 'clusters'. A cluster is a group of related behaviours that show some family resemblances. They identify three clusters: the cooperation cluster, the empathy cluster, and the justice cluster (Bekoff and Pierce 2009, p. 8). What kinds of behaviour are covered by these clusters?

I start with the cooperation cluster. Bekoff and Pierce use the term 'cooperation' to refer to a suite of behaviours related to helping others and working together with others towards a common goal. Cooperative behaviour includes grooming, group hunting, communal care of the young, alliance formation, and play. They also mention various 'mechanisms' that grease the wheels of cooperation: honesty, trust, punishment and revenge, spite, and the negotiation of conflicts (p. 59).

The second cluster is the empathy cluster. The name already suggests that its focus is on a capacity underlying certain behavioural patterns, and not on the patterns themselves.

The third cluster of 'moral' behavioural patterns is the justice cluster. In Bekoff and Pierce's view, this cluster comprises several behaviours related to fairness, including a desire for equity and a desire for and capacity to share reciprocally. It also includes various behavioural reactions to injustice, including retribution, indignation, and forgiveness, as well as reactions to justice such as pleasure, gratitude, and trust (Bekoff and Pierce 2009, p. 113).

Flack and De Waal (2000) do not speak of behavioural patterns, but of tendencies and capacities already present in non-human species, which cannot be missed in human morality. These tendencies, they say, deserve to be called the four ingredients of morality. These tendencies and capacities are:

1. *Sympathy related:* Attachment, succourance, and emotional contagion. Learned adjustment to and special treatment of the disabled and injured. Ability to trade places mentally with others: cognitive empathy.
2. *Norm related:* Prescriptive social rules. Internalization of rules and anticipation of punishment. A sense of social regularity and expectation about how one ought to be treated.

3. *Reciprocity:* A concept of giving, trading, and revenge. Moralistic aggression against violators of reciprocity rules.
4. *Getting along:* Peace making and avoidance of conflict. Community concern and maintenance of good relationships. Accommodation of conflicting interests through negotiation (Flack and De Waal 2000, p. 22).

There clearly is an overlap between Bekoff and Flack's categorization and that of Flack and De Waal. Both are, however, a mixed bag of competences, attitudes, mechanisms, and behavioural patterns. I prefer to keep behavioural patterns, competences, and mechanisms (such as rules) apart. The term 'cluster of behavioural patterns' is useful, but I suggest other names, related to what I consider to be the most important foci of morality. The first focus is the prevention, containment and regulation of violence and aggression. I call the related cluster the *anti-violence and aggression cluster*. The second focus is the distribution and the sharing of food (which may or may not be the fruit of cooperative activities). An appropriate name for the related cluster might be the *sharing and distribution cluster*. The third focus is on behaviour that goes beyond direct reciprocity, such as helping, and caring for weak and vulnerable conspecifics. *Helping and caring cluster* is a suitable name. Anticipating my discussion in Sect. 6.8, the last focus is on behaviour that is shown when norms are violated. I call this last cluster the *social disapproval and punishment cluster.*

6.4 Empathy, Concern for Others, and Helping Behaviour

Many students of animal behaviour suggest that the basic moral competences or capacities are already present in nonhuman primates. One of these competences or capacities is empathy. Bekoff and Pierce even call empathy the cornerstone of what in human society is called morality (2009, p. 87). As is well known from the literature in developmental psychology (Eisenberg 2000; Hoffman 2000), empathy is not a single behaviour. There is a whole class of behavioural patterns with varying degrees of complexity (Preston and De Waal 2002; Bekoff and Pierce 2009). It occurs in nested levels, with the inner core a necessary foundation for the other layers. The simplest forms of empathy are body mimicry and emotional contagion, largely automatic physiological responses. The next layer consists of somewhat more complex behaviours such as emotional empathy and targeted helping. Empathy of the two lowest levels can be found in mice for example. More complex is cognitive empathy, the capacity to feel another's emotion and to understand the reasons for it. Cognitive empathy appears to emerge developmentally and phylogenetically with other 'markers of mind', including perspective taking (PT), mirror self-recognition (MSR), deception, and tool use (Preston and De Waal 2002). According to Preston and De Waal cognitive empathy may be found in a wider range of species, in the hominoid primates and perhaps elephants, social carnivores, and cetaceans (whales, dolphins, and porpoises). Most complex is the capacity of attribution, in which an individual can take the other's perspective, which requires the use of imagination. According to Koski and Sterck (2009) the capacities of chimpanzees to understand others' emotional state operate at the level of what Hoffman calls 'quasi-egocentric empathy' – a complete separation between the own distress and that of the other has not yet been established. They would also be able to show initial other-regard. There is some evidence, for instance, that chimpanzees can attribute goals (Premack and Woodruff 1978; Call and Tomasello 1998). Research also suggests that nonhuman primates are sensitive to a conspecific's distress signals (e.g. Miller et al. 1963).

More insight in the role of empathy and concern for the distress of others in human morality is provided by Shaun Nichols (2004). He examined the moral capacities of very young children. Nichols builds on the distinction by Turiel and his colleagues between conventional and moral rules (Turiel 1983; Turiel et al. 1987). They contend that it is characteristic for moral persons to regard violation of moral rules as special on the dimensions of seriousness, wide applicability, authority, independence and justification. Violation of moral rules is above all serious when it causes harm to other people. Although the domain of morality is probably wider than that of harm-based violations, Nichols assumes that rules whose violation bring about harm constitute the core of morality. The capacity to see harm-based violations as very serious, generalizable, authority-independent and wrong because of well-being considerations, appears, according to Nichols, early in children's ontogenetic development – before their third year – and seems to be cross-culturally universal. Nichols calls this capacity the capacity for Core Moral Judgment (CMJ). CMJ depends on two mechanisms: a 'normative theory' prohibiting harming others and a basic altruistic motivation that is activated by representing suffering in others. In referring to the studies of psychologist Robert Blair (1995, 1997), Nichols contends that psychopaths, known to be deficient in affective response to the distress of others, do have a normative theory prohibiting harming others. A striking feature of psychopaths is that they provide conventional-type justifications for why violating moral rules is wrong, rather than offering justifications in terms of harm suffered by the victim. This leads Nichols to the conclusion that the normative theory is at least dissociable from the affective system. As far as I understand, a normative theory is for Nichols simply a system of norms. Whether animals can be said to have norms, we discuss in next section.

Nichols wants to know the cognitive and affective mechanisms underlying altruistic motivation. He argues that altruistic motivation depends on the minimal mind-reading (or empathic) capacity for an enduring representation of pain or some other negative affective or hedonic states in others. This minimal mind-reading capacity does not suffice for perspective taking. How can attributing distress to others lead to altruistic motivation? Nichols assumes that the altruistic motivation is mediated by an affective response. He gives two accounts of this affect. The available evidence does not really decide between these two. The first account is that there is a distinctive basic emotion of sympathy. The other is that distress attribution might produce a kind of second-order contagious distress in the subject. Representing the sorrow of another person may lead one to feel sorrow. This would produce an empathic response – to help for example. Nichols suggests that perhaps both affective mechanisms are operative. He introduces an overarching term for these two affective mechanisms: Concern Mechanism. Neither reactive distress nor concern requires, according to Nichols, sophisticated mind-reading abilities.

Nichols thinks it possible that at least some nonhuman animals have the mind-reading capacity to attribute distress to another (Nichols 2004, p. 60). Nichols rightly notes that it is unclear from the available data which mechanism is operative in nonhuman primates – whether it is a form of concern or reactive distress (Nichols 2004, p. 61).

6.5 Behavioural Regularities and Norms

Rules become visible in behavioural regularities, but not all behavioural regularities indicate the existence of a social rule. Habits are also behavioural regularities. For a group to have a habit it is enough that the behaviour of most of its members at certain occasions in fact converges. According to the highly influential philosopher of law Herbert Hart, a common behavioural regularity must be explained by a rule (norm) if (1) deviation of the regularity elicits criticism, (2) the deviation is generally accepted as a good reason for criticism, and (3) the norm is seen as binding and obligatory – Hart speaks of the internal aspect of rules, or looking at rules from the internal point of view (Hart 1961, pp. 53ff). In the next sections I examine whether the first and the third criterion are also useful for distinguishing norm-based behavioural regularities of animals from 'mere' behavioural regularities such as habits. I leave out the second criterion since reasons for criticism can only be expressed in language. Criticism as such can also be expressed in non-verbal form. Even humans often use other than linguistic means to show their disapproval of a given behaviour. They also express it by gestures, facial expressions, and sounds – all of them means of communication also available to animals. While the first criterion points to reactions of group members to norm transgression, the third criterion refers to the attitude of the agent towards the norm. When an agent only follows rules out of fear for sanctions, they are not binding and obligatory for him.

Hart's third criterion for distinguishing a statistical behavioural regularity from a rule-based one is the presence of the 'internal point of view'. The internal point of view with regard to norms or rules is the point of view taken by someone who has internalized the norm, or, in more technical terms, has the practical attitude of norm-acceptance. Someone who has internalized a norm is motivated to follow the norm by the reasons that are the rationale of the norm – the reasons why it exists.

In the next two sections I discuss the internal point of view with regard to norms in more detail. Section 6.6 deals with norm internalization, while Sect. 6.7 discusses the motivation to follow norms that results from the internalization of these norms. In Sect. 6.8 I deal with Hart's modified first criterion which says that it is a characteristic of social rules that their violation meets social disapproval.

6.6 Guidance by Norms in Human Morality

In his paper ‘Normative guidance’ Peter Railton explores central features of normative guidance, the mental states that underlie it, and its relation to our reasons for feeling and acting, using fictive examples describing everyday activities involving all sorts of norms (Railton 2006). He develops in several steps what he calls ‘a partial, largely functional characterization of conditions a piece of behaviour must meet to be norm-guided.’ This characterization applies to all norm-guided behaviour, not only to behaviour guided by moral norms. I skip these steps, and go right to the last formulation he gives, which I adapt – in his spirit – because I am here only interested in moral norms:

> Agent *A*’s conduct *C* is guided by norm *N* only if 1) *C* is the manifestation of *A*’s disposition to act in a way conducive to compliance with *N*, so that 2) *N* plays a regulative role in *A*’s *C-ing,* where this involves some disposition on *A*’s part 3) to notice failures to comply with *N*, 4) to feel shame or guilt[4] when this occurs, and 5) to exert effort to comply with *N* even when the departure from *N* is unsanctioned and non-consequential.

Condition 1 – the disposition to act in a way conducive to compliance with *N* – expresses that ‘To be norm-guided is a matter of how one is disposed to think, act, and feel, not simply of how one sees oneself, or would like to’ (2006, p. 7). Condition 2 – *N* plays a regulative role in *A*’s *C-ing* – says that reference to *N* must be a necessary part of the explanation of A’s behaviour. Condition 3 – the disposition to notice failures to comply with *N* – refers to the fact that A must monitor his behaviour because compliance with *N* matters to him. That it matters to him explains that he takes pains to comply with the norm even if non-compliance doesn’t cause a disadvantage to him and goes unnoticed by other people (condition 5). The sanctions are internal: feelings of shame and guilt (condition 4).

Railton is not satisfied with a functional characterization of conditions a given behaviour must meet to be norm-guided, and goes on to explore the distinctive role of norm-guidance in an agent’s psychology. He wants to know what mental acts or states of mind give a norm this sort of role in his life. He reviews several candidates that are discussed in recent philosophical literature: acceptance of norms, endorsing norms, and identification with norms. None of these attitudes accounts for the role of norms in shaping our lived world and contributing to the reasons why we act:

> Humble *internalization* of norms without the self’s permission, approval, or identification, like humble acquisition of beliefs without the benefit of judgment or reflection, provides much of our substance as agents. And the critical assessment and revision of norms that saves us from mere conformity and inertia, like the critical assessment and revision of what we believe, proceeds more often by trial-and-error feedback and unselfconscious readjustment over the course of experience than by spontaneous higher-order acts of endorsement or self-definition. (Railton 2006, pp. 31f.)

[4] Railton speaks here of discomfort, not of shame and guilt which are more specific moral feelings.

To this he adds that these higher-order acts do play a crucial role in making us candidates for moral agency and moral accomplishment. The distinction between humble internalization of norms and higher-order acts of endorsement or self-definition is important for our subject. Humble internalization might describe the role that norms play in guiding animal behaviour. Unlike humans, animals are not capable of endorsing norms and of self-identification with norms.

6.7 Motivation by Moral Norms

I said that guidance by (moral) norms does not require more than humble internalization of these norms. Humble internalization is all that is needed for norms to motivate behaviour. Suppose that someone confronts us with the argument that only agents are capable of being motivated by moral norms. Whatever else they may be, even our evolutionary next of kin, the nonhuman primates, are not capable of being motivated by moral norms. Mark Rowlands (2011) has developed an interesting response to this argument. Moral motivation, he says, is the mark not of being a moral *agent*, but of being a moral *subject*:

1. X is a moral *subject* if and only if X is, at least sometimes, motivated to act by moral considerations (p. 519).

The notion of a moral subject is, according to Rowlands, typically run together with that of a moral agent:

2. X is a moral *agent* if and only if X can be morally evaluated – praised or blamed (broadly understood) – for its motives and actions (p. 519).

In Rowlands's view, motivation and evaluation are conceptually distinct. Reasons that explain someone's action need not be *his* reasons. A reason can only be someone's reason if he has control over it. A reason is someone's reason if he understands that he has that reason. This is what is meant by an internal reason. We can blame or praise an agent for an action if the reasons that motivated the action were also his internal reasons. Motivating reasons need not be internal, they can also be external. An external reason exists for a subject whether or not he is aware of it or would endorse it. It is a reason *for him*, but not *his* reason. Rowlands argues that being – at least sometimes – motivated by (external) moral reasons is a necessary and sufficient condition for being a moral subject. This condition also applies to animals. It suffices for establishing that animals are moral subjects to determine that their behaviour is at least sometimes motivated by moral reasons. Rowlands realizes that his view is in agreement with Kantian as well as Aristotelian theories which regard moral subjects as reflective scrutinizers of their motivations and actions. His article is meant to show that it is implausible to suppose that the existence of a reflexive subject is a *necessary* condition of the possibility of moral motivation. I cannot reconstruct his complete argument. For reasons that will become clear later on, I focus on that part of the argument where he introduces someone whom he calls Mishkin[5] – after the prince in Dostoevsky's *The idiot*:

> *Prima facie*, Mishkin has the soul of a prince: throughout his life, he performs many acts that seem to be kind or compassionate. He performs these acts because he is the subject of sentiments that – again, at least *prima facie* – seem to be kind or compassionate ones. When he sees another suffering, he feels compelled to act to end or ameliorate that suffering. When he sees another happy, he feels happy because of what he sees. If he can help someone get what they want without hurting anyone else, he will help because he finds that he enjoys doing it. In short, Mishkin deplores the suffering of others and rejoices in their happiness. His actions reflect, and are caused by, these sentiments. What Mishkin does not do, however, is subject his sentiments and actions to critical moral scrutiny. Thus, he does not ever think to himself things like: 'Is what I am feeling the right feeling in the current situation – that is, is what I am feeling what I *should* be feeling?' Nor does he think to himself things like: 'Is what I propose to do in this circumstance the (morally) correct thing to do (all things considered)?' (Rowlands 2011, p. 528)

Rowlands supposes that Mishkin is not incapable of reflection, but operates on 'a more visceral level'. This is the picture that Rowlands gives us, after some discussion, of Mishkin:

> (i) Mishkin performs actions that are good, and (ii) Mishkin's motivation for performing these actions consists in feelings or sentiments that are the morally correct ones to have in the circumstances, and (iii) Mishkin has his own good reasons for having these feelings and performing these actions in these circumstances, and (iv) Mishkin is not aware of these reasons. (p. 531)

Mishkin has reasons. These reasons are internal, but not available to his conscious, rational scrutiny. They are embodied in his non-conscious, sub-personal processing operations. The terminology Rowlands uses here stems from social psychology. Most psychologists nowadays agree that there are two types of cognitive processes or 'reasoning systems'. Roughly, one system is associative and its computations reflect similarity and temporal structure; the other system is symbolic, and its computations reflect a rule structure (Sloman 1996). Stanovich and West labelled these systems or types of processes 'System I' and 'System II' (Stanovich and West 2000). There is now considerable agreement on the characteristics that distinguish the two systems. The operations of System I are fast, automatic, effortless, associative, and difficult to control or to modify. System I is cognitively impenetrable. The operations of System II are slower, serial, effortful, and deliberately controlled; they are also relatively flexible and potentially rule governed. The perceptual system and the intuitive operations of System I generate *impressions* of the attribute of objects of perception and thought. System II is uniquely human. Recent studies show that

[5] The correct spelling of the name of Dostoevsky's prince in *The idiot* (1868) is 'Mysjkin', but I will follow Rowlands's spelling. After a long stay abroad, prince Myshkin returns to Russia where he finds people under the spell of money. Myshkin is called the idiot because of his epileptic seizures.

most of human judgments are not simply the outcome of conscious – System II – reasoning. To a large extent, they are intuitive and automatic – System I – responses to challenges, elicited without awareness of underlying mental processes (Bargh 1996; Bargh and Chartrand 1999).[6]

6.8 Disapproval and Punishment

In their article ‘Evolutionary precursors of social norms in chimpanzees’ Claudia Rudolf von Rohr, Judith Burkart and Carel van Schaik (2011) develop a theoretical framework for recognizing different functional levels of social norms and distinguishing them from mere statistical regularities. They define social norms as behavioural regularities that are normative (i.e. entail a sense of oughtness in the moral sense) to a varying degree, and generate social expectations. These expectations do not have to be experienced consciously. This aligns with what I said about humble internalization of norms. Their satisfaction or violation might, according to Von Rohr et al., produce distinct reactions observable from the outside. Since meeting expectations is the normal situation, no reactions have to follow. But when a given behaviour violates expectations, nearly always negative reactions ensue. Most important are the negative reactions by uninvolved bystanders (Von Rohr et al. 2011, pp. 3ff.). Von Rohr et al. distinguish three types of negative reactions from bystanders on the violation of three different types of norms:

1. *Quasi social norms*. The negative reactions might simply be caused by specific cues. E.g., when an infant that is attacked, screams, bystanders flow to the scene and harass the perpetrator. This type of bystander reaction does not reflect violated social expectations, and most likely does not involve emotions such as indignation towards the perpetrator. Bystanders in this category probable do not possess any specific inference on how the distress of an infant and the behaviour of the perpetrator are linked together and thus are not able to perceive harming an infant as norm violation per se (2011, p. 16).
2. *Proto social norms.* If bystander reactions cannot be explained by simple stimulus-response mechanisms, it might be that they respond to norm violation as such. In this case, bystander reactions might also involve emotions comparable to indignation in humans. Bystander reaction on norm violation per se requires

[6] Rowlands’s Mishkin is someone whose morality seems to operate completely on the subconscious level. Hubert and Stuart Dreyfus (1991) would characterize him as a ‘moral expert’, the product of successful moral education and training whose judgments and decisions are the product of intuitive thinking. I don’t have problems with Rowlands’s characterization of Mishkin as a moral subject, but his picture of Mishkin is over-simplified. Even moral experts sometimes have to reason consciously, e.g., when their intuitions conflict or are indeterminate. Moreover, it is hard to imagine that a person is never challenged to give reasons for his judgments and actions. Therefore Mishkin must be capable to reflect on his motivation. Mishkin may most of the time function as a moral subject, but he must be capable of moral agency.

the capacity to exhibit some empathetic competence because this would enable bystanders to understand the mistreated infant's and its mother's distress to some extent, and also its cause. Von Rohr et al. assume that apes but not monkeys have empathetic competence, because monkeys seem to lack the capacity to attribute mental states to others (2011, pp. 16ff).

3. *Collective social norms.* Humans are endowed with sophisticated empathetic and cognitive abilities, which enable them to grasp the full extent and far-reaching consequences of mistreating children. Moreover, they are able to reason that infants are completely defenceless and therefore highly vulnerable creatures (2011, p. 17). An important difference between the reactions of chimpanzees and humans on norm violation is that chimpanzees might experience 'indignation' in a fairly individualistic way, while humans are able to share their feelings of indignation. Referring to Tomasello and Carpenter (2007), Von Rohr et al. state that, in analogy to shared intentionality, shared indignation goes beyond simultaneous experience by different individuals and includes the awareness of a collective experience which may lead to collective protest against and condemnation of the violator. This exemplifies the collective nature of a social norm. 'It is this collectivity upon which the viability and the enforceability of a social norm ultimately rests and which on current evidence appears to be absent in chimpanzees' (Von Rohr et al. 2011, p. 18).

Negative reactions by non-involved bystanders on the deviation of a socially expected behavioural regularity, accompanied by feelings comparable to human indignation, indicate that a social norm lies at the base of the behavioural regularity. According to Von Rohr et al. this kind of negative reactions requires the capacity to exhibit some empathetic competence; a capacity which is according to them present in chimpanzees, and not in monkeys and other species which lack the capacity to attribute mental states to others (Tomasello and Call 1997). They conclude that norms might play a role in guiding the behaviour of chimpanzees; these norms are not collective social norms however. They are proto-social norms. Since in their view moral norms are collective social norms, the conclusion must be that only humans have a morality.

Although I am inclined to accept that only humans are capable of shared indignation, I am not convinced that shared indignation marks the violation of the kind of social norms we call moral norms. An important step in their argument is the distinction that Von Rohr et al. make between personal norms and social norms. A personal norm refers to a personal expectation about how an individual wants to be treated. Personal norms are precursors of social norms because it seems implausible that one would form expectations about how others should be treated before forming expectations about how one wants to be treated oneself. Moral behaviour, they say, starts when personal expectations are generalized and extended to others. It seems that they call norms personal if violation of a norm elicits a negative reaction only from the individual that is negatively affected. I find this concept of a personal norm implausible. If I punish my neighbour when he does not bow to me, because I personally expect that younger persons should bow to older ones, this clearly is a personal norm – provided that I myself also bow to older persons. Nobody else punishes youngsters who do not bow for older persons. The norm is not shared by others. Suppose I am talking to some neighbours at the back of my house, when I see a stranger climbing over my fence. I get angry at that person and shout that he has no right to enter my garden without my consent. Although I am the only one who starts shouting, my neighbours approve of my reaction and would do the same if someone climbed over their fence. I consider a norm to be a social norm if every individual in a group reacts negatively when the violation of a norm directly affects them.

6.9 Animal Morality and Human Morality

In the previous sections I argued that there is evidence that some animal species, chimpanzees for example, display disapproval of norm-violating behaviour. I do not know of evidence proving that behaviour of higher animal species, say, the nonhuman primates, is motivated by humbly internalized moral norms. It might even be impossible to prove that the behaviour of non-linguistic animals results from following moral norms. Rowlands only showed that moral motivation is not restricted to the class of moral agents. He made it plausible that a species has a morality, if its members show – at least sometimes – behaviour that is morally motivated. If I understand him well, Rowlands argues that animals that belong to a species which might have a morality are similar to Mishkin. Like Mishkin, they are, at least sometimes, unconsciously motivated by moral reasons of which they are not aware. Also in Railton's view, guidance of human behaviour by norms largely takes place unconsciously and automatically. This type of guidance can be found with animals as well. Not only a large part of human behaviour generally, but also a large part of their moral judgments results from an embodied, unconscious, and pre-reflective morality. It is plausible to assume that this pre-reflective morality is still influenced by the human evolutionary heritage – (social) motivations, emotions, learning and other dispositions, and inhibitions. But it is also probable that the influence of this evolutionary heritage can be controlled, inhibited, modified, extended by reflective capacities. Moreover, the embodied non-reflective morality of humans contains more than what they have inherited of their evolutionary predecessors. Many of a person's intuitive judgments are the result of indirect or direct learning processes.

Railton's characterization enables us to get a clearer picture of how norm guidance with animals relates to that with humans. A large part also of human morality is non-reflective and functions in the same manner as animal morality. But there are important differences. Animals do not experience shame or guilt when they don't comply with their moral rules. Mishkin, who doesn't show moral agency, is still a different moral subject than e.g. a chimp. It is important to stress that unconscious guidance by rules doesn't imply that the guidance is irrational. It isn't rational in the sense of consciously based on reasons. It can however be rational in the sense that it conforms to standards of rationality.

Non-reflective animal morality can only engage System I-processes. Human morality can be both non-reflective and reflective, and makes use of both System I and System II processes. In contradistinction to System II cognitive processes which are rule based and computational, System I processes are said to be associative and/or heuristic based. This image is, according to Peter Carruthers (2012), wrong. He refers to research by Gallistel and colleagues on conditioning of animals that shows that the behaviour of animals involved in conditioning experiments is best explained in rule-governed, computational terms, rather than in terms of associative strengths (Gallistel and Gibbon 2001; Gallistel and King 2009). Carruthers concludes that non-human animals engage in unreflective processes that can be both flexible and rule governed. Otherwise learning by animals could not be explained.

Many of the rules and principles that guide my behaviour are 'humbly internalized', even principles such as autonomy – in the Kantian sense – and self-determination. From time to time however we are confronted by situations in which 'humble internalization' doesn't suffice, for example, a confrontation with people with diverging moral beliefs and practices. Only then the question arises whether we can accept, endorse, or identify with our internalized rules and principles. Only then conscious reflection is needed. That is when System II processes need to be activated. Reflection may lead to revising rules and principles. What distinguishes humans from moral animals is that humans are *capable* of moral agency, although they rarely *show* moral agency.

According to Carruthers, System I and System II processes cooperate. He holds that rational reflection operates in this way: action schemata are selected and activated, and mentally rehearsed, while overt action is suppressed. This gives rise to conscious images which are globally broadcast and thus made available as input to the full suite of intuitive systems. The latter draw inferences from them, activate relevant memories, and issue in emotional reactions. Thus, reflection doesn't exist alongside of intuitive systems, but is partly realized in cycles of operation of the latter, utilizing pre-existing mechanisms and capacities (Carruthers 2012, p. 8).

6.10 Animal Ethics and Animal Morality

Animal ethics works, as I said in the Introduction, with the distinction between moral agents and moral patients. Animals are moral patients if they have sentience. Lacking reflective capacities, animals cannot be moral agents. This is currently the dominant view among ethicists. I have argued that some animals can be said to have not just a 'morality', but a morality in the full sense. In this last section I only explore what being a moral animal means for the moral status of the animal. I plan to go into this issue in a subsequent article. Do humans have other duties to moral animals than to animals which only have a 'morality'? Do moral animals themselves have moral duties? I start with the first question. Only vegetarians reject killing animals for food. The more received view is that humans have a duty to withhold from all kinds of actions that cause harm to animals, but have no duty not to kill them. Thus, it is morally permissible to kill animals, provided that it is done painlessly. This view assumes that there are morally relevant differences between humans and animals. In my opinion, moral animals do not differ from humans in morally relevant respects. Moral animals are, as are humans, worthy of our respect. Here I cannot work out this view in detail.[7]

The second question asks whether moral animals are duty bearers. The received view is that only agents are duty bearers. By speaking of proto-morality Flack and De Waal avoid difficult questions such as whether (some) animals have agency, questions which Bekoff and Pierce are forced to answer. Many people assume, they state, that by claiming that animals have morality, we are also claiming that they are moral agents (2009, p. 144). Bekoff and Pierce accept what they regard as the philosophical implications of their position: One cannot argue that (some) animals have a morality while denying that they have agency. However, they argue that it would be naïve to assert that other animals are moral agents *in the same sense* in which most adult humans are. Moral agency is species specific and context specific; animals are moral agents *within the limited context of their own community.* Wolf morality reflects a code of conduct that guides the behaviour of wolves within a given community of wolves. The predatory behaviour of a wolf towards an elk is *amoral* (2009, pp. 144 ff.).

By embracing relativism, Bekoff and Pierce avoid difficult questions. They concede that animals who are moral agents in principle deserve blame or praise for their behaviour. From a relativistic point of view, moral agents are not to be blamed for actions that are not wrong within the morality of their own group (or species). Animal moral agents can only be blamed for what they do to members of their own community. An elk does not belong to the moral community of wolves. Killing an elk is, in the view of Bekoff and Pierce, not a moral issue for a wolf. This is, of course, a very simple case. It is hard to imagine that predatory animals are to be blamed for killing their prey. A more difficult case would be that of chimpanzees who are systematically murdering the members of a rival group. Assuming that Bekoff and Pierce are consistent, they would regard the behaviour of these chimpanzees as not blameworthy, because the rival group doesn't belong to their moral community. Suppose that a subordinate chimpanzee takes food that is not seen by the dominant one. Is his behaviour blameworthy? Of course, praising and blaming is a linguistic practice which is not available for animals. They are neither capable of praising and blaming, nor of understanding praise and blame of others. Even if it doesn't make sense to blame someone, it still might make sense to ask whether the individual is blameworthy. In my view, an individual is only blameworthy if he is aware of the norm he transgressed, or can be made aware of that. Therefore, moral animals who are only unconsciously motivated by moral norms, cannot be blamed for moral transgressions. That is why Rowlands calls them moral subjects, and not moral agents.

[7] An important implication of this view is for me that moral animals cannot be used for food, nor for research purposes.

6.11 Conclusion

I argued that some animal species may have not just a morality-analogous system for the regulation of social behaviour, but a morality. These species disapprove of behaviour that conflicts with moral rules. There are also important similarities between animal morality/moralities and human morality/moralities. The moral rules and behavioural patterns of both animal and human morality can be clustered in the same categories: the anti-violence and aggression cluster, the sharing and distribution cluster, the helping and caring cluster, and the social disapproval and punishment cluster. Animal morality regulates behaviour automatically and unconsciously. However, a large part also of human morality is non-reflective and functions in the same manner as animal morality. Humans differ from animals in that they possess rational and reflective capacities. These capacities enable them to deliberate about what is right and appropriate to do and constitute the core of agency. Humans can be held accountable and responsible for what they do. In the last part of the paper I explored what being a moral animal (having a morality) means for the moral status of the animal. I argued that moral animals are worthy of the same respect that we owe to humans, which implies that we should not kill them for food or use them in medical trials that benefit only humans.

References

Bargh, J.A. 1996. Automaticity in social psychology. In *Social psychology: Handbook of basic principles*, ed. E.T. Higgins and A.W. Krugalski, 169–183. New York: Guilford.

Bargh, J.A., and T.L. Chartrand. 1999. The unbearable automaticity of being. *American Psychologist* 54: 462–479.

Bekoff, M., and J. Pierce. 2009. *Wild justice. The moral lives of animals*. Chicago/London: The University of Chicago Press.

Blair, R.J.R. 1995. A cognitive developmental approach to morality: Investigating the psychopath. *Cognition* 57: 1–39.

Blair, R.J.R. 1997. Moral reasoning and the child with psychopathic tendencies. *Personality and Individual Difference* 22: 731–739.

Call, J., and M. Tomasello. 1998. Distinguishing intentional from accidental actions in orangutans, chimpanzees, and human children. *Journal of Comparative Psychology* 112: 192–206.

Carruthers, P. 2012. The fragmentation of reasoning. http://www.philosophy.umd.edu/Faculty/pcarruthers/The%20Fragmentation%20of%20Reasoning.pdf. In *La coevolución de mente y lenguaje. Ontogénesis y filogénesis*, ed. P. Quintanilla. Lima: Fondo Editorial de la Pontificia Universidad Católica del Perú.

De Waal, F.B.M. 2009. In *Primates and philosophers. How morality evolved*, ed. S. Macedo and J. Ober. Princeton: Princeton University Press.

Dreyfus, H.L., and S.E. Dreyfus. 1991. Towards a phenomenology of ethical expertise. *Human Studies* 14: 229–250.

Eisenberg, N. 2000. Emotion, regulation, and moral development. *Annual Review of Psychology* 51: 665–697.

Evans, E. [1906]/1987. *The criminal prosecution and capital punishment of animals*. Whitefield: Kessinger Publications Company.

Flack, J.C., and F.B.M. De Waal. 2000. ‘Any animal whatever’. Darwinian building blocks of morality in monkeys and apes. *Journal of Consciousness Studies* 7: 1–29.

Gallistel, C.R., and J. Gibbon. 2001. Time, rate and conditioning. *Psychological Review* 108: 289–344.

Gallistel, C.R., and A. King. 2009. *Memory and the computational brain*. New York/ Chichester/Malden: Wiley-Blackwell.

Girgen, J. 2003. The historical and contemporary prosecution and punishment of animals. *Animal Law* 9: 97–133.

Hart, H.L.A. 1961. *The concept of law*. Oxford: Oxford University Press.

Hoffman, M.L. 2000. *Empathy and moral development: Implications for caring and justice*. Cambridge: Cambridge University Press.

Koski, S.E., and E.H.M. Sterck. 2009. Empathetic chimpanzees. A proposal of the levels of emotional and cognitive processing in chimpanzee empathy. *European Journal of Development Psychology* 7: 38–66.

Miller, R., J. Banks, and N. Ogawa. 1963. Role of facial expression in ‘cooperative-avoidance conditioning’ in monkeys. *Journal of Abnormal and Social Psychology* 67: 24–30.

Nichols, S. 2004. *The sentimental rules*. Oxford: Oxford University Press.

Peterson, D. 2011. *The moral lives of animals*. New York: Bloomsbury Press.

Premack, D., and G. Woodruff. 1978. Does the chimpanzee have a theory of mind? *The Behavioral and Brain Sciences* 1: 516–526.

Preston, S.D., and F.B.M. De Waal. 2002. Empathy. Its ultimate and proximate bases. *Behavioural and Brain Sciences* 1: 1–20.

Railton, P. 2006. Normative guidance. In *Oxford studies in metaethics*, vol. 1, ed. R. Shafer-Landau, 3–35. Oxford: Oxford University Press.

Rowlands, M. 2011. Animals that act for moral reasons. In *The Oxford handbook of animal ethics*, ed. T.L. Beauchamp and R.G. Frey, 519–547. Oxford: Oxford University Press.

Sloman, S.A. 1996. The empirical case for two systems of reasoning. *Psychological Bulletin* 119: 3–22.

Stanovich, K.E., and R.F. West. 2000. Individual differences in reasoning: Implications for the rationality debate? *Behavioural and Brain Sciences* 23: 645–665.

Tomasello, M., and J. Call. 1997. *Primate cognition*. New York: Oxford University Press.

Tomasello, M., and M. Carpenter. 2007. Shared intentionality. *Development Science* 10: 121–125.

Turiel, E. 1983. *The development of social knowledge: Morality and convention*. Cambridge: Cambridge University Press.

Turiel, E., M. Killen, and C. Helwig. 1987. Morality: Its structure, functions, and vagaries. In *The emergence of morality in young children*, ed. J. Kagan and S. Lamb, 155–244. Chicago: University of Chicago Press.

von Rohr, C.R., J. Burkart, and C. van Schaik. 2011. Evolutionary precursors of social norms in chimpanzees: A new approach. *Biology and Philosophy* 26: 1–30.

Warnock, G.J. 1971. *The object of morality*. London: Methuen.

Chapter 7
Two Kinds of Moral Competence: Moral Agent, Moral Judge

Florian Cova

7.1 What Makes Us Moral? And the Continuism/Discontinuism Debate

There are many ways of understanding the Big Question 'what makes us moral?' One way to understand it is to understand it as bearing on the psychological capacities that allow us to be moral beings. Thus, it can be paraphrased in the following way:

(BQ) *What psychological capacities allowed us to become moral beings?*

Note that the question is not what capacities are *necessary* for a being to count as a moral being. For all we know, there might be different capacities that are separately sufficient, but not necessary, to make a being a moral being. What we are interested in here is what capacities *actually* made us (and continue to make us) moral beings.

This question is undoubtedly fascinating, and this is why it is currently investigated by philosophers, psychologists, neuroscientists, evolutionary biologists, economists, anthropologists and others I surely forgot. It would be preposterous to claim that I actually have the answer to this question, so my goal won't be to answer it. My goal is rather to point out that a preliminary question must be answered if we are to find the answer to our Big Question. To discover what made us moral, we must first understand what it is 'to be moral' and what it means to be a 'moral being'.

Indeed, answers to (BQ) come into two very different forms: *Continuists* believe that the psychological capacities required for being a moral being are not unique to humans and that at least rudimentary traces of these capacities can be found in other species (most likely apes). *Discontinuists*, on the contrary, believe that these psychological capacities are unique to humans and that human beings are the only moral beings we know of.

F. Cova (✉)
Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland
e-mail: florian.cova@gmail.com

B. Musschenga and A. van Harskamp (eds.), *What Makes Us Moral? On the capacities and conditions for being moral*, Library of Ethics and Applied Philosophy 31,
DOI 10.1007/978-94-007-6343-2_7, © Springer Science+Business Media Dordrecht 2013

A paradigmatic example of this debate is offered in Frans de Waal's book *Primates and philosophers: How morality evolved*. In this book, De Waal takes a continuist stance in describing the school of thought he defends:

> [This school] views morality as a direct outgrowth of the social instincts that we share with other animals. In the latter view, morality is neither unique to us nor a conscious decision taken at a specific point in time: it is the product of social evolution. (De Waal 1997, p. 6)

De Waal champions this theory by pointing to the different psychological capacities he claims we share with apes (perspective taking, emotions like gratitude) and to reports of 'moral behaviours' (help, consolation). This line of argument directly clashes with the conception of certain of his respondents. For example, Christine Korsgaard (De Waal 1997, p. 104) claims that 'morality represents a break with our animal past'. And indeed, in accordance with this conception, some psychologists have argued that morality is the product of new psychological capacities that are not to be found in non-human animals, some 'sense of fairness' (Baumard 2010) or 'universal moral grammar' (Mikhail 2011).

These two theses seem to me equally plausible, though they seem to cancel one another: when someone describes the delightful stories of seemingly moral behaviour in the animal realm, we are moved, and find the conditions set by the discontinuist theory too high. But when we listen to advocates of the human specificity, we can't help notice that there's really something special about human beings, i.e. the complexities of their interrogations about what is right or wrong. After all, we are the only known species to do moral philosophy.

Here, I want to loosen this tension by suggesting that the disagreement between continuists and discontinuists can be (partly) resolved because it (partly) stems from the conflation between two ways of counting as a moral being. First, one can be a moral being because one is morally responsible of (some of) his action: thus, one is a moral being in the sense of being a *moral agent*. Then, one can be a moral being in the sense that one is able to judge whether something is right or wrong: in this sense, one is a moral being in the sense of being a *moral judge*. Indeed, once this distinction made one can argue that both sides are right (and wrong) respective to one sense of counting as a moral being. More precisely, I will argue that it is possible to be a continuist about moral agency while being a discontinuist about moral judgment, and argue that we share the psychological bases for moral agency with other animals while we are the only known species able to form moral judgments.

Consequently, the main point I propose to defend here will be that these two ways of being moral can be dissociated and that one can be a moral agent without being a moral judge. I won't try to argue for a *double* dissociation and to prove that there are also cases of moral judges that are not moral agents. Maybe such cases can be found in children (that we do not always judge responsible for their action but that are able to form at least rudimentary moral evaluations)[1] or in patients suffering from

[1] Famous psychological studies led by Turiel suggest that, by the age of 4, children are able to understand moral concepts and use them to form moral judgments. For example, they are

particular neuropsychological deficits.[2] Nevertheless, such cases being about non-matured or pathological individuals that are not representative of the full potential of their species, they won't be relevant to the present discussion. I will then stick to the more plausible claim that a species can be endowed with the psychological requirements for moral agency while being unable to form moral judgment.

7.2 The Epistemic Argument Against the Moral Agency/Moral Judgment Dissociation

I will take as a starting point that it is not *prima facie* implausible that one can be a moral agent without being a moral judge. As a proof, we can see that one of the foundational myths of occidental culture, in which many people have believed, relies on such a dissociation. Let's go back to the famous episode of *Genesis*, when Adam and Eve eat the forbidden fruit. What happens then? What is said is that, by eating the fruit, their 'eyes opened' and they came to 'know good and evil' (*Genesis* 3:5). Now, remember that Adam and Eve are finally punished for having eaten the fruit: so, by eating the fruit, they did something wrong. Nevertheless, at the time they weren't able to recognize right from wrong yet. So, what is suggested by this story is that a being can be morally responsible for an action he performed at a time he wasn't able to recognize right from wrong.

Once granted that it is not obvious that moral judgment is required for moral agency and that one cannot be a moral agent without being a moral judge, I will argue that the burden of the proof lies with those who consider moral judgment as a requirement for moral agency. The thesis that one cannot be a moral agent without being a moral judge amounts to the thesis that the two concepts of 'moral agency' and 'moral ignorance' are incompatible.[3] But, since each concept can be understood separately, and since two logically independent concepts are *prima facie* compatible, it seems to be the position by default that moral agency and moral ignorance are compatible. Thus, if there is no argument in favour of the thesis that moral agency and moral ignorance are incompatible, we are justified in thinking that they are compatible.

able to distinguish moral rules (that are universally valid and independent from authority) from conventional rules (that are only locally valid and dependent on authority) or prudential rules (see Turiel 2002 for a review).

[2] For example, patients with lesions in the prefrontal cortex are still able to make moral judgments, but are much less prone to act according to these judgments, due to emotional and motivational deficits (Damasio 1995). Patients suffering from aboulia, who have lost all motivation, are also plausible cases of people able to form moral judgments but lacking moral motivations (Marit and Wilkosz 2005).

[3] I define 'moral agency' as the ability to be morally responsible for (some of) one's action and 'moral ignorance' as the inability to judge something as morally right or wrong.

Nevertheless, there are such arguments. So, before concluding that moral agency and moral ignorance are compatible, we must examine these arguments. I will distinguish two main kinds of arguments, each one corresponding to a certain kind of condition for moral responsibility. It is usual in the literature about freedom and moral responsibility to distinguish two categories of conditions for moral responsibility: the *epistemic* conditions (what one has to know for being responsible for one's action) and the *freedom* conditions (the type of control one has to exert on one's action for being responsible).

As far as the epistemic conditions are concerned, it is commonplace to say that one had to know and understand what one was doing for being morally responsible. Some have understood this condition as implying that one must be able to understand the moral significance of one's action to be morally responsible for them. But if this is true, then moral judgment is a condition for moral agency, and one cannot be a moral agent without the capacity to form moral judgment.

7.2.1 The Epistemic Conditions for Moral Responsibility

To give an example of the debate about the epistemic conditions for moral responsibility, let's have a look at the debate on the moral responsibility of psychopathic individuals. Psychological studies have recently suggested that psychopaths are complicated cases: it is not even sure that they are able to form genuine moral judgments. Studies suggest that they fail to distinguish between moral rules and conventional rules, a distinction 'normal' people master around the age of 4.[4] For this reason, people like Levy have suggested that we should not hold psychopaths responsible for their action: 'Moral responsibility requires moral knowledge; because psychopaths lack this knowledge through no fault of their own, we must refrain from blaming them. Psychopaths are victims, as well as victimizers' (Levy 2008, p. 136).

The principle according to which moral responsibility requires moral knowledge is widely endorsed. For example, a similar principle is part of the *Model Penal Code* of the American Law Institute:

> A person is not responsible for criminal conduct if at the time of such conduct as a result of mental disease or defect he lacks substantial capacity either to appreciate the wrongfulness of his conduct or to conform his conduct to the requirements of the law.

Now, this principle is just a way among others to state that moral agency requires moral knowledge – but what reasons do we have to endorse this principle?

One reason I see is that this principle seems to correspond to a kind of excuses we use to accept for wrongdoing. For example, someone who has harmed another accidentally can say that 'he didn't know he was doing something wrong'. I don't think that this is enough to support this principle. Surely, if we imagine that someone

[4] See Blair (1995, 1997) and footnote 1 in this text.

offers a child a peanut butter sandwich without knowing that the child is allergic to peanut butter, and that this action results in the child being driven to the hospital, we would accept an excuse such as ‘I didn’t know I was doing something wrong’ (meaning: I didn’t know that my action would have such dreadful consequences). But we would be less willing to accept it if the person actually intended to kill the child by poisoning her with peanut butter and tried to escape blame by saying she didn’t know that poisoning was wrong.

Contrasting these two examples, we can draw a distinction already made by Aristotle (2011, III, 2) between two kinds of moral ignorance: ignorance of the particular circumstances of an act (what exactly has happened) and ignorance of the moral value of this act (if the act was permissible or forbidden). According to him, only the first kind of ignorance can be a valid excuse, the second kind of ignorance being the mark of vice. Granted, things are complicated because Aristotle also thinks that, ultimately, moral ignorance in this second sense is the result of the agent’s choice. So, if moral ignorance in this sense is not an exculpatory condition, that would be because the agent is ultimately responsible for it. I do not wish to enter in an exegesis of Aristotle’s conception of moral ignorance: all I am interested in here is the distinction between moral ignorance of the particular circumstances of an action (*factual ignorance*) and moral ignorance of the moral value of this action (*moral ignorance* in the strong sense). This distinction allows us to see that even if it is intuitive to say of someone that he should be forgiven because he didn’t know what he was doing, such expressions are more commonly used in cases of factual ignorance, when we excuse an agent because he didn’t have a full understanding of the impact and consequences of his action.

Thus, I don’t think it is obvious that moral ignorance is a mitigating circumstance. Though this has been accepted as a valid principle by philosophers and courts, the idea seems to have no clear intuitive appeal. One explanation for its acceptance is that it is the product of an overgeneralization of the intuitive principle according to which ignorance of the nature of an action can be a valid excuse, that ended up including in the ‘nature’ of an action its moral value.

7.2.2 Moral Knowledge and Acting for Good Reasons

Another reason to endorse this principle is that it might help us explain the difference between good actions done from good intentions (that are praiseworthy) and good actions that are done without good intentions (and often are not praiseworthy). Let’s take the following (paradigmatic) example[5]:

> (Pond) As a man walks by a pond, he notices a young child drowning. He dives into the pond to save the child and brings the child back to the shore.

[5] This case is inspired by Singer (1972).

Now, should we praise this man for having saved the child? It depends on *why* he saved the child. As a philosopher will justly point out, we shouldn't (and wouldn't) praise this man if he didn't care at all about the child and saved it only because he hoped to receive a reward or to be considered as a hero. This means that, to be praiseworthy, our man must have acted *for the good reasons*.

But what does it mean to act *for the good reasons*? Following Kant (2011), a philosopher could advance the following thesis:

> (Reflectivism) An agent deserved to be praised (or blamed) for having performed a given action A only if he did so because he considered this action to be good (or thought this action to be bad).

This principle gives a very simple answer to our problem: the man who saves the child only because he cared about the reward does not deserve praise because he did not do it because he thought it to be good (i.e. by reason of respect for the moral law, to use a Kantian expression). Though very simple, Reflectivism has an important implication: that an agent can be morally responsible for an action only if he has the capacity to judge something as morally good (or wrong). Thus, the need to distinguish between praiseworthy good actions and good actions not performed for the good reasons might be a support for this principle, because it would be the best way to distinguish the two kinds of good actions.

But are we forced to endorse Reflectivism if we are to make a difference between these two kinds of good actions? I don't think so. To show why, I'll propose a number of counter-arguments. All of them rely on appeal to intuitions, and so are not strong enough to conclude that Reflectivism is false. Nevertheless, I'll argue they are sufficient to show that we have no reason to endorse Reflectivism.

Let's start with (i) the *argument from impulsivity*. Here is a slightly modified version of Pond:

> (Pond*) As a man walks by a pond, he notices a young child drowning. Understanding that the child will soon die, he immediately dives into the pond to save the child, without taking the time to figure whether it's the right thing to do. He succeeds in retrieving the child.

And here is a second variant:

> (Pond**) As a man walks by a pond, he notices a young child drowning. Due to his strict moral education, this man has taken the habit (and is motivated) to do what he thinks is right. Realizing that the children will soon die and that saving him would be something right, he dives into the pond to save the child. He succeeds in retrieving the child.

These two men have different reasons for acting. Our first man doesn't take time to think: he doesn't classify his action as 'right' before acting. If we asked him about his reasons for acting, he would answer something like: 'because the child would have died'. Our second man goes through some kind of moral reasoning, and saves the child because he categorizes this action as 'right'. If we asked him about his reasons for acting, he would answer: 'because it was the right thing to do'. According to Reflectivism, only the second one should be considered praiseworthy, because he's the only one to act on the basis of a moral judgment (the judgment that it is right to save the child and wrong to let him die).

But does this conclusion seem right? I don't think so. I rather think that, if we compare those two cases, the man in Pond* is at least as praiseworthy as the man in Pond**. In fact, there even seems to be something wrong with our second man: shouldn't he be more concerned about the child rather than whether it is right or wrong to save him?

Reflectivism, in the way I presented it, seems to discredit every good action that would look like a 'moral reflex'. If your friend is about to stumble, and you instinctively grab him to prevent his fall, then you shouldn't be praised for your action, since you did that only because you cared about your friend, and not because you judged that helping your friend was the right thing to do. And Reflectivism doesn't stop there: it also discredits actions that come from emotional reactions and are not mediated by moral reasoning. If a man, seeing a homeless person freezing in the winter is suddenly overwhelmed by compassion and gives him his coat, without wondering whether this is right or not, but just wanting to help this particular person, then Reflectivism should conclude that his action is not praiseworthy. However, we tend to praise and even to be moved by such actions. Finally, Reflectivism seems to lead to the conclusion that, when a friend or a parent helps us, he is all the more praiseworthy for helping us because it was the right thing to do. But, this doesn't seem right: we do not want our friends or our parents to help us because they think it's the right thing to do – we want them to help us because they actually *care* about us. So, to sum up, Reflectivism goes against most of our basic moral appreciations.[6]

One could object that we focused on cases in which we praise agents for right actions, and that Reflectivism is much more plausible when it comes to wrong actions: is it not intuitive that we shouldn't blame (and punish) those who didn't know that what they did was wrong?

This is far from clear. Let's imagine the following case (drawn from Pizarro et al. 2003):

> (Smash) Because of his overwhelming and uncontrollable anger, Jack impulsively smashed the window of the car parked in front of him because it was parked too close to his.

Let's say that Jack had a bad day, was irritated, and smashed the window without taking the time to assess whether it was right or wrong. Let's also say that, though he realized afterwards that it was the wrong thing to do, he did not regret this action at great length. Should we say that Jack is not responsible and does not deserve blame for what he did? That he hasn't the duty to pay for repairs? That seems very counter-intuitive.[7]

[6] One might say that these appreciations are not really *moral* and that the praise we attribute agents for caring about their relative has nothing to do with moral praise. This is indeed a possibility; nevertheless it seems me very unlikely: a mother who doesn't care about her children elicits from us blame that is very likely *moral* blame.

[7] The results obtained by Pizarro et al. show that many participants considered Jack responsible and blameworthy for having smashed the window. That Jack had not taken the time to realize that what he was doing was wrong was no excuse.

Another argument that could be opposed to Reflectivism are the cases of '*inverse akrasia*'. By 'inverse akrasia', Arpaly and Schroeder (1999) mean cases in which the agent fails to do what he thinks is right and does what he thinks is wrong, but in which we actually consider wrong what he failed to do and right what he actually did. A famous example is the case of Huckleberry Finn, who keeps helping Jim the runaway slave while thinking that the right course of action is to return the slave to his lawful owner. In this case, Kantianism should predict that we shouldn't praise Huckleberry for helping Jim, since he doesn't think that it is the right thing to do.

Still, many people still judge that Huckleberry is responsible and deserves praise for his actions. So, it is still not intuitive that one has to know that he's doing something right (or wrong) to be responsible for his actions.

One might object that inverse akrasia is no definitive objection to Reflectivism if we switch to a weaker version according to which for an agent to be morally responsible for his actions, this agent has just to assess the moral value of what he's doing, but not to *correctly* assess this value. If we drop the 'correctly', then Reflectivism can accommodate these cases, because Huckleberry (incorrectly) assesses the moral value of their actions.

This less demanding version still can't account for cases of impulsive actions. Surely, Reflectivism could once again lower its demands to become able to accommodate such cases. For example, it could say that people don't actually have to act upon the basis of moral judgments to be responsible for their actions – but that they must only have the ability to judge that what they have done is right or wrong. In this version however, Reflectivism loses one of its main advantages and motivations: it can no longer explain the difference between the man who saves the child because he cares about it and the man who saves the child because he expects a reward – both have the ability to judge *a posteriori* that what they did (saving the child) was the right thing to do. Of course, Reflectivism can try to account for this difference by another feature of these cases, but if it does so, then we have no longer any reason to endorse it, because our reason to endorse it was to account for this difference.

Another objection to this weaker version of Reflectivism is that it is hard to see why the mere fact of having this ability should make a difference if the fact of using it doesn't. We can compare this objection to Frankfurt's famous argument against the Principle of Alternate Possibilities (Frankfurt 1969). The Principle of Alternate Possibilities is a principle that states that an agent cannot be morally responsible for what he did if he couldn't have done otherwise. Against this principle, Frankfurt has us imagine cases in which an agent has no alternative (and thus cannot do otherwise), but in which this fact doesn't motivate the subject (e.g. John decides by himself to kill his wife while, unbeknownst to him, Black has implanted in his brain a chip that would have compelled John to kill his wife if he hadn't chosen to do it by himself). In this case, according to Frankfurt, the agent's action is exactly the same as in a parallel case in which he has an alternative (the case in which John decides by himself to kill his wife but Black and the chip do not exist). So, Frankfurt asks, why should we make any difference between these two actions? If John is responsible when he has an alternative, he should also be responsible when he has none, provided that the lack of alternative does not cause his action in any way. On this model, we can take our case Pond* and imagine a parallel case Pond*** in which the agent has an innate psychological defect that prevents him to understand what is good and what is bad. Since in Pond* such knowledge doesn't play any role in the production of the action, why should we treat the agent's action in Pond*** differently and deny praise to the agent?

To follow this idea further, let's imagine an individual who, due to very specific and focal brain damages he suffered at birth, is deprived of moral concepts: he doesn't understand what words like 'right', 'wrong', 'fair' or 'unfair' mean. This individual has no other cognitive issues whatsoever: he's fully capable of understanding what he's doing and the consequences of this action. Let's now say that this individual is also (and was before suffering from brain damages) a dangerous criminal: he enjoys kidnapping people and takes pleasure torturing them. Remember that he perfectly understands what he's doing: he knows what pain is, and he knows that his victims suffer from the treatment he inflicts on them. He also knows that they don't want to suffer and that torture is a highly traumatizing experience. He just doesn't know whether it is 'right' or 'wrong' to torture people. Is it obvious that this individual is not responsible for what he did? And that he doesn't deserve blame? I don't think so. We have the feeling that, even if he can't form moral judgment about his actions, this individual is a terrible person. He does something wrong (torturing people) for clearly vicious motives (because he enjoys watching people suffering).

Let's now imagine the reverse individual (what I call a 'reverse psychopath'): a man who does not have access to moral concepts but deeply cares about other persons (because he's deeply empathetic), so that he spends most of his time helping people. Should we be reluctant to praise him? Should we consider that the people he helps don't have to be grateful? That seems a wrong conclusion.[8]

Of course, these arguments are far from being inescapable: they are mostly appeals to our intuitions. But the fact that our intuitions go against Reflectivism is *prima facie* a reason to reject it, unless its adherents provide a convincing argument. And if we have no reason to endorse Reflectivism, then I can't think of any other reason to accept the thesis according to which moral ignorance is incompatible with moral agency.

[8] One could consider that these example are not sufficient because it is one thing to evaluate persons (as nice or vicious) and another to attribute them responsibility for their action. This is true. Nevertheless, I think we also have the intuition in those cases that these persons are responsible for their actions. Let's take the reverse psychopath and imagine that he helps you (by saving your life). It seems quite natural to feel grateful and consider that you owe him something. But this feeling cannot be accounted by a mere judgment of 'niceness' (you are not indebted to all people you find nice): rather, you would feel indebted in such a case because the reverse psychopath deserves credit for what he did – that is to say: because he was responsible for what he did.

7.3 Reasons and the Freedom Conditions for Moral Responsibility

In the previous section, I have examined what in the epistemic conditions for moral responsibility could make moral ignorance incompatible with moral agency. Nevertheless, considerations about the freedom conditions for moral responsibility could give support to the thesis that moral agency requires the capacity to form moral judgment.

There are many incompatible ways of framing the freedom conditions for moral responsibility – I will focus here on its definition as 'reason-responsiveness'. Fischer and Ravizza (1998) have given 'reason-responsiveness' as a criterion for freedom: basically, the idea is that we can identify whether a given action is free by considering whether the agent would have acted differently had his reasons been different. If I would have had a good reason not to get out of bed this morning, for instance, I would not have got out of bed. Thus we can say that my getting out of bed this morning is something I did freely. To be free is to act on the basis of reasons.

The question is now what it means to act on the basis of reasons. In a strong reading, acting on the basis of reason R amounts to act because we consider R to be a good reason. But assessing whether a reason is a good reason or not seems to immediately drive us in the realm of practical reason and of morality, and thus to make moral evaluation of reasons a requirement for acting for reasons and moral responsibility. Such a strong reading can be found in Korsgaard's conception of morality:

> Our purposes may be suggested to us by our desires and emotions, but they are not determined for us by our affective states, for if we had judged it wrong to pursue them, we could have laid them aside. Since we choose not only the means to our ends but also the ends themselves, this is intentionality at a deeper level. For we exert a deeper level of control over own movements when we choose our ends as well as the means to them than that exhibited by an animal that pursues ends that are given to it by its affective states, even if it pursues them consciously and intelligently. Another way to put the point is to say that we do not merely have intentions, good or bad. We assess and adopt them. We have the capacity for normative self-government, or, as Kant called it, 'autonomy'. It is at this level that morality emerges. The morality of your action is not a function of the content of your intentions. It is a function of the exercise of normative self-government. (De Waal 1997, p. 112)

According to Korsgaard, being moral implies to be reason-responsive in a particular way, i.e. to act according to our reasons and the judgments we make about them. Under such a strong reading of reason-responsiveness, the ability to form moral evaluations is a key component of moral agency.

However, I would like that a weaker account of reason-responsiveness is sufficient to capture our intuitions about moral agency. Let's first return to the problem mentioned in the second section: why do we praise the man who saves the child 'because the child would have died' but not the man who saves the child 'to become famous or get a reward'? The Reflectivist answer was that the man who saves the child for a reward does not act on the basis of a moral judgment. Nevertheless, we rejected this condition as too high: the man who impulsively dives to save the child does not act upon the basis of a moral judgment either – still, he deserves praise for his actions.

Reflectivism might be on the right track when it says that a moral agent deserves praise when he acts *for good reasons.* But a good reason doesn't need to be the judgment that an action is morally good. The man who dives 'because the child would have died' acts precisely for a *good* reason, though 'the child would have died' is not a moral judgment.

So, what makes a reason a morally good reason? I think that what makes our man praiseworthy is that he's motivated by the fact he somehow *cares* about the child: he dives because he wants to prevent the child to be harmed. Thus, one acts for the good reasons when one actually cares about the person one is trying to help. In the same way, Huckleberry Finn deserves praise for helping Jim, since he helps Jim because he *cares* about him, whatever he might think of the moral appropriateness of such an attitude. Similarly, our 'reversed psychopath' also deserves praise for helping people because, though he doesn't realize that what he's doing is right, he *cares* about people.

The question is now: what does it take to be able to *care* about someone? One condition is to be able to understand that this person has *interests* – that some things have (positive or negative) values for this person. For example, if you understand that a person can suffer and doesn't want to suffer, then you understand that it is in his interests not to suffer. Nevertheless, understanding that persons have interests is not enough to actually care about them. For example, psychopaths totally understand that people have interests, but do not care about them. To care about someone is to give an intrinsic (i.e. non-instrumental) positive value to the fact that this person's interests are preserved and augmented, and to give an intrinsic negative value to the fact that this person's interests are damaged.

So, to deserve praise for a right action involving a moral patient, one just has to perform this action because one cared about this moral patient's interest. This is why, the case of a man that would help the child because he has been taught it's the right thing to do but not because he cares about the child is so disturbing.

These conditions also allow us to determine when an agent deserves praise for punishing a criminal that harmed a moral patient. One does not immediately deserve praise for 'having done justice'. For example, if I punish a criminal not because I care about what he did to the victim, but because I just like punishing people, I don't deserve praise – here again, I must care about the moral patient, the victim, to deserve praise for what I did.

Going further, it also allows us to understand what it takes to deserve blame for having done something wrong. To deserve blame, I must harm a person and, though I understand this person has interests, not care about them. For example, if I stomp on someone's toes, just because I'm in a hurry and don't care about harming this person, it's enough for me to deserve blame. Otherwise said, I deserve blame when my action is expressive of the fact I don't care about others.[9]

[9] For a development of the psychological theory underlying this account of moral responsibility, see Cova et al. (submitted).

My proposal is then that being a moral agent only requires understanding that others have interests and the capacity to be motivated by this understanding (that is: to act in accordance with how much I care about others). If I don't care about these interests, I have bad motives. If I care about these interests, I have good motives – but the main point is to understand that some entities have *interests*, and that makes them moral patients. All it takes to be a moral agent is thus to be able to act according to what we attach importance to and a bit of theory of mind.[10]

With this theory, we can construe a less demanding reason-responsiveness for the freedom condition: to be reason-responsive is to be able to be sensitive to the well-being of others, and thus able to act (or refrain from acting) on the reason that someone's well-being is at stake.[11] But this doesn't mean that agents are required to act because they consider that improving the well-being of others is a good thing. If it did, that would run against our intuitions about the Pond cases presented in Sect. 7.2.

Thus, we think there is a plausible version of the epistemic and freedom conditions for moral responsibility according to which moral responsibility might not require the ability to form moral judgment. In the last section of the paper I return to the implication of such a possibility.

7.4 Conclusion: Moral Animals and Twice-Moral Human Beings

I began this paper by presenting a conflict between the continuists, who think we share with animals the psychological requirements for morality, and discontinuists, who think that we are endowed with unique moral capacities. In this paper, I've argued that both perspectives could be reconciled by distinguishing two components of our moral life: moral agency (we are morally responsible for our actions) and moral judgments (we are able to evaluate our behaviour and that of others).

To side with the discontinuist, it is hard to deny that moral life is much richer in human beings than in any other moral animals: we are able to ask tough moral questions and reason about difficult moral situations (such as moral dilemmas). Nevertheless, the continuist can also cite cases in which it is hard to deny that other species can have a real moral life. Let's consider the following story:

[10] Some (e.g. Knobe 2006) have argued that our theory-of-mind is already suffused with moral considerations and evaluations, which goes directly against our argument that presupposes that theory of mind is independent from the faculty of moral judgment. However, there are reasons to doubt this claim. See Cova et al. (2010) for a rebuttal.

[11] Does it mean that empathy is necessary to be a moral agent? No, empathy is what makes us care about others' interests and be *good* moral agents. But, a *bad* moral agent, one who doesn't care about others, is still a moral agent, so empathy is not necessary to be a moral agent. Note also that empathy might not even be necessary to be a good moral agent: there might be other emotional or cognitive ways to care about others' interests.

> The Arnhem chimpanzees spend the winters indoors. Each morning, after cleaning the hall and before releasing the colony, the keeper hoses out all the rubber tires in the enclosure and hangs them one by one on a horizontal log extending from the climbing frame. One day Krom was interested in a tire in which water had been retained. Unfortunately, this particular tire was at the end of the row, with six or more heavy tires hanging in front of it. Krom pulled and pulled at the one she wanted but could not move it off the log. She pushed the tire backward, but there it hit the climbing frame and could not be removed either. Krom worked in vain on this problem for over ten minutes, ignored by everyone except Otto Adang, my successor in Arnhem, and Jakie, a seven-year- old male chimpanzee to whom Krom used to be the 'aunt' (a caretaker other than the mother) when he was younger.
>
> Immediately after Krom gave up and walked away from the scene, Jakie approached. Without hesitation he pushed the tires off the log one by one, as any sensible chimpanzee would, beginning with the front one, followed by the second in the row, and so on. When he reached the last tire, he carefully removed it so that no water was lost and carried the tire straight to his aunt, where he placed it upright in front of her. Krom accepted his present without any special acknowledgment and was already scooping water with her hand when Jakie left. (De Waal 1997, p. 83)

In this case, it seems that Jakie understood what Krom wanted, and helped her get it, and it is hard not to perceive this story in moral terms. According to the account developed in Sect. 7.3, it is the fact that Jakie understood that Krom had interests that makes him look like a moral agent.[12]

By stressing that moral agency can exist without moral judgment, I hope to have contributed to diminishing the gap between continuists and discontinuists. Indeed, it is interesting to note that continuists often emphasize action and emotions, while discontinuists stresses the uniqueness of human moral reflexion. These two insights can be reconciled in a more complex and fine-grained view of moral life, a view that opens the interesting possibility that moral agency can have evolved independently from moral judgment.

[12] This claim immediately raises several questions. (i) First, if Jakie is really a moral agent, does that immediately make him a moral patient (i.e. someone it is wrong to harm). I was tempted to say 'yes' until an anonymous reviewer gave me the following argument I found quite convincing: 'Are all moral agents moral patients? Probably, but: one could imagine someone who has racist beliefs and thinks, for example, that black people are inferior to white people, and thus think that they do not bear rights in the same way and to the same extent as white people, yet think that they are equally morally responsible for their actions, and are bound by the same norms. In this case, being a moral agent would not be sufficient for being a moral patient. This option does not seem conceptually incoherent.' (ii) Second, if Jakie can be responsible for his action, does it necessarily entail that he can be punished? I am not sure either, for punishment seems to require something more than moral responsibility. For example, it seems to me that we want the people we punish to understand why they are punished and that it is essential for punishment that the punished one understands it as such. This intuition is supported by experimental studies showing that people consider revenge satisfactory only if the offender understands (and acknowledges) that revenge was taken against him because and in virtue of a prior unfair behaviour (Gollwitzer and Denzler 2009). If it is the case, then one has to be both a moral agent and a moral judge to be an appropriate target of punishment. (Even if you take a consequentialist stance on punishment, then you must admit that an agent that cannot understand why he has been punished and on whom deterrence won't work is not a suitable target of punishment.)

Acknowledgments This research was supported in part by a Grant from the French Agence Nationale de la Recherche (ANR) (ANR Blanche: SoCoDev). I thank François Jacquet and an anonymous reviewer for useful comments on previous versions of this paper.

References

Aristotle. 2011. *Nicomachean ethics*. Chicago: University of Chicago Press.

Arpaly, N., and T. Schroeder. 1999. Praise, blame and the whole self. *Philosophical Studies* 93(2): 161–188.

Baumard, N. 2010. *Comment nous sommes devenus moraux: Une histoire naturelle du bien et du mal*. Paris: Odile Jacob.

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

Blair, R.J. 1997. Moral reasoning and the child with psychopathic tendencies. *Personality and Individual Differences* 26: 731–739.

Cova, F., E. Dupoux, and P. Jacob. 2010. Moral evaluation shapes linguistic reports of others' psychological states, not theory-of-mind judgments. *The Behavioral and Brain Sciences* 33: 334–335.

Cova, F., P. Jacob, and E. Dupoux. Submitted. The more you deserve blame, the more you deserve blame: The "valence matching" heuristic.

Damasio, A. 1995. *Descartes' error*. New York: Putnam.

De Waal, F. 1997. *Good natured: The origins of right and wrong in humans and other animals*. Cambridge, MA: Harvard University Press.

Fischer, J.M., and M. Ravizza. 1998. *Responsibility and control: An essay on moral responsibility*. Cambridge, MA: Cambridge University Press.

Frankfurt, H.G. 1969. Alternate possibilities and moral responsibility. *Journal of Philosophy* 66(23): 829–839.

Gollwitzer, M., and M. Denzler. 2009. What makes revenge sweet: Seeing the offender suffer or delivering a message? *Journal of Experimental Social Psychology* 45: 840–844.

Kant, I. 2011. *Groundwork of the metaphysics of moral*. Cambridge, MA: Cambridge University Press.

Knobe, J. 2006. The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies* 130: 203–231.

Levy, N. 2008. The responsibility of psychopaths revisited. *Philosophy, Psychiatry and Psychology* 14(2): 129–138.

Marit, R.S., and P.A. Wilkosz. 2005. Disorders of diminished motivation. *The Journal of Head Trauma Rehabilitation* 20(4): 377–388.

Mikhail, J. 2011. *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. Cambridge, MA: Cambridge University Press.

Pizarro, D., E. Uhlmann, and P. Salovey. 2003. Asymmetry in judgments of moral blame and praise. *Psychological Science* 14(3): 267–271.

Singer, P. 1972. Famine, affluence and morality. *Philosophy and Public Affairs* 1(3): 229–243.

Turiel, E. 2002. *The culture of morality: Social development, context and conflict*. Cambridge, MA: Cambridge University Press.

Chapter 8
Humean Moral Motivation

Andrés Luco

8.1 Introduction: The Problem of Moral Motivation

One of the more redeeming features of human nature is that people are capable of doing the right thing – of acting with exceptional benevolence and justice. Philosophers have long been trying to understand the motivations that underlie moral conduct, and for good reason. Discoveries about the facts of moral motivation could overturn some of the most fundamental postulates of moral philosophy. Suppose *psychological egoism* turns out to be true. Psychological egoists hold that all human behaviour is motivated ultimately by self-interest. But as Russ Shafer-Landau has noted, if psychological egoism is true, then altruism is impossible (Shafer-Landau 2010, pp. 88–89). And yet, many of the moral duties we have seem to require altruism. Surely, however, we cannot have moral duties to do the impossible. Thus, if psychological egoism is true, it may entail the unpalatable consequence that there are no moral duties which require us to act altruistically. Another rationale for philosophical interest in moral motivation has to do with the practical aims of ethics. It is all very fine to theorize about the right and the good, but we would of course like to see people actually *do* right and *pursue* the good. Accordingly, ethicists from Plato to Rawls have put considerable effort into understanding the levers of human psychology which can be pulled to generate ethical behaviour.

In this essay, I shall defend the *Humean theory of motivation* (hereafter 'Humeanism') as the best account of moral motivation. Humeanism is often called the 'belief–desire model' of action because it explains all intentional behaviour – including moral behaviour – by citing a combination of a desire for some end and a belief that an action is a means to achieve that desired end. I offer three arguments in favour of Humeanism. First, I suggest that Humeanism meets one criterion

A. Luco (✉)
Philosophy Group, Nanyang Technological University, Singapore, Singapore
e-mail: acluco@gmail.com

B. Musschenga and A. van Harskamp (eds.), *What Makes Us Moral? On the capacities and conditions for being moral*, Library of Ethics and Applied Philosophy 31,
DOI 10.1007/978-94-007-6343-2_8, © Springer Science+Business Media Dordrecht 2013

of theory choice better than its theoretical competitors. This is the *continuity constraint*. Humeanism is consistent with a *continuous natural history of moral motivation*, i.e. an evolutionary account which explains how the moral motivations of modern humans evolved from simpler precursors resembling the psychological traits of nonhuman animals (especially nonhuman primates). The same cannot be said for the rivals of Humeanism. Then, in order to ward off the objection that nonhuman primates are too different from humans to provide any insight into moral motivation, I offer a second argument for Humeanism. It draws on another criterion of theory choice called *Morgan's Canon* – a standard of theoretical parsimony according to which a psychological mechanism should be attributed to an organism only if it's the *sole* mechanism which can cause some behaviour. I contend that Humeanism meets this criterion more successfully than rival accounts of moral motivation. I will so argue on the grounds that the Humean belief–desire model can successfully explain both moral action and a variety of non-moral actions. Finally, I shall argue that recent findings in neuroeconomics furnish empirical evidence in favour of Humeanism, and against anti-Humeanism.

8.2 Theories of Moral Motivation: Humeanism Versus Anti-Humeanism

In this section I describe the dominant philosophical theories of moral motivation. First, I canvass a few preliminary observations and assumptions about moral motivation in general. *Moral action* is action *judged by the agent* to be either morally praiseworthy or required. By this definition, an agent's own moral judgments determine whether his or his actions count as *moral* actions. The definition does not presuppose that an agent must *correctly* judge his or her actions to be morally praiseworthy or required in order for those actions to be moral actions. *Moral motivation* refers to the motivations that explain moral action. As Connie Rosati notes, the 'basic phenomenon' of moral motivation is that moral motivations reliably change in accordance with changes in moral judgment. For instance: if someone judges that φ is morally right, then he or she will ordinarily be motivated to φ; but if this person becomes convinced that φ is morally wrong and that ψ is morally right, he or she will ordinarily become motivated to ψ rather than φ (Rosati 2006, §1). It is also thought that moral motivation varies with moral judgment due to practical reasoning. *Practical reasoning* refers to the process of thinking about how one should act. All the theories of moral motivation canvassed below assume that moral conduct is the result of a process of practical reasoning initiated by moral judgment.

It is also commonly assumed that moral action is a species of intentional action. *Intentional action* is action motivated by the intentional states of the agent. Intentional states are psychological states that are *about*, *of*, or *for* something else. They include beliefs, desires, and emotions – e.g. a belief *about* the existence of God, a desire *for* pistachio ice cream, a fear *of* snakes, and so on. The 'content' of an intentional state is whatever the state is about, of, or for. I have a belief about the existence of God; so, the content of that belief includes the concept of God.

It will also become clear that the rival theories of moral motivation all distinguish between cognitive and affective states of mind. *Cognition* is associated with thinking, believing, reasoning, learning, decision making, and memory. *Affect*, or emotion, is linked to feeling and wanting. However, there is no consensus definition of either cognition or affect. Despite this unfortunate fact, I will rely on a promising account of the cognition–affect distinction articulated by Jesse Prinz (cf. Prinz 2004, pp. 41–51). According to Prinz,

> cognitive states and processes are those that exploit representations that are under the control of an organism rather than under the control of the environment. A representation is under organismic control if the organism has activated it or maintains it in working memory. A cognitive state is one that contains such a representation, and a cognitive process is one that activates, maintains, or manipulates such a representation. (Prinz 2004, pp. 45–46)

Prinz uses the term 'organismic control' as a synonym for 'executive control', He identifies the 'executive' brain structures as those centred in the prefrontal cortex (Prinz 2004, p. 47). The sorts of representations that Prinz takes to be under organismic control are *concepts*, as opposed to *percepts*. When you see a shape, you have a percept of that shape. But you can also store a copy of that percept in your memory. The copy of that percept recalled from memory is a concept of the shape. Thus, a cognitive state, in Prinz's view, is a mental state which contains concepts (Prinz 2004, p. 46).

Prinz's definition of a cognitive state accounts for many intuitions about what mental activities count as 'cognitive' (Prinz 2004, p. 48). For instance, cognition is associated with the conscious processing of information ('access consciousness'). Such processing includes acts of thinking, like doing long division, and deliberation about actions. It therefore involves organismic control. Also, cognition is thought to be effortful. Choosing a goal, for instance, seems to require some kind of mental work. The frontal cortex contributes to making strategic choices among competing goals. By contrast, affect occurs unbidden. You don't have to make an effort to feel sad or angry; these emotions are induced by events in the surrounding environment.

Bearing in mind the differences between cognitive and affective states, we shall see that competing accounts of moral motivation can be distinguished by the way they respond to two questions:

1. *Can a cognitive (i.e. belief-like) psychological state produce a motivation at time t, without being accompanied by an affective (i.e. emotional or desire-like) psychological state which exists prior to t?*
2. *Are moral judgments cognitive states or affective states?*

The first theory of moral motivation on the table is the Humean theory of motivation – also designated here as 'Humeanism'. The theory bears this name because its adherents claim an intellectual debt to David Hume. Humeanism is also called the *belief–desire model* of action, because it holds that intentional action is explained by a combination of a desire for some end, together with a belief that some action is a means to satisfying the desire. A belief that some action is a means to achieving some desired end is called a *means–ends belief.*

The Humean answer to the first question above is ‘no’ – it is not psychologically possible for a cognitive state to produce a motivation in an agent, unless it is accompanied by some existing affective state. Humeans treat beliefs as paradigmatic cognitive states, and desires as paradigmatic affective states. Additionally, Humeans assume that cognitive and affective states are mutually exclusive. Thus, Humeans hold that no belief is a desire. The basic Humean position with respect to the first question, then, is that no *belief* can motivate intentional action without being accompanied by an existing desire.

Regarding the second question, Humeans are split into two camps: cognitivists and non-cognitivists. *Cognitivism* is the view that a moral judgment expresses a cognitive state – in particular, a belief. The defining feature of beliefs is that they are *truth-apt*, meaning they are capable of being either true or false. *Non-cognitivism*, by contrast, is the view that a moral judgment expresses a non-cognitive state, such as an emotion or desire. Non-cognitive states are supposedly not truth-apt. Desires, for instance, cannot be true or false – they can only be satisfied or unsatisfied. Non-cognitivist Humeans include such authors as Simon Blackburn (1984) and Allan Gibbard (1990), whereas proponents of cognitivist Humeanism include John Mackie (1977) and Nick Zangwill (2003).

Humeans maintain that desires play a role quite independently of beliefs in generating moral motivation (Zangwill 2008a, b). At least three kinds of desires can contribute to moral motivation: (1) self-regarding desires, (2) other-regarding desires, and (3) moral desires. First, agent *S* has a *self-regarding desire* to achieve some end *E*, if *E* is exclusively a state of *S*. Second, agent *S* has an *other-regarding desire* to achieve *E* if *E* is exclusively a state of some person or group *other than S*. For instance, suppose that I donate blood purely out of a desire to win the respect of my peers through conspicuous acts of benevolence. If you were to ask me, I would honestly admit that if donating blood did not do wonders for my reputation, I wouldn’t have any desire to do it. In this case, I have a self-regarding desire to donate blood, where the desired end is an improved reputation. An improved reputation is a state in which I find myself; it is not a state of anyone else. On the other hand, suppose instead that I donate blood because of a desire to promote the health of people I’ve never met. Suppose my desired end is to change the states that other people are in, and not my own state, by improving the health of those in need of blood transfusions. To the extent that my end is exclusively to improve the health of others, I have an other-regarding desire to donate blood.

Finally, and perhaps most importantly, moral motivation can result from moral desires. Agent *S* has a *moral desire* to achieve end *E* if *S* judges that *E* bears a set of moral properties, and *S* desires the moral properties of *E*. For instance, I will have a moral desire to donate blood if I judge that it is morally right to donate blood, and I want to do what I judge to be right. The content of a moral desire can be understood as a concept or representation of moral properties such as *rightness*, *goodness*, *praiseworthiness*, etc. Conversely, the content of a *non-moral* desire would not include any representation of moral properties. In this essay, no position is taken on the existence of mind-independent moral properties. Even a moral error theorist could assent to the definition of moral desires just introduced; but would simply add the caveat that the contents of moral desires are *moral concepts* that only *purport* to represent moral properties. However, the error theorist will add that moral concepts *misrepresent* reality, because moral properties do not exist. Likewise, a non-cognitivist Humean like Blackburn could accept the foregoing definition of moral desires. But he would specify that moral properties are projections of sentiment, and that no mind-independent moral properties exist.

The second group of theories about moral motivation is classified under the heading of *anti-Humeanism*. With regard to the first question above, the anti-Humean answer is 'yes'. Anti-Humeans claim that a cognitive state, or a state with both cognitive and affective features, can be sufficient to generate motivation without being paired up with any independently existing affective state. Regarding the second question, anti-Humeans treat moral judgments as either cognitive psychological states, or as hybrid psychological states with both cognitive and affective features.

Margaret Olivia Little (1997) defends two prominent versions of anti-Humeanism. Each version advances a different theory of the cognitive (or semi-cognitive) psychological states that motivate moral action. According to the first version of anti-Humeanism, moral judgments are partly cognitive, and partly non-cognitive; they are hybrid psychological states that have both cognitive (belief-like) and affective (desire-like) features. Accordingly, these hybrid states have been dubbed '*besires*' (Little 1997, p. 254).[1] Besires are claimed to be both belief-like and desire-like because they supposedly have both of two *directions of fit* that distinguish beliefs from desires. Beliefs have a 'mind-to-world' direction of fit, because they are the sort of intentional state that goes out of existence when the subject perceives evidence that the world is not how he believes it to be. If, for instance, I believe that it's 9 am on Monday, but my friend informs me that it's actually 10 am, my belief that it's 9 am will be extinguished because it is contradicted by my friend's (presumably reliable) testimony. On the other hand, desires have a 'world-to-mind' direction of fit, since the content of a desire is a representation of how the world *is to be*. Whenever there is a discrepancy between how the world is and how the subject wants the world to be, a desire has the effect of motivating the subject to *change* the world until his desire is satisfied (cf. Smith 1987, p. 56).

Besires are said to have both directions of fit. David McNaughton, another anti-Humean, suggests that moral judgments are besires. A moral judgment has a mind-to-world direction of fit, in the sense that an agent can perceive the moral properties of actions, persons, institutions, etc. For instance, one can perceive that an action is morally required (McNaughton 1988, p. 109). But at the same time, a moral judgment has a world-to-mind direction of fit, i.e. a moral judgment can motivate an agent to change the world and carry out what he judges to be morally required.

[1] The term 'besire' was coined by J.E.J. Altham (1986).

Another group of anti-Humeans postulate a different kind of cognitive state to explain moral motivation. Little calls them *desire-entailing beliefs* (Little 1997, pp. 254, 259–260). According to the anti-Humeans in question, moral judgments are beliefs, and at least some moral beliefs are desire-entailing beliefs (Little 1997, p. 261). Moral beliefs 'entail' desires in the sense that 'possession of certain desires is a necessary condition for possession of certain moral beliefs' (Little 1997, p. 260). Here it's crucial to bear in mind that the explanatory direction of the entailment is *from* moral beliefs *to* desires. Possessing certain moral beliefs *necessarily explains* the formation of certain desires to act in accordance with those beliefs. Thus, if one lacks the desires which would be entailed by a desire-entailing moral belief, either one does not genuinely have the relevant moral belief in the first place, or one is afflicted by some form of irrationality which interferes with the normal process of entailment (Little 1997, pp. 260–261; Zangwill 2008b, pp. 96–97, 111). For instance: suppose that I believe I have a moral duty to minimize avoidable suffering, and I notice that a shortage of blood supplies causes avoidable suffering. If my moral belief is a desire-entailing belief, I must form the desire to donate blood *because* of that moral belief.

Unlike besires, desire-entailing beliefs are not hybrid states with dual directions of fit. They are ordinary beliefs with mind-to-world direction of fit. Also, the desires entailed by desire-entailing beliefs are ordinary desires with world-to-mind direction of fit. However, the idea of a desire-entailing belief is contrary to Humeanism. It suggests that a belief unaccompanied by an existing desire state is by itself sufficient to produce a new desire. Humeans think this is impossible.

Now I turn to three arguments in favour of the Humean account of moral motivation. My remarks are not designed to favour either cognitivist or non-cognitivist Humeanism. Instead, the arguments are intended to provide a general case for Humeanism and against anti-Humeanism.

8.3 For Humeanism: The Argument from Continuity

Humeanism and anti-Humeanism are rival theories in the psychology of motivation. As such, they are subject to the same criteria of adequacy as any psychological hypothesis. In this section I suggest that Humeanism meets one theoretical criterion better than anti-Humeanism. This criterion derives from the theory of evolution, and I will call it the *continuity constraint*:

Continuity Constraint

> Given that human beings and other animals evolved from common ancestors, a theory of human psychology should be consistent with the best available explanation of how human psychological traits evolved from simpler precursors resembling the psychological traits of nonhuman animals (especially nonhuman primates).

The continuity constraint is based on one of the most profound insights of evolutionary theory. It presupposes that organic structures, including psychological structures, are products of a slow, incremental process of 'descent with modification', in which newer, more complex structures arise as piecemeal modifications of older, often simpler ones.[2] The continuity constraint advises us to prefer psychological theories which are consistent with explanations of how human psychological traits evolved from simpler antecedents resembling the psychological traits of nonhuman animals. The psychologies of nonhumans – especially nonhuman *primates* – furnish important sources of evidence for the motivations that were probably characteristic of (now extinct) ancestor species we modern humans have in common with the other animals.

Many significant philosophers and scientists have defended the continuity constraint. Darwin himself argued for the psychological continuity of humans and nonhumans (cf. Sober 1998, p. 228). In the second edition of *The Descent of Man* (1874), Darwin wrote:

> The difference in mind between man and the higher animals, great as it is, certainly is one of degree and not of kind ... the senses and intuitions, the various emotions and faculties, such as love, memory, attention, curiosity, imitation, reason, &c., of which man boasts, may be found in an incipient, or even sometimes in a well-developed condition, in the lower animals. (Thompson 1995, p. 71)

In the same treatise, Darwin suggests that the precursors of human moral psychology could be found in other social animals:

> Any animal whatever, endowed with well-marked social instincts, the parental and filial affections being here included, would inevitably acquire a moral sense or conscience, as soon as its intellectual powers had become as well, or nearly as well developed, as in man. (Thompson 1995, pp. 41–42)

Daniel Dennett also defends a commitment to the continuity constraint. He urges that the continuity constraint applies even when an organism exhibits unique and unprecedented traits, as is the case with modern humans. Even when the emergence of a new organism or biological structure at some moment in evolutionary history appears radically different from what existed previously, Dennett maintains that theorists should search for

> a subprocess or special feature of a design process that can be demonstrated to permit the local speeding up of the basic, slow process of natural selection, *and* that can be demonstrated to be itself the predictable (or retrospectively explicable) product of the basic process. (Dennett 1995, p. 76)

In sum, the continuity constraint requires that the origins of an observed psychological trait should be explicable as a product of descent with modification from

[2] Evolutionary change does not *only* proceed from the simple to the complex. As Elliot Sober notes, 'the history of life is peppered with cases of evolutionary simplification. For example, the evolution of parasites typically involves a transition from complex to simple, as the parasite loses features of its free-living ancestor' (Sober 1998, p. 225). When it comes to understanding the evolutionary origins of human psychological features, though, we obviously seek an explanation for how something as complex as the human mind arose from simpler precursors.

pre-existing traits. Since the theory of evolution implies that humans and other animals share common ancestors, the older traits on which natural selection operated to produce the psychological capacities of modern humans can probably be found in nonhuman creatures – most especially, the extant nonhuman primates. Even when a trait appears to be radically new and different, it should still be explicable as the product of evolutionary processes operating on pre-existing traits.

I argue that Humeanism meets the continuity constraint, while anti-Humeanism does not. The motivations which Humeans posit to account for moral action are the same as, or very similar to, the motivations that researchers have invoked to explain what I call *proto-moral behaviours* in nonhuman primates. Human beings are capable of *altruism*, i.e. of helping and caring for others. We extend aid to our family members, friends, neighbours, and even strangers, sometimes at great sacrifice to ourselves. When we ask ourselves why we help and care for others the way we do, we often resort to moral language – we speak of the *duties* of a good mother or son, the *virtues* of a loyal friend, and so on. Thus, some altruistic behaviours, at least when carried out by human beings, are instances of *moral* action. But some nonhuman animals – particularly the other primates – engage in similar forms of helping and caring. Such animal behaviour can be called *proto-moral*, since it may furnish insight into the evolutionary precursors of an important class of morally significant actions exhibited by humans. The continuity constraint supports the expectation that the origins of the psychological capacities associated with human moral motivation should be traceable back to ancestors that *homo sapiens* shares with other primates. As we shall see, Humeanism meets the continuity constraint because it explains how proto-moral motivations could have been shaped by natural selection into the fully fledged moral motivations of modern humans. The same cannot be said for anti-Humeanism.

Primatologist Frans de Waal has observed compelling parallels between helping/caring behaviours in human beings and nonhuman primates. In fact, De Waal himself voices a commitment to the continuity constraint. He argues that some primates – particularly the great apes – are motivated by the same psychological mechanisms which underlie human propensities for empathy and altruism (De Waal 2006, p. 28, 2008). In particular, De Waal asserts the following:

> In human behavior, there exists a tight relation between empathy and sympathy, and their expression in psychological altruism . . . It is reasonable to assume that the altruistic and caring responses of other animals, especially mammals, rest on similar mechanisms. (De Waal 2006, p. 28)

De Waal characterizes empathy and altruism as necessary conditions for moral conduct. Furthermore, he holds that empathy and altruism are fundamentally affective motivations based on an ‘emotional interest in others’ (De Waal 2006, pp. 20–21). In line with the continuity constraint, De Waal contends that the same affective states which generate moral motivation in humans also motivate proto-moral behaviour in nonhumans.

De Waal uses the term ‘*intentional altruism*’ to denote the suite of motivations which explain helping behaviour (De Waal 2008, p. 281). Intentional altruism consists of *intentional states* that motivate helping. De Waal argues that *empathy* is the proximate mechanism through which intentional altruism operates (cf. De Waal 2008, p. 282). Empathy is defined by De Waal as ‘the capacity to (*a*) be affected by and share the emotional state of another, (*b*) assess the reasons for the other’s state, and (*c*) identify with the other, adopting his or her perspective’ (De Waal 2008, p. 281).

De Waal proposes what he calls a ‘Russian Doll Model’ of empathy. The model consists of three levels of empathy: (1) *emotional contagion*, (2) *sympathetic concern*, and (3) *empathetic perspective-taking* (De Waal 2008, pp. 287–288). These three levels of empathy are distinguished according to the complexity of the psychological mechanisms needed for an organism to experience each level. Emotional contagion – the lowest level of empathy – can be attributed to rats, birds, and monkeys. These animals are much less psychologically complex than large-brained mammals like cetaceans and the great apes. Emotional contagion consists of the matching of emotional states between one individual (the ‘subject’) and another (the ‘object’) (De Waal 2008, p. 282). For example, mice exhibit intensified pain responses when they see other mice exhibiting similar responses. Also, rhesus macaque monkeys tend to terminate projected pictures of conspecifics in a fearful pose (De Waal 2008, pp. 283, 288). In each of these examples, the emotional state of one individual ‘spreads’ to another.

The second level of empathy is *sympathetic concern*, which involves emotional concern about a distressed or needy other (De Waal 2008, p. 283). Additionally, the sympathizing subject’s concern motivates attempts to ameliorate the other’s distress. A manifestation of sympathetic concern is consolation behaviour, which de Waal defines as ‘reassurance provided by an uninvolved bystander to one of the combatants in a preceding aggressive incident’ (De Waal 2006, p. 33). For example, De Waal observed a juvenile chimpanzee approach and put its arms around a screaming adult male who had just been defeated in a fight. Quantitative studies have found that in chimpanzees, third parties direct consolation behaviour more at recipients of aggression than at aggressors, and more at recipients of *intense* aggression than mild aggression (De Waal 2006, p. 35).

De Waal takes sympathetic concern to involve more complex psychological capacities than emotional contagion, for it involves both an emotional concern which is sensitive to the emotional state of another being, and a motivational disposition to try and reduce the distress of the other. Animals capable of emotional contagion, but not sympathetic concern, do not engage in consolation behaviour. For instance, macaque monkeys do not even comfort their own offspring after a fight. By contrast, the reassurance of distressed others is typical in the great apes, which are closer phylogenetic relatives to modern humans than monkeys (De Waal 2008, p. 285).

De Waal’s third level of empathy is *empathetic perspective-taking* (henceforth EPT). According to De Waal, EPT consists in ‘the capacity to take another’s perspective, e.g. understanding another’s specific situation and needs separate from one’s own, combined with vicarious emotional arousal’ (De Waal 2008, p. 285). Like the first two levels of empathy, EPT involves emotional engagement with the other, along with a disposition to reduce the other’s distress. However, EPT is more complex because it requires the ability to attribute mental states to the other, and to recognize the causes of those states. The best evidence for EPT is the phenomenon of *targeted helping*, which is helping behaviour fine-tuned to the needs and situation of the other (De Waal 2008, p. 285). De Waal cites hundreds of observations of targeted helping in apes, dolphins, and elephants. For instance, researchers have frequently observed that when juvenile orangutans get stuck in a tree, they are rescued by their mothers. The mother orangutans drape their bodies between one branch and another, thereby creating a bridge to safety for their whimpering offspring. This behaviour clearly requires an understanding of the causes of the juvenile's emotional distress. De Waal holds that the mechanisms and cognitive abilities associated with empathetic perspective-taking must be complex. For this reason, EPT is exhibited by only a few large-brained animal species other than humans – particularly the great apes, elephants, and dolphins (De Waal 2008, p. 286).

De Waal argues that the more complex levels of empathy are descended from, and even enabled by, the simpler levels (De Waal 2008, pp. 287–288). He hypothesizes that EPT evolved through processes of reciprocity and kin selection from simpler mechanisms related to emotional contagion. Emotional contagion and sympathetic concern could enable the organism to share in the emotional state of the distressed subject while developing a motivation to alleviate that distress. The work left for EPT, then, would be to identify the cause of the subject's distress and to determine how to alleviate it. Hence, De Waal suggests that the capacity for EPT would have been a relatively new trait that emerged with the common ancestor of humans and apes, whereas the capacity for emotional contagion was inherent in our common ancestor with monkeys and probably in far more ancient organisms (De Waal 2008, p. 292).

The mechanisms that De Waal theorizes to explain directed altruism in nonhuman primates, and ultimately in humans as well, are consistent with the Humean theory of motivation. For instance, a chimpanzee with the capacity for empathetic perspective-taking shares in the emotional state of the one in need or distress, assesses the causes of the need or distress, and then takes the means necessary to meet the needs of the other. The chimp can be said to be motivated by Humean motivational mechanisms, including a desire to meet the needs of a conspecific, and a belief about the means necessary to meet those needs.

In contrast, De Waal's Russian Doll Model is not consistent with the motivations postulated by anti-Humeanism. Anti-Humean mechanisms depend on a level of cognitive sophistication that is far beyond that of any nonhuman animal. As we saw, both desire-entailing beliefs and besires require an ability to make moral judgments, i.e. to apply moral concepts such as *forbidden*, *permissible*, *obligatory*, *virtuous*, *vicious*, *just*, and *unjust* to actions, persons, and states of affairs. But nonhuman animals do not have moral concepts. Richard Joyce convincingly argues that nonhumans lack moral concepts, because they cannot carry out a 'semantic ascent' through which they regard certain categories of actions as *worthy* or *unworthy* of acceptance, rather than merely *accepted* or *not accepted* (Joyce 2006, pp. 82–85). So, the cognitive limitations of nonhuman animals render them incapable of making moral judgments. If nonhumans cannot form judgments about what is morally required of them, or about what behaviour is worthy or unworthy of acceptance, they could not be motivated to act on such judgments. Thus, besires and desire-entailing beliefs could not have a place in the psychologies of even the most intelligent nonhumans.

On the other hand, it may be tempting to interpret some of De Waal's experimental findings as evidence that primates have primitive moral concepts. For instance, when a capuchin monkey is given a cucumber and a neighbouring capuchin is given a more attractive grape, the first monkey often reacts negatively, e.g. by throwing the cucumber (cf. De Waal 2006, pp. 44–49). However, Sarah F. Brosnan suggests that there is insufficient evidence to justify attributing such reactions to a 'sense of fairness'. She points out that it is impossible to tell whether nonhuman animals have an understanding of fairness, because there is no way to get them to describe the nature of that understanding (Brosnan 2010, p. 80).

If anti-Humean motivations exist, they would have appeared after the evolution of moral judgment in either modern human beings or in some hominid ancestor of *homo sapiens*. Given Joyce's argument that a semantic ascent is necessary for possession of moral concepts, a hominid ancestor species capable of moral judgment would need a robust capacity for language. Thus, anti-Humeanism implies a discontinuity in evolutionary history. It suggests that a new trait – either besires or desire-entailing beliefs – would have emerged without identifiable precursors. The continuity constraint counsels us to minimize such discontinuities, provided that it does not result in explanatory loss. Even where evolutionary discontinuities are necessary to explain what is observed, the continuity constraint teaches us to show that they are products of some evolutionary process. It is far from clear why desire-entailing beliefs or besires would have been selected for. However, perhaps the anti-Humean motivations did not themselves evolve from any process of selection. They may have been byproducts ('spandrels') of other selected traits that are unique to modern humans, such as large brain size or cerebral complexity. But even if that's the case, it is not clear how besires or desire-entailing beliefs could have been a *predictable*, or *retrospectively explicable*, byproduct of some selective force operating on pre-existing traits. Spandrels may not be selected for, but they can still be explained as products of selection for other traits. Ultimately, beliefs and desires, as Humeans understand them, stand a far better chance of being included in a continuous evolutionary etiology of moral motivation. For this reason, the continuity constraint supports a presumption in favour of Humeanism.

8.4 For Humeanism: The Argument from Morgan's Canon

For all that has been said about continuity, human beings are probably unique among the animals in possessing the ability to make moral judgments. An anti-Humean may point out that the argument from continuity overemphasizes the extent to which humans are *descended* from ape-like ancestors, while it underplays the extent to which we are marvellously complex *modifications* of our ancestors. In answering this objection, I shall appeal to another criterion for theory choice to show that there is no good reason to suppose that moral action must be explained by anti-Humean mechanisms which humans do not share with other animals. This principle is known as *Morgan's Canon*. It is so named after comparative psychologist C. Lloyd Morgan. In his *Introduction to Comparative Psychology* (1894), Morgan states his principle as follows:

> In no case may we interpret an action as the outcome of the exercise of a higher psychical faculty, if it can be interpreted as the outcome of the exercises of one which stands lower in the psychological scale. (quoted in Sober 1998, p. 224)

Morgan's Canon applies to theories that explain a creature's behaviour by citing its internal psychological mechanisms. This is, of course, how the dominant theories of moral motivation purport to explain moral action. It's quite appropriate, therefore, to evaluate theories of moral motivation with respect to Morgan's Canon.

Although Morgan himself believed his principle to be justified by Darwinian considerations, Elliot Sober argues that it can be substantiated by deductive logic. To demonstrate this, Sober offers a novel interpretation of Morgan's Canon, summarized as follows (Sober 1998, pp. 236–237):

> **Morgan's Canon**
>
> Let one internal mechanism, H, be higher than another, L, if and only if the behavioral capacities entailed by H properly include the behavioral capacities entailed by L. Suppose, for instance, that L entails the behaviors in set B1 and H entails the behaviors in both sets B1 and B2 (where B1 and B2 are not equivalent sets).
>
> Then:
>
> if an organism performs the behaviors in B1, but not the behaviors in B2, the organism lacks H.

On Sober's interpretation of Morgan's Canon, so-called 'higher' psychological mechanisms enable an organism to do more than 'lower' ones. Accordingly, *H* is a higher mechanism than *L*, because an organism with *H* can perform the behaviours in both sets *B1* and *B2*, whereas an organism with only *L* can only perform the behaviours in set *B1*.

So construed, Morgan's Canon states that attributions of psychological mechanisms are constrained by the following deductive inference:

1. Mechanism *H* entails the behaviours in *B1* and *B2*.
2. Organism *O* does not display the behaviours in *B2*.
3. Therefore, by *modus tollens*, *O* does not have *H*.

Thus, Morgan's Canon advises us to attribute a psychological mechanism to an organism only if it's the *sole* mechanism which can explain some behaviour. This is not to say that all behaviours of the same type must be explained by *the same* mechanism. To use Sober's example, stickleback fish and chimpanzees both build nests, but Morgan's Canon would not imply that the same psychological process must motivate nest-building in both species (Sober 1998, p. 230). Instead, the

Canon allows the psychological process which motivates nest-building in chimps to be different, but only if this process *alone* could explain additional behaviours exhibited by chimps and not by sticklebacks.

Sober offers an example to illustrate how Morgan's Canon should guide theory choice in comparative psychology. When a species of bird called the piping plover sees a predator approach its nest, it moves away from the nest, sits down, and starts peeping loudly. In so doing, the bird makes it appear as though it's sitting on eggs, when really it isn't (Sober 1998, p. 238). This behaviour has been called 'false nesting'. What explains this curious behaviour? Consider two possibilities. Let *H1* be the hypothesis that piping plovers want to protect their young, and they believe false nesting will accomplish this without knowing exactly *how*. Let *H2* be the hypothesis that the plovers want to protect their young, and they believe false nesting will accomplish this *by* causing the predator to believe that the eggs are not in their true location. Unlike *H1*, *H2* attributes *second-order intentionality* to the plovers. Second-order intentionality is the capacity to form intentional states (e.g. beliefs or desires) *about* the intentional states of others. If piping plovers have second-order intentionality, then a plover can form the belief that the false nesting display will make an approaching predator incorrectly believe that the plover is sitting on its eggs. However, Morgan's Canon instructs us to reject *H2* in favour of *H1*, since the plover does not behave in ways that are explained *only* by second-order intentionality. False nesting can be exhaustively explained by *H1*, the hypothesis of *first-order intentionality*. The plover may simply believe that the false nesting display will lure the predator away from the true location of the eggs, without believing that the display will lure away the predator *by* manipulating the predator's beliefs about the location of the eggs. By contrast, the plover does not exhibit any behaviours which are *uniquely* explained by second-order intentionality. And so, by Morgan's Canon, it should be concluded that the piping plover lacks second-order intentionality.

Morgan's Canon can be deployed in an argument against anti-Humeanism. The argument begins with an observation that is hard to dispute: the Humean belief–desire model can successfully explain at least *some* human behaviours. When explaining the actions of other people, we often appeal to the beliefs and desires that motivated them. Philosophers call this explanatory strategy *folk psychology*. Folk psychology is commonly deployed in explanations of *non-moral* action. By 'non-moral action' I mean action which is not morally significant. Consider the act of tying your shoe. You are under no *moral* obligation to tie your shoe. Neither is it morally wrong, nor morally praiseworthy, nor blameworthy for you to tie your shoe. In and of itself, shoe-tying is a non-moral act. Folk psychology provides a straightforward explanation for your decision to tie your shoe: you have a desire to walk about with shoes on, and you believe that tying your shoe will enable you to do so successfully. According to Morgan's Canon, a 'higher' psychological mechanism should be attributed to an agent only if it can explain behaviours that a 'lower' mechanism could not. If we follow the Canon, then the burden is on the anti-Humean to show what behaviours, if any, can *only* be explained by anti-Humean mechanisms such as besires or desire-entailing beliefs. If there are no such behaviours, then Morgan's Canon would call for dispensing with the anti-Humean mechanisms.

The most likely anti-Humean rejoinder is to argue that there is no adequate way of explaining *moral* action other than by appeal to besires or desire-entailing beliefs. Yet I submit that motivational Humeanism can fully account for moral action. From a Humean perspective, moral action is explained by either a self-regarding, other-regarding, or moral desire plus a belief about how to satisfy the desire. Of these three types of desire, moral desires are the least well-understood. Accordingly, in the remainder of this section I shall elucidate the role that moral desires can play in a Humean explanation of moral action.

In my view, moral desires are induced by what Jonathan Haidt and other researchers call *moral emotions* (cf. Haidt 2003; Moll et al. 2008). According to Haidt, moral emotions are 'those emotions that are linked to the interests or welfare of society as a whole or at least of persons other than the judge or agent' (Haidt 2003). In his review of moral emotions research, Haidt says that moral emotions are generally elicited by situations and events that affect others (Haidt 2003). But judging from the details of Haidt's presentation, another critical factor which sets moral emotions apart from other emotions is their link to moral judgment. For instance, *self-critical* moral emotions, such as guilt, shame, and embarrassment, are elicited when one judges oneself to have committed some kind of moral violation, while *other-critical* moral emotions, such as indignation, anger, contempt, and disgust, are elicited when one judges someone else to have acted immorally (Haidt 2003; Moll et al. 2008, pp. 2–3, 6). Furthermore, moral emotions have effects on motivation. They are said to generate *action-tendencies,* i.e. they put a person in a motivational state in which 'there is an increased tendency to engage in certain goal-related actions (e.g., revenge, affiliation, comforting, etc.)' (Haidt 2003).

The action-tendencies of moral emotions provide evidence for the link between moral emotions and moral desires. Moral emotions have motivational effects by virtue of inducing moral desires in the agent. Consider anger. The action tendencies associated with anger include the motivation to redress situations the subject judges to be a moral violation. Evidence that anger motivates people to redress perceived moral wrongs can be drawn from an experimental economics paradigm involving the *ultimatum game*. In the ultimatum game, two players are shown a sum of money, say $10. The first player, called the 'proposer', is instructed to offer any whole number of dollars, from $1 to $10, to the second player, who is called the 'responder'. The proposer can make only one offer, and the responder can either accept or reject this offer. If the responder accepts the offer, the money is shared according to the terms of the offer. If the responder rejects the offer, neither player receives anything.

When the ultimatum game is played by people in non-repeated experimental trials and with varying amounts of money (including large sums of money), proposers most often offer 50 % of the original sum and respondents reject offers below 20 % about half of the time (Gintis et al. 2007, p. 608; Sanfey et al. 2003, p. 1755). Participants in these experiments reported that they reject low offers because they are angered by offers they judge to be unfair (Sanfey et al. 2003,

p. 1756; cf. Pillutla and Murnighan 1996). In addition, Sanfey et al. (2003) used fMRI scans on subjects playing the ultimatum game, and found that brain areas associated with anger were activated in responders by low or ‘unfair’ offers of $2:$8 or less. Activation of the bilateral anterior insula is involved in the neural realization of specific other-critical moral emotions, including disgust and anger (Sanfey et al. 2003, p. 1757; Moll et al. 2008, p. 15). Greater activations of the bilateral anterior insula were correlated positively with the degree of an offer’s unfairness. The anterior insula exhibited greater activations in response to offers of $1:$9 than to $2:$8, and offers of $2:$8 generated greater activations than ‘fair’ offers of $5:$5. Moreover, the magnitude of activation in this brain region correlated with subsequent decisions to reject the offer. Participants with stronger anterior insula activations in response to unfair offers were more likely to reject an unfair offer. Also, the anterior insula exhibited more intense activation in response to unfair offers from a human proposer than unfair offers from a computer. From these results, Sanfey and colleagues conclude that activation of the anterior insula is involved in the neural realization of anger directed at persons who intentionally violate a principle of fairness. The experience of anger in turn motivates resistance to unfairness in the form of a rejected offer (Sanfey et al. 2003, p. 1756).

The Humean can offer an intuitive explanation for Sanfey et al.’s findings: anger generated by an offer judged to be unfair induces in subjects a moral desire to reject the offer. People in Sanfey’s experiments understood themselves to be rejecting a low offer because they judged it to be unfair. So here we have instances of a moral judgment generating a moral action. Interestingly, fMRI imaging suggests that the mechanism through which judgments of unfairness motivated rejections of offers is entirely consistent with the Humean belief–desire model. Thus, the moral actions observed in Sanfey’s study are not of a sort that *only* the anti-Humean machinery of besires or desire-entailing beliefs would be able to produce. And since the anti-Humean motivations are not uniquely necessary to moral action, Morgan’s Canon would have us discard an anti-Humean account in favour of a Humean belief–desire model which is sufficient to explain both moral and non-moral behaviour.

8.5 Explanatory Deficits of Anti-Humeanism

I have been arguing that moral action can be exhaustively explained by the Humean belief–desire model. Since anti-Humean motivations are not necessary to explain such action, Morgan’s Canon delivers yet another reason to reject anti-Humeanism. Now the argument will take a more pointed turn. It will be shown that anti-Humeanism should be abandoned outright, because it is incompatible with observations drawn from recent work in neuroscience. On the other hand, the mechanisms which actually operate to produce both moral and non-moral action *only* fit a Humean mould.

Studies in neurobiology and neuroeconomics have attempted to isolate the brain mechanisms responsible for moral action. For instance, a large body of evidence shows that, in animal brains, the ventral striatum and nucleus accumbens, along with the insula and the orbitofrontal cortex, respond to the satisfaction of basic biological needs such as food, shelter, and social contact. In humans, the same brain areas respond to these same goods, *and* to abstract rewards like receiving money. Furthermore one study by Moll and colleagues reports that increased activity in the ventral striatum is observed with fMRI scans when human subjects anonymously and voluntarily choose to donate money to a charity (Moll et al. 2006). In the experiment, subjects had the choice of donating up to US$128 to a real charitable organization, or not doing so. The ventral striatum was activated both by the receipt of pure monetary rewards and by decisions to donate. Indeed, the ventral striatum was *more* active when subjects made donations than when they received monetary rewards (Moll et al. 2006, p. 15624). Since the voluntary choice of donating to charity often issues from the moral judgment that donating is *the right thing to do*, such a choice can be regarded as a moral action. Moll et al.'s findings suggest that the same mechanisms which respond to the satisfaction of non-moral desires (e.g., for attaining food and money) also respond to the performance of moral action.

In another neuroimaging study, researchers compared human subjects' neural reactions to receiving money and being forced to give money to charity (Harbaugh et al. 2007). In a first treatment, the researchers forced subjects to donate money to a charity in a tax-like condition, and measured consequent increases in brain activity in the ventral striatum, the head of the caudate, and the nucleus accumbens. In a second treatment, the researchers measured increased activity in the same brain areas when subjects were *given* a sum of money. The authors discovered that the difference between these two measures reliably predicts people's willingness to donate to charity on a *voluntary* basis. People who exhibited greater neural responses (in the mentioned brain regions) to mandatory donations, relative to receiving money, were more willing to donate voluntarily (Mayr et al. 2009, p. 308).

Harbaugh et al. also suggest that activity in the ventral striatum and other brain regions is a neural indicator of the *utility* a subject receives both from getting money and from voluntarily giving money to charity (Harbaugh et al. 2007, pp. 1623–1624; Mayr et al. pp. 308–309). 'Utility' is a technical term behavioural scientists use to refer to the satisfaction of preference or desire. The authors support this conjecture by using measured increases in neural activity in the ventral striatum and insula as a basis for comparing the relative strength of an individual's preference for receiving money as compared to his or her preference for donating money. A selfish person who strongly prefers getting money to donating money would only be willing to give up very little of his money in order to donate an additional unit of it. But an altruistic person who has a stronger preference for donating money would be willing to give up considerably more of his money in order to donate. Harbaugh and colleagues found that by modeling their subjects as *utility-maximizers* in this way, they could accurately predict how much money each subject was willing to donate voluntarily to charity. It's important to appreciate that the theory of utility-maximization is itself a version of motivational Humeanism. Utility-maximizing agents are agents who act to maximize the satisfaction of their preferences or desires. Accordingly, Harbaugh's findings provide compelling empirical evidence in favour of Humeanism, because they show that a belief–desire model can successfully *predict* moral action.

At the same time, the neuroeconomics research makes trouble for anti-Humeanism. The most prominent arguments for anti-Humean motivations emphasize that these motivations are needed to explain *moral* thought and action specifically. Thus, for example, David McNaughton argues that besires must be postulated in order to explain how one can be motivated by a 'purely cognitive' awareness of a moral requirement:

> To be aware of a moral requirement is, according to the realist, to have a conception of a situation as demanding a response. Yet to conceive of a situation as demanding a response, as requiring one to do something, is to be in a state whose direction of fit is: the world must fit this state . . . .But the realist also wishes to insist that the agent's conception of the situation is purely cognitive. That is, the agent has a belief that he is morally required to act and so his state must have the direction of fit: this state must fit the world . . . . (McNaughton 1988, p. 109)

In McNaughton's view, besires are needed to account for how a moral judgment can motivate action, *given* the hypothesis that awareness of a moral requirement is a 'purely cognitive' state. Additionally, Little notes that virtue ethicists like McDowell (1998) invoke the notion of a desire-entailing belief for a very similar purpose:

> The virtue theorist's claim is that a kind of cognitive state – a kind of state that does satisfy a belief direction of fit – necessarily brings with it the motivation to act as it says we ought. There are certain *ways* of seeing or of conceiving the world, as many have put it, that one cannot have without reacting affectively in a certain way . . . . (Little 1997, p. 261)

Thus, from an anti-Humean perspective, when someone acts to meet the requirements of morality, i.e. whenever someone acts morally, he or she will be acting on a special type of cognitive state that both apprehends the morally relevant features of a situation and generates a motivation. Moreover, these anti-Humean motivations *only* motivate moral action. They are not necessary to explain non-moral actions like tying one's shoe, because these actions are not (necessarily) motivated by the recognition of any moral requirement.

Now if anti-Humean motivations are realized in brain processes, we should expect to see something *special* going on in people's brains when they undertake moral actions – something that would *not* be going on when they perform non-moral acts. If there were besires or desire-entailing beliefs, they would have special neural correlates, and those neural correlates would be at work when anti-Humeans say they are at work – viz. when people act in conformity with their moral judgments.

However, neural imaging of what happens in people's brains when they donate to charity – an act that can be construed as a moral act – have so far shown that *nothing* special is going on in the brain when people engage in moral action. Instead, the same neural mechanisms that are activated when people (*and* animals, for that matter) receive rewards, eat food, find shelter, or partake in social bonding are also at work when test subjects make morally significant charitable donations. By contrast, Humeanism is not in any way threatened by Harbaugh et al.'s conclusion that people give money to charitable organizations because 'these transfers are associated with neural activation similar to that which comes from receiving money for oneself' (Harbaugh et al. 2007, p. 1624). Humeans regard the motivations driving moral action to be of the same kind as non-moral (intentional) motivations: they are desire-like or affective states. So it is to be expected, from a Humean viewpoint, that the same brain processes which underlie motivations to acquire food and money would also underlie motivations to act morally. Furthermore, when activity in the striatum was modelled as an indicator of the utility one derives for charitable donations, donating behaviour was successfully predicted. The model assumes that people are motivated to voluntarily donate to charity because it is rewarding to them. The neural reward for voluntary giving is registered in the same way as the neural reward for getting money, since both events elicit activity in the ventral striatum. And yet, receiving money does not call for recognizing the moral requirements of a situation. So, the motivation to donate to charity stems from a neural reward that occurs independently of any state of mind which involves recognizing the moral requirements of a situation. The independence of the motivation-inducing reward for donating from any moral judgment about the rightness of donating is compatible with Humeanism, but quite contrary to anti-Humeanism (Zangwill 2008a, b).

8.6 Conclusion

Three considerations have been adduced as evidence in favour of the Humean theory of motivation, and against motivational anti-Humeanism. First, Humeanism is much more compatible with De Waal's theory of how motivations to act morally could have evolved from simpler precursors resembling the proto-moral motivations of nonhuman primates. Second, anti-Humean motivations are not uniquely necessary to explain any behaviour that could not be explained by the Humean belief–desire model. Indeed, anti-Humean motivations aren't even necessary to explain *moral* behaviour. And third, despite the anti-Humeans' insistence that only besires or desire-entailing beliefs can explain motivations to act in accordance with the recognition of moral requirements, neuroimaging studies of people engaging in moral action yield no indication of any special neural process which is not successfully explained by a Humean framework of utility maximization.

Humeanism carries significant implications for moral philosophy. It suggests that moral motivation cannot be a purely cognitive achievement. For even if there were mind-independent moral truths, knowledge of these truths wouldn't be sufficient to direct behaviour. People can't simply know how the world ought to be; they must also *want* to change the world accordingly. The study of how to increase people's desires to do justice is no less a worthy enterprise than the study of what justice is.

Acknowledgments Many thanks to David Wong, an anonymous referee, and the participants of the "What Makes Us Moral?" Conference, VU University Amsterdam, for their helpful feedback on this essay.

References

Altham, J.E.J. 1986. The legacy of emotivism. In *Fact, science, and morality: Essays on A.J. Ayer's language, truth and logic*, ed. G. Macdonald and C. Wright, 275–288. Oxford: Basil Blackwell.

Blackburn, S. 1984. *Spreading the word*. Oxford: Oxford University Press.

Brosnan, S.F. 2010. Fairness and other-regarding preferences in nonhuman primates. In *Moral markets: The critical role of values in the economy*, ed. P.J. Za and M.C. Jensen, 77–104. Princeton: Princeton University Press.

De Waal, F. 2006. Morally evolved: Primate social instincts, human morality, and the rise and fall of 'veneer theory'. In *Primates and philosophers: How morality evolved*, ed. S. Macedo and J. Ober, 1–58. Princeton: Princeton University Press.

De Waal, F. 2008. Putting the altruism back into altruism: The evolution of empathy. *Annual Review of Psychology* 59: 290–300.

Dennett, D.C. 1995. *Darwin's dangerous idea: Evolution and the meanings of life*. New York: Simon & Schuster.

Gibbard, A. 1990. *Wise choices, apt feelings: A theory of normative judgment*. Cambridge, MA: Harvard University Press.

Gintis, H., S. Bowles, R. Boyd, and E. Fehr. 2007. Explaining altruistic behavior in humans. In *Oxford handbook of evolutionary psychology*, ed. R.I.M. Dunbar and L. Barrett, 605–619. New York: Oxford University Press.

Haidt, J. 2003. The moral emotions. In *Handbook of affective sciences*, ed. R.J. Davidson, K.R. Sherer, and H.H. Goldsmith, 852–870. Oxford: Oxford University Press.

Harbaugh, W.T., U. Mayr, and D.R. Burghart. 2007. Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science* 316: 1622–1625.

Joyce, R. 2006. *The evolution of morality*. Cambridge, MA: MIT Press.

Little, M.O. 1997/2007. Virtue as knowledge: Objections from the philosophy of mind. In *Foundations of ethics*, ed. R. Shafer-Landau and T. Cuneo, 252–264. Malden: Blackwell Publishing Ltd.

Mackie, J.L. 1977. *Ethics: Inventing right and wrong*. London: Penguin.

Mayr, U., W.T. Harbaugh, and D. Tankersley. 2009. Neuroeconomics of charitable giving and philanthropy. In *Neuroeconomics: Decision making and the brain*, ed. P.F. Glimcher, C.F. Camerer, E. Fehr, and R.A. Poldrack, 303–320. London: Elsevier.

McDowell, J. 1998. *Mind, value, and reality*. Cambridge, MA: Harvard University Press.

McNaughton, D. 1988. *Moral vision: An introduction to ethics*. Oxford: Basil Blackwell.

Moll, J., F. Krueger, R. Zahn, M. Pardini, R. Oliveira-Souza, and J. Grafman. 2006. Human fronto-mesolimbic networks guide decisions about charitable donation. *Proceedings of the National Academy of Sciences* 103(623): 15623–15628.

Moll, J., R. Oliveira-Souza, R. Zahn, and J. Grafman. 2008. The cognitive neuroscience of moral emotions. In *Moral psychology*, vol. 3, ed. W. Sinnott-Armstrong, 1–18. Cambridge, MA: MIT Press.

Pillutla, M.M., and J.K. Murnighan. 1996. Unfairness, anger, and spite: Emotional rejections of ultimatum offers. *Organizational Behavior and Decision Processes* 68(3): 208–224.

Prinz, J.J. 2004. *Gut reactions: A perceptual theory of emotion*. Oxford: Oxford University Press.

Rosati, C.S. 2006. Moral motivation. In *Stanford encyclopedia of philosophy*, ed. E.N. Zalta. http://plato.stanford.edu/entries/moral-motivation/. Accessed 29 June 2012.

Sanfey, A.G., J.K. Rilling, J.A. Aronson, L.E. Nystrom, and J.D. Cohen. 2003. The neural basis of economic decision-making in the ultimatum game. *Science* 300: 1755–1758.

Shafer-Landau, R. 2010. *The fundamentals of ethics*. New York: Oxford University Press.

Smith, M. 1987. The Humean theory of motivation. *Mind* 96(381): 36–61.

Sober, E. 1998. Morgan's canon. In *The evolution of mind*, ed. D. Dellarosa Cummins and C. Allen, 224–242. Oxford: Oxford University Press.

Thompson, P. (ed.). 1995. *Issues in evolutionary ethics*. Albany: State University of New York Press.

Zangwill, N. 2003/2007. Externalist moral motivation. In *Foundations of ethics*, ed. R. Shafer-Landau and T. Cuneo, 243–251. Malden: Blackwell.

Zangwill, N. 2008a. Besires and the motivation debate. *Theoria* 74: 50–59.

Zangwill, N. 2008b. The indifference argument. *Philosophical Studies* 138: 91–124.

Chapter 9
Whispering Empathy: Transdisciplinary Reflections on Research Methodology

Harry Wels

9.1 Introduction

'What makes us moral?' is the leading question of this volume. Morality can be considered a *sine qua non* for animals, both human and non-human, to live a social life (cf. De Waal 1997; Ridley 1996; Dugatkin 2006). This morality seems based on the ability of social animals, again human and non-human alike,[1] to be empathetic towards others (Bekoff and Pierce 2009, p. 1). *Ergo*, empathy might be considered a precondition for morality. An abundance of rather convincing (empirical) evidence is being brought forward nowadays, primarily from the biological sciences, stating that empathy is certainly not uniquely human, but is widespread amongst many other social animals like elephants, dolphins, primates and even animals like vampire bats (see Bradshaw 2009a; De Waal 2009; Bekoff and Pierce 2009; Balcombe 2010). A review of a recent publication on 35 years of elephant research in Amboseli, Kenya, by Cynthia Moss and her team, says that 'Moss's team has individually tracked about 2,500 elephants, observing and recording their lives over the decades. They conclude there is 'no doubt that elephants display empathy for one another . . . . The sheer range of emotions and their ability to use tools shown in the new study has led researchers to conclude that elephants should be considered at least as

[1] In this paper 'social animals' refers to both human and non-human animals, unless indicated otherwise. This is in line with Calarco (2008), who is following Donna Haraway's 'A cyborg manifesto' in which she states that 'By the late twentieth century . . . the boundary between human and animal is thoroughly breached . . . many people no longer feel the need for such a separation' and argues that '*we could simply let the human–animal distinction go*' (Calarco 2008, pp. 148, 149, italics in original). For a thorough historical exploration of what it means to be human, see Bourke (2011).

H. Wels (✉)
Department of Organisational Sciences, VU University, Amsterdam, The Netherlands
e-mail: h.wels@vu.nl

B. Musschenga and A. van Harskamp (eds.), *What Makes Us Moral? On the capacities and conditions for being moral*, Library of Ethics and Applied Philosophy 31,
DOI 10.1007/978-94-007-6343-2_9, © Springer Science+Business Media Dordrecht 2013

similar to humans as some of the most advanced animals.'[2] However, this idea is complicated by questions about the boundaries of empathetic capabilities in social animals (cf. Coetzee 1999). Can any given social animal be empathetic with any 'other', even across the species divide? Can an elephant feel empathy for a wild dog? Can a human being feel empathy for a cow? Can a wolf feel empathy for rodents? Or is empathy limited to *certain* others? The closer the other is to you in terms of species, the more empathy is to be expected? Therefore you can feel empathy for your pet dog that is considered part of the family, but less for the pig on its way to the abattoir? These are all questions about empathy *in* social animals, but what about the research methodologies used to study social animals? To what extent does this type of research depend on the empathetic capabilities of the researcher, as is the case when studying human animals? And how does that link to my earlier questions? Should a cognitive ethologist have as many empathetic capabilities for studying animal behaviour as an ethnographer studying human behaviour?[3] In this paper I want to explore the concept of empathy in relation to research methodologies in the social and biological sciences, and develop the argument that a research approach to studying social and moral behaviour in human and non-human animals based on empathy could be explored as a transdisciplinary research methodology that will bring ethologists and ethnographers together to further our understanding of social behaviour. Furthermore I will argue that Deleuze and Guattari's (1980) concept of 'becoming (animal)' can be used as a stepping-stone for deepening our conceptual understanding of empathy as a research methodology.

9.2 Ethological and Ethnographic Research and the Need for Empathetic Understanding

A great breakthrough in the study of social animals, certainly in popular imagery, was brought about by the ethological fieldwork of Jane Goodall on chimpanzees, Dian Fossey on mountain gorillas and Biruté Galdikas on orangutans (Montgomery 2009). It was considered revolutionary because 'their methods of study were much more like those approved for anthropologists than like those approved for wildlife biologists' (Marshall Thomas in Montgomery 2009, p. xiv). It was also seen as groundbreaking because with their particular fieldwork approach they challenged 'the masculine world of Western science' (p. xix), and their work was therefore to be heralded as a triumph of the feminine approach to science (p. 238). An approach that could be captured by the Japanese word *kyokan*, derived from Kawai Masao (1969), a primate researcher from Japan. *Kyokan* means '*becoming* fused

[2] *The Sunday Times*, 5 June 2011, 'Jumbos just like us.'

[3] 'Cognitive ethology' is a concept coined by Donald Griffin (1978), meaning 'an approach to animal behavior which attributes "mentality" to animals' (Skipper 2004, p. 483). Whenever I refer to 'ethology' or 'ethologist', this implies 'cognitive ethology' or 'cognitive ethologist'.

with the monkeys' lives where, through an intuitive channel, feelings are mutually exchanged', that is, 'to feel one with them' in a shamanistic way (pp. 238–239, my italics). A shamanistic way of 'becoming animal' that is also proposed by Deleuze and Guattari (1980): 'A similar ontological process to shamanism which undermines fixed identities, as well as crossing thresholds' (in Westwood 2008, p. 4; also see the extensive discussion on 'becoming' in Ten Bos 2008, pp. 75–91). This view is described in a tongue-in-cheek manner by Michael Ryan writing about Merlin Tuttle's interest in bats, saying that he had 'rarely met anyone with such *a feel* for their study animal' (Ryan 2010, p. 478, italics added), and jokingly remarking that he thought that 'Merlin was part bat' (p. 477). At the same time it should be recognized that this close proximity to and scientific study of animals can also lead to a rather mechanistic perspective on animals, as the life and work of famous and Noble prize winning ethologist Niko Tinbergen illustrates. He is regarded as the father of the strand of ethology that assumes that 'we cannot know what an animal feels or what it intends, so scientists should not speculate on its subjective experience' (Kruuk 2003, p. 3). In the same biography though, it is observed that this particular position and argument may be attributed to the fact that

> Niko never felt comfortable in the presence of dogs, and did not keep any himself in later life. It is quite possible that if he had grown up with one, it would have been more difficult for him to see animals mechanistically as he did in his later science, and it would have been more difficult to sideline animal emotions and feelings. (Kruuk 2003, p. 19)

One of the most well-known scientists studying empathy in social animals (chimpanzees) is probably Dutch primatologist Frans de Waal, who defines empathy as the capacity to: (a) be affected by and share the emotional state of another; (b) assess the reasons for the other's state; (c) identify with the other, adopting his or her perspective (De Waal 2008, p. 281).

This definition seems to refer as much to the object of study, that is, empathy in chimpanzees, as to empathy as a research methodology, without De Waal making the distinction explicit. It becomes clear from this example that we need to make a distinction between researching empathy in social animals and using empathy as a research methodology. My suspicion is that many scientists interested in empathy in animals, actually also use empathy as a research methodology without overtly stating so, and without defining clearly what empathy means as a research methodology. De Waals's approach actually seems to be very similar to the definition of *kyokan* given above. This shouldn't come as a surprise, as throughout his many books De Waal has always written very highly of his Japanese colleagues, who, according to him, were also the first to start researching 'culture' among primates. This research was inspired by the work of Kinji Imanishi, who argued that non-human animals are not instinct-driven, i.e. Cartesian machines, and do not learn and know by instinct (i.e. genetically), but through social learning (i.e. culture) (De Waal 2001). This has resulted in various culture studies among animals, led by Japanese researchers (De Waal 2010). De Waal's appreciation for the specific contribution of Japanese primate researchers is shared by others, Asquith (2002) among others. In the foreword of his 2002 book, Hiroyuki Takasaki, one of the translators of the work of Imanishi, writes that the 'discovery of cultural behaviours [of primates] is also traceable to his worldview, which encourages anthropomorphism when judged appropriate' (Asquith 2002, p. xii).[4] It seems that De Waal's interaction with his Japanese colleagues over the years has almost inevitably led him to the concept of empathy, to which he has devoted one of his latest books (2009). De Waal is interested in the extent to which empathy can be observed in non-human animals, and more particularly primates. He argues convincingly, as others do (see for instance Bekoff 2007), that empathy is widespread amongst social animals, but he does not seem to be explicitly aware of his use of the concept as a research methodology.

Interestingly enough, it seems that 'empathy as an approach to fully understand others, especially non-human animals' has penetrated popular culture more deeply than it is broadly recognized or accepted as a route to knowledge in science. Popular culture abounds with stories, articles, documentaries and books about people who through sheer empathy learn the ways of an animal and are able to literally live with them. The rather scientific ring of the word empathy has found its expression in popular discourse, in all sorts of (wild) animal whisperers, whispering to dogs (Fennell 2001), horses (Roberts 2008), lions (Richardson 2010), wolves (Ellis 2009), bears (Treadwell and Palovak 1997), and elephants (Anthony 2010). It is important to note here that the various whisperers talk about 'understanding' the other in the realm of their ordinary and everyday life. Whisperers do not focus on extravagant events, dramas, festivities, or celebrity aspects of the animals' lives. It is all about the everyday. Naturally, to every human animal the everyday of the non-human animal is out of the ordinary. It is this methodological approach of the peculiarity of the everyday that shows that ethnography, as the study of the everyday of human interaction, and ethology, which investigates the everyday in non-human animal behaviour, actually share the same roof.

The proof of this pudding could be the following quote from a recent edited volume on organizational ethnography:

> Ethnographic fieldwork typically involves the development of close connections between the ethnographer and the subject and situations being studied; (. . .) in order to understand 'what goes without saying', intimate knowledge of other people's lifeworlds is indispensible. (Ybema and Kamsteeg 2009, p. 101)

If I change this quote just a little, I suspect that no readers' eyebrows will be raised:

> *Ethological* fieldwork typically involves the development of close connections between the *ethologist* and the subject and situations being studied . . . in order to understand 'what goes without saying' . . . intimate knowledge of *the animals'* lifeworlds is indispensible.

[4] Liebenberg reports similar observations about the San in southern Africa. 'In order to understand animals, the [San] trackers must identify themselves with an animal' (Liebenberg 1990a, p. 88). Therefore, 'the [San] knowledge of animal behaviour essentially has an anthropomorphic nature' and 'although their knowledge is at variance with that of European ethologists, it has withstood the vigorous empirical testing imposed by its use.' Therefore, 'anthropomorphism may well have its origins in the way trackers must identify themselves with an animal' (Liebenberg 1990a, p. 83).

Ethologists and ethnographers share a belief in and a passion for fieldwork on everyday lifeworlds. They are 'out there' together. I could pull the same trick (the other way around) with the following long quote from Marc Bekoff on what fieldwork means to him as an ethologist:

> If I begin my research, as I often do, with a deceptively simple question like, 'what is it like to be a dog in such-and-such situation?' then I must try to understand how dogs get through their [every]day and nights from their dog-centric view of the world. On many occasions I've walked around on all fours, done play bows, howled, barked, bitten their scruffs, and rolled over on my back – though I draw the line at mimicking the all-important hindquarter sniff (I gladly leave that to the dogs). I try to go where the animals live to observe them, and as I study them, I also try to empathise with them. How would I feel if I were in the same situation? Of course, I always remember that *my* view of their world is not necessarily *their* view of their world, but the closer I can get to their view, even by personal analogy, the better I might be able to understand it. (Bekoff 2007, pp. 37–38)

Ethologists and ethnographers not only seem to live under the same roof, but even go in through the same door; both starting from the assumption that 'empathy increases understanding' (Masson and McCarthy 1995, p. 36). Lestel et al. argue for an 'ethno-ethology' that 'grants all living beings the status of relational beings, that is, agents interacting on the phenomenon of "culture" that was hitherto reserved for human beings' (Lestel et al. 2006, p. 168). According to Lestel et al. this approach should be complemented with 'etho-ethnology', in which an animal can be 'defined as a natural or artificial, human or non-human agent that attempts to control its actions and those of others as a result of the *significations* it ascribes to their behaviors'. Taken together this approach would be able to study the '*shared lives*' of human and animals, 'on the paradigm of *convergences* between the two, of *life shared* by *intentional* agents belonging to different species' (Lestel et al. 2006, p. 156). Other voices from the anthropological discipline speak of 'multispecies ethnographies' in which 'becomings' [cf. Deleuze and Guattari] – new kinds of relations emerging from non-hierarchical alliances, symbiotic attachments, and the mingling of creative agents ... – abound' (Kirksey and Helmreich 2010, p. 546). There is a clear overlap with Lestel's work (although he and his colleagues are not cited), since multispecies ethnographers, like in the approach that Lestel et al. propose, put 'a fresh emphasis on the subjectivity and agency of organisms whose lives are entangled with humans' (Kirksey and Helmreich 2010, p. 566).

But empathy not only offers shelter to ethologists and ethnographers, as many other disciplines increasingly make use of the same door. Take environmental historians for instance. In the conclusion of her latest book, which analyses the role of horses in South African history, Sandra Swart (2010) argues for studies that look at particular histories through the horse's eyes: 'It is an interesting and helpful exercise to write history through the eyes of the horse, forcing the human historian to adopt a new and sympathetically imaginative perspective' (Swart 2010, p. 217). Another example is psychology, a discipline that has already paid a lot of attention to the concept of empathy in its thinking about understanding the (ways of the) human psyche (for a fascinating overview, see Hakonsson 2003). To move from the psyche of men to the psyche of non-human animals is but a small step:

> Trans-species psychology allows us to imagine – without undue anxiety about anthropomorphism – what it might be like to walk in elephant 'shoes' and experience what these awesome herbivores might be thinking and feeling, in much the same way that we think about ourselves and other people . . . . (Bradshaw 2009a, p. 18)

'In merging ethology with psychology we recognize what humans and elephants share in brain and behavior, and we learn to ask more expansive questions about elephants than are usually included in ethological investigation' (Bradshaw 2009a, p. 72). The 'trick' I used above, playing around and swapping words in a quote, actually comes from Bradshaw, who did more or less the same when presenting an obvious case of Post-Traumatic Stress Disorder (PTSD) in elephants to five mental health specialists (that is, human mental health) without mentioning that the case was about elephants instead of humans (Bradshaw 2009a, pp. 95–98 and 108–112). All experts reached the PTSD diagnosis and suggested similar treatment. No one gathered from the neutrally formulated case description that it was not actually about humans. Bradshaw (2009b) actually argues for a trans-species psychology, as 'psychology and psychotherapy apply not only to the human psyche but equally and seamlessly to the psyches of our animal relatives' (Bradshaw 2009b, p. 157). Approaching and trying to understand another through empathy makes no distinction between the human and non-human other; empathy does not seem to be curtailed by 'speciesism' (which, according to some animal rights activists, should be compared to racism, see LaFolette and Shanks 1996)[5]; empathy seems to recognize social distress (in this case PTSD) across species.

Nevertheless, in *The lives of animals* (1999), Nobel laureate John. M. Coetzee lets his characters discuss the limits of our sense and applicability of empathy beyond our own species in a fictious setting, inspired by the famous essay by American philosopher Thomas Nagel, entitled 'What is it like to be a bat?' (1974). Here another aspect is added to our thinking about empathy, because the question Nagel tries to answer is not whether we can imagine how it would be for us as *humans* to be a bats, but 'what it is like for a *bat* to be a bat' (Nagel in Coetzee 1999, p. 31). Can we become another life form? Coetzee's main character, Professor Elisabeth Costello, thinks that we can, and argues that:

> there is no limit to the extent to which we can think ourselves into the being of another. There are no bounds to the sympathetic imagination. If you want proof, consider the following. Some years ago I wrote a book called *The House on Eccles Street*. To write that book I had to think my way into the existence of Marion Bloom . . . . *Marion Bloom never existed*. Marion Bloom was a figment of James Joyce's imagination. If I can think my

[5] Speciesism is defined by Joanne Bourke (2011, p. 132) as 'discrimination based on membership of a species'. As Bourke in her book (2011) makes abundantly clear, 'historically the two [speciesism and racism] are inextricably intertwined, the former being used to bolster, explain, and justify the latter' (LaFolette and Shanks 1996, p. 41). Based on Bourke's (2011) analysis we can add sexism as a third 'inextricably and intertwined' thread to LaFolette and Shanks quote.

> way into the existence of a being who has never existed, then I can think my way into the existence of a bat or a chimpanzee or an oyster, any being with whom I share the substrate of life. (Coetzee 1999, p. 35)

Hilary Mantel, Booker Prize winner for her novel on Thomas Cromwell in 2009, entitled *Wolf Hall*, tells us in a recent interview in a Dutch national newspaper: '*Ik stap in Cromwells schoenen*' (I step into Cromwell's shoes) and, *'het is me gelukt om* Wolf Hall *te schrijven alsof ik in hem zit. Ik overlap in tijd en ruimte*' (I managed to write *Wolf Hall* as if I lived inside of him [Cromwell]. I overlap in time and space). On the process of writing the novel, she tells the interviewer: '*was [ik] bezig Cromwell te worden*' (I was becoming Cromwell) (Roodnat 2011). In other words, training empathy is a training in imagination and becoming: 'We need to exercise our imaginative faculties, stretch them beyond where they have already taken us, and observe things we have never been able to see before' (Masson and McCarthy 1995, pp. xxi–xxii).

Training the imagination is maybe best done by reading stories on human–animal relations, in a way as suggested by Pierce's 'narrative ethology' (Bekoff and Pierce 2009, pp. 36–38). I followed that route when I read all kinds of popular stories about animal whisperers and looked at television series and movies about the same animals and whisperers. Another training device is reading philosophical studies, especially on human–animal relations, like for instance the posthumously published works of Jacques Derrida on human–animal relations, famously based on his experience of the gaze of his cat (2008) which brings him to an 'acceptance of the point of view of the cat' (Westwood 2008, p. 3). In this difficult text Derrida basically shows how the European philosophical tradition, from Kant to Lacan, to Descartes' animal-as-machine, to Heidegger, to Levinas, has systematically marginalized animals (see for a similar argument Calarco 2008). By not granting them a plural ('the animal'), as 'though all animals from the earthworm to the chimpanzee constituted a homogeneous set to which "(the hu)man" would be radically opposed' (Marie-Louise Mallet in Derrida 2008, p. x); by not ever granting animals the independent agency to respond (react yes, but not respond) to humans. As they do not speak our language, and as we are still far away from understanding the numerous languages animals speak (see for instance O'Connell 2007; Shanor and Kanwal 2009), we tell ourselves that they do not respond to us, and with that deny them independent agency. By not granting them a law that would make it possible to label the killing of an animal as murder, an animal is basically denied its own death, as it is denied its own life: 'The animal doesn't die, that is, ... one can put it to death without "killing" or murdering it, without committing murder ....' By not recognizing its ability to speak on its own behalf, in a sense the animal does not exist. Heidegger argues that stones are worldless (*weltlos*), animals are 'poor in world' (*weltarm*) and humans are world-forming (*weltbildend*) (in Derrida 2008, p. 151). Humans give meaning to the world, a widespread notion in most if not all the social sciences, especially since the 'interpretive turn' (see Yanow and Schwartz in Shea 2006). Typically for European philosophy, animals are sidelined with one blow: 'The animal can *mitgehen* with us in the house; a cat, for example ... can inhabit the same place as us, it can "go with us", "walk with us", it can be "with us" in the house, live "with us" but it doesn't "exist with us" in the house' (Derrida 2008, p. 145); according to Heidegger, it cannot *mitexistieren* (for interesting discussions on this work of Derrida, see Westwood 2008; Ten Bos 2008; Calarco 2008). Perhaps because of this persistent philosophical tradition, most people still feel that it is morally legitimate to eat animals in the large quantities produced in a mechanical way through the meat industry (cf. Pachirat 2011); as long as we cannot kill or murder animals, they cannot suffer by our hands and therefore it is no problem to eat them (see Safran Foer 2009). This philosophy inspired by Derrida and his post-modern contemporaries will maybe trigger our imagination in order to train our empathetic capabilities, but it will at the same time confront us with the fact that we have to transcend our species in a post-humanist way in order to be able to become the other. Would it still be possible to be warned against or accused of anthropomorphism, once you practice empathy and try to transcend species boundaries?

9.3 Developing Empathy as a Transdisciplinary Research Methodology

Any scientist daring to suggest this type of empathy-based research approach is often warned by fellow scientists not to go that route because of the danger of anthropomorphism, ultimately considered 'a form of scientific blasphemy' (Masson and McCarthy 1995, p. xviii), and an accusation in academia that is considered serious enough to hold your breath and think twice before going on. Nevertheless, a rather persistent challenge of this accusation seems to be going on both in the biological sciences and in the human sciences (Westwood 2008, pp. 5–8). We probably refrain from fully embracing cross-species empathy as a research method due to something that is strongly related not so much to the anthropocentrism as to the *logo*-centrism in social-science research.

It can be argued that a lot of ethnographic fieldwork, especially since the discursive turn, has become strongly focused on the audiovisual senses in terms of recording only (spoken) words and (written) texts. This is to the detriment of other research methods in empirical fieldwork, especially those making use of the other senses; smell, sounds and sights (other than words), touch and taste (cf. Pink 2009). While it is generally acknowledged that "*jij en ik (. . .) juist door te praten een hoop dingen niet [laten] zien*" (You and I . . . precisely by talking [leave] many things unseen) (Ten Bos 2008, p. 127), and that body language is far more important in communication than the spoken word (Mehrabian 1981), *logos* seems to be the centre of research attention in much of the ethnographic enterprise (nowadays); logo-centrism is the word that reigns. Contrary to ethnographers, ethologists in general, and those studying wild animals during extensive periods of fieldwork in particular, cannot rely on words and human language in their research and therefore develop all their senses in a broad range of observational skills, trying to understand the language animals 'speak' in a myriad of ways, ranging from body posture to smells, sounds, signals and combinations of these. Like a tracker they have to 'read the signs' (cf. Walker 1996) with all their senses. Louis Liebenberg (1990a) even argues that 'the art of tracking' is to be considered a form of pre-scientific thinking: on the basis of a range of tracks (i.e. 'signs', from prints to smell, to broken branches, to sounds, many of them incomplete), a tracker has to hypothesize where the animal has gone. If a tracker ever wants to catch up with the animal, he cannot look for every successive spoor, but has to try to define in what direction and where the animal is heading. Only finding the animal will prove the hypothesis. If the animal is not found, a new hypothesis has to be developed, until the animal is 'tracked down'. Although this may sound as a rather straightforward process, it is fraught with difficulties and complexities, as 'the art of tracking involves each and every sign of animal presence that can be found in nature, including ground spoor, vegetation spoor, scent, feeding signs, urine, feces, saliva, pellets, territorial signs, paths and shelters, vocal and other auditory signs, incidental signs, circumstantial signs and skeletal signs' (Liebenberg 1990b, p. 3). In other words, a data set that seems at first chaotic and is often incompletely constructed, needs to be interpreted in order to be able to find the animal. Clive Walker, perhaps with a little sardonic smile on his face, observes about a party that is taking a course in the art of tracking, that '*it* [is] *interesting* how many of the party ventured different interpretations of what animal had passed (Walker 1996, p. 10, italics added). Tracking is a methodological approach that makes use of all the senses and that can metaphorically be considered as a stepping-stone in developing empathy; tracking could then be seen as the basic methodological skill that ethnographers and ethologists alike should master, as 'tracks ... give an account of the animal's undisturbed *everyday life*' (Liebenberg 1990b, p. xi, italics added).[6] In this approach the researcher is attempting, by almost literally tracking footprints, to step into the footsteps of the Other, in order to find him or her; it is an attempt at 'becoming Other' in order to come to a better understanding of his or her wanderings and whereabouts, i.e. about 'finding' their everyday life. In this respect we as researchers could learn something from the most domesticated animal, at least according to ethologist and Nobel laureate Konrad Lorenz, who wrote in his classic 1949 study on domestic dogs and cats that particularly 'the degree to which dogs understand human expressions of feeling is little short of a miracle', due to his understanding that 'everything that socially living animals ... "have to say" to each other belongs exclusively to the plane of those interlocking norms of action and reaction which are innate in the animals of a species' (Lorenz 2002, pp. 129 and 127).

[6] The skills of especially the San people in southern Africa in tracking animals have been extensively and notoriously used both in Zimbabwe (by the Selous Scouts) and in South Africa (by Koevoet) to track down humans in the context of counter-insurgency operations, where so-called 'terrorists', or 'terrs', were 'tracked down' in order to be eliminated; people were 'hunted down' like animals (cf. Kamango 2011).

'Becoming Other' is a philosophical concept most explicitly developed by Giles Deleuze and Felix Guattari (1980), while at the same time forming part of a broader discourse in postmodern (primarily French) philosophy that seeks to capture and conceptualize the volatility of social reality. Deleuze and Guattari participate in this discussion by introducing the concept of 'becoming', which stands for the ultimate fluidity and flux of social reality, a reality that never reaches any final state or destination; 'becoming' never totally 'becomes', as it always remains an exploration of the other (Janssens and Steyeart 2001, p. 131). In that sense 'becoming' always lingers 'in the middle of difference'. 'Becoming animal' is therefore not an attempt to ultimately become the animal itself, but to try and understand the animal from the middle of one's relation with it; from the 'middle of difference' (Ten Bos 2008, p. 80). 'Becoming' aims to avoid looking at the Other from a dominant position of the self. Describing animals as 'non-human animals' for instance, categorizes them as something that is not similar to the dominant human, instead of trying to approach the animal from its own self (cf. Neolan in Janssens and Steyaert 2001, p. 128). 'Becoming' steers clear of using difference as an absence, or failure, of similarity, but aims instead at studying the other from the perspective of difference itself, from the middle. This means that becoming is a process of anonymizing the human subject, trying to reach the middle of difference in the relation with the other. In a way, 'becoming' leaves the self behind in its exploration of the other; in its journey to the middle. Interestingly, Deleuze and Guattari refer to children as an example of how to relate to animals without taking the self, which is not yet developed as such in children, along (in Ten Bos 2008, p. 89). This brings to mind the Biblical notion that to 'really' believe in God, is to *become* like a child, as only they will enter the Kingdom of God (cf. Mark 10:13). This parallel is even more suggestive as Ten Bos, following George Kampis (2008, p. 87), suggests that 'becoming' basically asks for a 'knowing without knowing', which seems to echo the Biblical dictum that to have faith is 'to be certain of the things we cannot see and to be sure of the things we hope for' (Hebrews 11:1). 'Becoming' asks for 'intensities of relationship', not with a single animal, because in order to try and understand the animal we must appreciate the totality or collective of contexts in which the animal lives, which in itself is continually 'becoming' (cf. Ten Bos 2008, pp. 87–89). It is an argument and interpretation that seems to fit and conceptually frame the stories about 'Tippi of Africa', a child born in the bush in Namibia who grows up loving, interacting and communicating with all the wild animals that surround her (Robert and Degré 1996) – in a way becoming them. She was called 'the real-life Mowgli', after Rudyard Kipling's famous jungle boy (*The Telegraph*, 12 November 2008). 'Intensive relationships' with the other, be it animals, plants, or other, facilitates 'becoming'. Barbara McClintock (1902–1992), the famous geneticist, basically asked of her students to work towards 'becoming plant', without the vocabulary of Deleuze and Guattari yet developed or available to her at the time. The students had to stay with a maize plant when it germinated and grew, cell by cell, into a full-grown maize plant, and she herself did the same. As McKlintock said, 'I don't really know the story if I don't watch the plant all the way along Only by cultivating an intense relation with the plants' becoming was she able to understand them and to work on their genetic modification: 'Over and over again, she tells us one must have the time to look, the patience to "hear what the material has to say to you", the openness to let it come to you". In short, one must have '"a feeling for the organism"' (McClintock in Fox Keller 1983, p. 198). McClintock maintained that 'good science cannot proceed without a deep emotional investment on the part of the scientist.' She sounds a bit like the trackers described above: 'At any time, for any plant, one who has sufficient patience and interest can see the *myriad signs* of life that a casual eye misses' (Fox Keller 1983, p. 200, italics added).

Barbara McClintock's example is significant for other reasons as well, since she was also a woman making a scientific career at a time when universities in the United States of America were dominated by white males. Critical feminists have therefore embraced Barbara McClintock as an example and icon of women's emancipation in science (see for example Tuana 1989). Of course more postmodern-oriented feminists have also taken up Deleuze and Guatarri's concept of becoming, emphasizing how it leads to the 'inter-connection of self and others, including the non-human or "earth" others' (Braidotti 2006, par. 36) in a way that resonates with 'the non-anthropocentric epistemologies of Donna Haraway' (ibid. par. 11; see also Haraway 2008). The Centre for Gender Research (GenNa) of the University of Uppsala in Sweden has started a specific research group on gender and animals, organized in the HumAnimal Group.[7] At a seminar in October 2011, they argued for the development of so-called 'zoo-ethnographies', which resonates much the same approach and themes as pointed out by Calarco (see above), and by Braidotti and Haraway in their call for papers for the seminar:

> Humanimal encounters are simultaneously creative and political as we ... open ourselves to the lively presences which make and disrupt our more-than-human social worlds and explore the politics and powers which infuse and define interactions. This diversity, fluidity, and creativity raises significant questions regarding how we approach the questions animals raise, what methods we employ to engage these issues, and how we write in a zoo-sensitive manner.[8]

9.4 Tentative Conclusions and Discussion

Deleuze and Guattari's concept of 'becoming (animal)' belongs to the same family of concepts as empathy, and its elaboration actually offers a pathway for developing empathy as a transdisciplinary and interpretive research methodology. I have argued that we are able to practice 'becoming' because of our empathetic capabilities and imagination as social animals, which gives rise to the exciting perspective that other (social) animals, being active agents in their own right who are part of our communities and share our livelihoods, could meet us halfway.

[7] www.genna.gender.uu.se/themes/animals/, accessed 21 April 2011.

[8] www.genna.gender.uu.se/themes/animals/events/zooethnographies/, accessed 21 April 2011. See also Bourke (2011).

In this paper I have argued that both ethologists and ethnographers use empathy as a research methodology, but rarely in an explicit way, and certainly do not develop empathy conceptually in this context. I have explored and argued that Deleuze and Guattari's (1980) concept of 'becoming (animal)' could offer us a way forward in developing empathy as a research methodology. This highly abstract exercise was, ironically, mainly informed by the hype in popular culture around all kinds of animal whisperers, who basically whisper 'empathy' and 'becoming' into the ears and minds of other social animals. A particular operationalization of this abstract exercise of becoming was offered by introducing the metaphor of 'tracking' for describing a possible transdisciplinary research process and its use of 'the senses' as tools, as methods to actually do the etho-ethnologies (cf 'doing ethnography', Geertz 1973).

This chapter should also be read as my first attempt to try and reflect on formulating a transdisciplinary research methodology, essentially trying to 'merge' the social and biological sciences. Other disciplines like psychology, literary criticism and philosophy are equally included, as are other knowledge producers, for example the various 'whisperers' from the domain of popular culture. It is a transdisciplinarity that lives up to the notion that:

> The non-specialist's representations of animals are no less 'right' and more 'popular' than those of the scientist because they are contextual and not objectified. It would be bold to claim that the ethologist knows dogs 'better' than the best dog-owners, or that he or she knows more about deer than the best hunters. (Lestel et al. 2006, p. 169)

This perspective echoes the observations made by Liebenberg about the San trackers' knowledge of animal behaviour as compared to that of ethologists (see note 3). Combining biology and the social sciences has been tried before, probably most famously and controversially by E.O. Wilson in his highly contested 1975 publication, entitled *Sociobiology. The new synthesis*.[9] But as far as I know, the more encompassing transdisciplinarity that I argue for here and its focus on synthesizing research methodologies along the lines of the concept of empathy has not been explored extensively before. Although I hope to have shown in this chapter that, conceptually at least, this approach could offer a fruitful pathway towards a transdisciplinary research methodology, I have not (yet) offered an operationalization of the methodology into concrete empirical methods that can be used in the field. Although I hint at a broader use of all the senses in fieldwork, in order to prevent an overly constraining reliance on (spoken and written) words and language, that avenue has to be explored in (far) more detail, to see how it could be included into a conceptual approach of 'empathy as becoming'.

I return once more to the key question of this volume 'What makes us moral?' The answer, following the argument developed in this chapter, would be: Certainly not the fact that we are human. Definitely not, in fact. It might instead be something we share and whisper together with other social animals: empathy.

[9] For an extensive critique of sociobiology from an anthropological perspective, see Sahlins (1977).

References

Anthony, L. 2010. *The elephant whisperer*. London: Pan Books.

Asquith, P. (ed.). 2002. *A Japanese view of nature. The world of living things by Kinji Imanishi*. London: RoutledgeCourzon.

Balcombe, J. 2010. *Second nature. The inner lives of animals*. New York: Palgrave MacMillan.

Bekoff, M. 2007. *The emotional lives of animals. A leading scientist explores animal joy, sorrow, and empathy – And why they matter*. Novato: New World Library.

Bekoff, M., and J. Pierce. 2009. *Wild justice. The moral lives of animals*. Chicago/London: University of Chicago Press.

Bourke, J. 2011. *What it means to be human. Reflections from 1791 to the present*. London: Virago Press.

Bradshaw, G.A. 2009a. *Elephants on the edge. What animals teach us about humanity*. New Haven/London: Yale University Press.

Bradshaw, G.A. 2009b. Transformation through service: Trans-species psychology and its implications for ecotherapy. In *Ecotherapy. Healing with nature in mind*, ed. L. Buzzell and C. Chalquist. San Francisco: Sierra Club Books.

Braidotti, R. 2006. Affirming the affirmative: on nomadic affectivity, *Rhizomes* 11/12, Fall 2005/Spring 2006. www.rhizomes.net/issue11/braidotti.html. Accessed 19 Apr 2011.

Calarco, M. 2008. *Zoographies. The question of the animal from Heidegger to Derrida*. New York: Columbia University Press.

Coetzee, J.M. 1999. *The lives of animals*. Princeton/Oxford: Princeton University Press.

De Waal, F. 1997. *Good natured: The origins of right and wrong in humans and other animals*. Cambridge, MA: Harvard University Press.

De Waal, F. 2001. *De aap en de sushimeester. Over cultuur bij dieren*. Antwerp/Amsterdam: Uitgeverij Contact.

De Waal, F. 2008. Putting the altruism back into altruism: The evolution of empathy. *Annual Review of Psychology* 59: 279–300. Antwerp/Amsterdam: Uitgeverij Contact.

De Waal, F. 2009. *The age of empathy. Nature's lessons for a kinder society*. New York: Harmony Books.

De Waal, F. 2010. *De aap uit de mouw*. Antwerp/Amsterdam: Uitgeverij Contact.

Deleuze, G., and F. Guattari. 1980. *A thousand plateaus. Capitalism and schizophrenia*. Minneapolis: University of Minnesota Press.

Derrida, J. 2008. I don't know why we are doing this. In *The animal that therefore I am*, ed. M. Marie-Louise. New York: Fordham University Press, pp. 141–160.

Dugatkin, L.A. 2006. *The altruism equation. Seven scientists search for the origins of goodness*. Princeton/Oxford: Princeton University Press.

Ellis, S. (with Junor, P.). 2009. *The man who lives with wolves.* New York: Harmony Books.

Fennell, J. 2001. *The dog listener. A noted expert tells you how to communicate with your dog for willing cooperation*. New York: Harper Collins.

Fox Keller, E. 1983. *A feeling for the organism. The life and work of Barbara McClintock*. New York/San Francisco: W.H. Freeman and Company.

Geertz, C. 1973. *The interpretation of cultures: Selected essays*. New York: Basic Books.

Griffin, D.R. 1978. Prospects for a cognitive ethology. *The Behavioral and Brain Sciences* 4: 527–538.

Hakonsson, J. 2003. *Exploring the phenomenon of empathy.* Dissertation, Department of Psychology, Stockholm University, Stockholm.

Haraway, D.J. 2008. *When species meet*. Minneapolis/London: University of Minnesota Press.

Janssens, M., and C. Steyeart. 2001. *Meerstemmigheid: Organiseren met verschil*. Leuven/Assen: Universitaire Pers Leuven/Van Gorcum.

Kamango, S. (with Bezuidenhout, L.). 2011. *Shadow in the sand. A koevoet tracker's story of an insurgency war.* Pinetown: 30° South Publishers.

Kawai, M.K. 1969. *Life of Japanese monkeys*. Publisher unknown.

Kirksey, S.E., and S. Helmreich. 2010. The emergence of multispecies ethnography. *Cultural Anthropology* 25(4): 545–576.

Kruuk, H. 2003. *Niko's nature. A life of Niko Tinbergen and his science of animal behavior*. Oxford: Oxford University Press.

LaFolette, H., and N. Shanks. 1996. The origin of speciesism. *Philosophy* 71: 41–61.

Lestel, D., F. Brunois, and F. Gaunet. 2006. Etho-ethnology and ethno-ethology. *Social Science Information* 45(2): 155–176.

Liebenberg, L. 1990a. *The art of tracking: The origin of science*. Cape Town: David Philip Publishers.

Liebenberg, L. 1990b. *A field guide to the animal tracks of southern Africa*. Cape Town: David Philip Publishers.

Lorenz, K. 2002 [1949]. *Man meets dog*. London/New York: Routledge.

Masson, J.M., and S. McCarthy. 1995. *When elephants weep. The emotional lives of animals*. New York: Bell Publishers.

Mehrabian, A. 1981. *Silent messages: Implicit communication of emotions and attitudes*, 2nd ed. Belmont: Wadsworth Publishing.

Montgomery, S. 2009 [2001]. *Walking with the great apes. Jane Goodall, Dian Fossey, Biruté Galdikas*. White River Junction: Chelsea Green Publishing.

Nagel, T. 1974. What is it like to be a bat? *Philosophical Review* 83(4): 435–450.

O'Connell, C. 2007. *The elephant's secret sense. The hidden life of the wild herds of Africa*. Chicago: University of Chicago Press.

Pachirat, T. 2011. *Every twelve seconds: Industrialized slaughter and the politics of sight*. New Haven: Yale University Press.

Pink, S. 2009. *Doing sensory ethnography*. Los Angeles/London: Sage Publications.

Richardson, K. 2010. *De leeuwenfluisteraar*. Amsterdam: De Boekerij.

Ridley, M. 1996. *The origins of virtue*. London: Viking.

Robert, S., and A. Degré. 1996. *Tippi of Africa. The little girl who talks to the animals*. Paris: Editions Michel Lafon.

Roberts, M. 2008. *The man who listens to horses: The story of a real-life horse whisperer*. London: Random House Publishing.

Roodnat, J. 2011. Schrijven over gezonde mannen geeft me kracht. *NRC Handelsblad*, January 7.

Ryan, M.J. 2010. An improbable path. In *Leaders in animal behavior. The second generation*, ed. L. Drickamer and D. Dewsbury, 465–496. Cambridge/New York: Cambridge University Press.

Safran Foer, J. 2009. *Dieren eten*. Amsterdam: Ambo/Athos uitgeverij.

Sahlins, M.D. 1977. *The use and abuse of biology: An anthropological critique of sociobiology*. Ann Arbor: University of Michigan Press.

Shanor, K., and J. Kanwal. 2009. *Bats sing mice giggle. Revealing the secret lives of animals*. London: Icon Books.

Skipper Jr., R.A. 2004. Perspectives on the animal mind. *Biology and Philosophy* 19: 483–487.

Swart, S. 2010. *Riding high. Horse, humans and history in South Africa*. Johannesburg: Wits University Press.

Ten Bos, R. 2008. *Het geniale dier. Een andere antropologie*. Amsterdam: Boom.

The Sunday Times. 2011. Jumbos just like us, June 5.

The Telegraph. 2008. The real-life Mowgli who grew up with Africa's wild animals, November 8.

Treadwell, T., and J. Palovak. 1997. *Among grizzlies. Living with wild bears in Alaska*. New York: The Ballantine Publishing Group.

Tuana, N. (ed.). 1989. *Feminism & science*. Bloomington: Indiana University Press.

Walker, C. 1996. *Signs of the wild. A field guide to the spoor & signs of the mammals of southern Africa*, 5th ed. Cape Town: Struik Publishers.

Westwood, W. 2008. *The animal gaze. Animal subjectivities in southern African narratives*. Johannesburg: Wits University Press.

Wilson, E.O. 1975. *Sociobiology. The new synthesis*. Cambridge, MA: Harvard University Press.

Yanow, D., and P. Schwartz-Shea (eds.). 2006. *Interpretation and method. Empirical research methods and the interpretive turn*. Armonk/London: M.E. Sharpe.

Ybema, S., and F. Kamsteeg. 2009. Making the familiar strange: A case for disengaged organizational ethnography. In *Organizational ethnography. Studying the complexities of everyday life*, ed. S. Ybema, D. Yanow, H. Wels, and F. Kamsteeg, 101–119. Los Angeles/London: Sage Publications.

Part III
Nativism and Non-Nativism

Chapter 10
The Origin of Moral Norms and the Role of Innate Dispositions

Jessy Giroux

10.1 Introduction[1]

When I hear the question ‘What makes us moral?’ I am reminded of a more ambitious question of which ‘what makes us moral?’ is probably a narrowed-down version. That more ambitious or perhaps simply broader question is ‘Where does morality come from?’ I am mainly interested in the latter question in this article, although my goal is to link that question to the former one. I intend to do this by investigating how ‘the things that make us moral’ – by which I mean our morally-relevant dispositions and capacities, such as empathy, vicarious distress, our natural propensity to cooperate and reciprocate, our ability to formulate and follow rules, etc. – bring societies to develop similar moral norms.

Let me begin by presenting two general ways of understanding where moral norms come from, how they originate or come about in human societies. First, one can argue that moral norms are no different from conventional norms in their origin. Like etiquette or sartorial norms, for instance, moral norms could be seen as the product of a variety of historical (geographical, religious, political, etc.) contingencies. On this minimalist account, asking why a society prohibits murder is no different from asking why it discourages eating spaghetti with one’s hands or wearing white socks with black shoes. Although there certainly is a difference in *degree* between moral and conventional infractions, there is no *qualitative* difference.

Second, one can describe moral norms as naturally emerging from human constitution. According to this model, moral norms, unlike spaghetti or sock-colour

[1] An ancestor of this article was published in 2011 (see Giroux 2011).

J. Giroux (✉)
PhD Student, Department of Philosophy, University of Toronto, Toronto, ON, Canada
e-mail: jessy.giroux@mailutoronto.ca

B. Musschenga and A. van Harskamp (eds.), *What Makes Us Moral? On the capacities and conditions for being moral*, Library of Ethics and Applied Philosophy 31,
DOI 10.1007/978-94-007-6343-2_10, © Springer Science+Business Media Dordrecht 2013

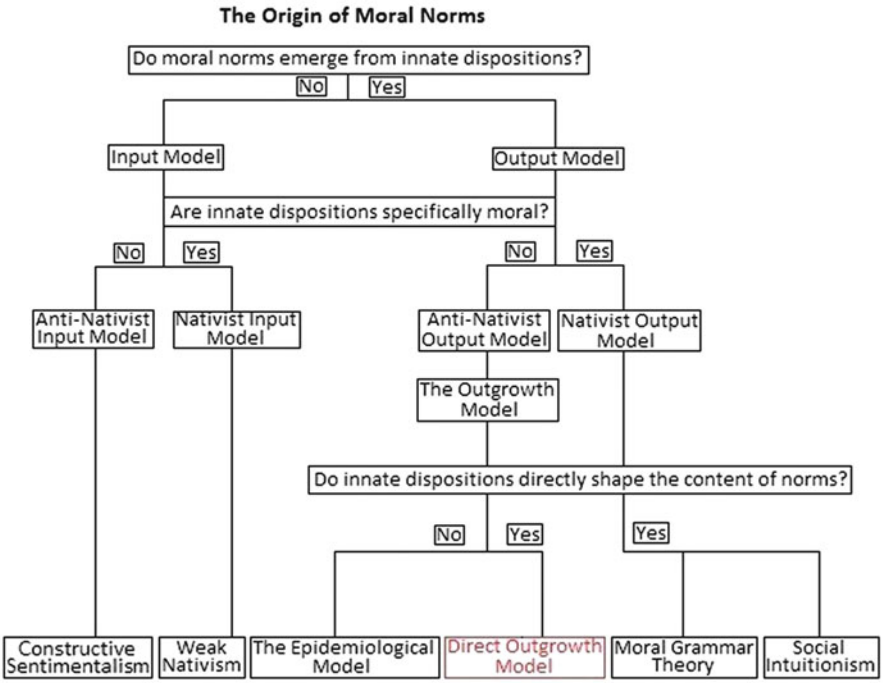
The Origin of Moral Norms
Do moral norms emerge from innate dispositions?
No
Yes
Input Model
Output Model
Are innate dispositions specifically moral?
No
Yes
No
Yes
Anti-Nativist Input Model
Nativist Input Model
Anti-Nativist Output Model
Nativist Output Model
The Outgrowth Model
Do innate dispositions directly shape the content of norms?
No
Yes
Yes
Constructive Sentimentalism
Weak Nativism
The Epidemiological Model
Direct Outgrowth Model
Moral Grammar Theory
Social Intuitionism

Fig. 10.1 Main questions and theoretical options in the debate on the origin of moral norms

norms, are not arbitrary social constructions but are rather the social extension of psychological traits shared by all human individuals. In this perspective, moral norms more closely resemble *prudential* norms, such as the (implicit) norm discouraging individuals from taking a walk in their underwear at zero degrees Fahrenheit. Such a norm is neither arbitrary nor contingent in that it derives from facts about human individuals, such as human beings' limited resistance to cold weather. One can also argue that human individuals would arrive naturally at such a norm even if it was not taught to them. For this reason, the norm is best described as extending from or emerging from 'human nature' rather than as being arbitrarily created by society and transmitted to individuals through socialization.

These two different kinds of explanations give us two general models for the origin of moral norms, which I will refer to as the Input and the Output models (Fig. 10.1).

In the first model, moral norms are seen as 'inputs' (from the individual's perspective) in that they are external entities transmitted to and assimilated by individuals. In this perspective, culture is the main purveyor of morality: were a given culture to provide radically different moral inputs, an individual from this culture would develop radically different moral judgments, emotions, and behaviours. Therefore, according to the Input model, society (S) creates moral norms (N) which are then assimilated by individuals (I):

Input Model: S → N → I

In the second model, the role of the independent and dependent variables is reversed. Moral norms are seen as 'outputs' (from the individual's perspective) in that they are seen as emerging from individuals – as instantiations of 'human nature' – and consequently imposing themselves on society. The dynamic is therefore a 'bottom-up' rather than a 'top-down' one:

Output Model: I → N → S

Of course, one can legitimately respond that these two models are in no way mutually exclusive. There is a trivial sense in which moral norms are always the product of innate dispositions and culture, of nature and nurture. For instance, a complex moral norm such as 'it is wrong to take money from the poor to give it to the rich' involves concepts that individuals can only acquire through experience. The emergence and transmission of such complex moral norms require not only a cultural context that provides concepts to individuals, but also innate capacities to formulate and grasp such concepts.

Although it is true that moral norms always require 'internal' and 'external' elements to some extent, the input/output dichotomy I am proposing is especially useful when it is applied to *basic* or *elementary* moral principles, which I conceive as the building blocks of more complex moral norms. There are reasons to believe that basic moral principles, such as the principle 'it is wrong to inflict pain on others',[2] are not merely *learned* by individuals but are truly constitutive of their nature – which would fit the Output model. This distinction between basic moral principles and complex moral norms will quickly become essential in the following discussion.

My goal in this article is to examine the Input and the Output models and determine which is the more plausible account of the origin of moral norms.[3] In doing so, I will discuss contemporary versions of each model, taking Jesse Prinz's 'Constructive Sentimentalism' as an example of the Input model (3), and Jonathan Haidt's 'Social Intuitionism' and the 'Moral Grammar' theory as examples of the Output model (4). I will then introduce my positive thesis, the 'Direct Outgrowth' model, which, although primarily an Output theory, captures important dimensions of Prinz's version of the Input model (5).

Before I can begin my presentation of the Input model, however, I need to set the scene by distinguishing three central and closely related debates.

[2] I will always presuppose a *pro tanto* – and perhaps *ceteris paribus* – clause behind any such basic moral principle.

[3] To provide a more complete picture of the different theoretical options, one should include the Realist model, according to which moral norms are or result from objective moral properties, often seen as 'supervening' on non-moral properties. I take an agnostic stance on the existence of moral properties in this article, and I will not discuss the Realist model, mainly because I wish to consider only properties whose existence is not controversial.

10.2 Origination, Nativism, and Universalism

The distinction I am proposing between an Input and an Output model may seem like a mere duplication of the more generally accepted distinction between ‘nativist’ and ‘anti-nativist’ theories. Indeed, both dichotomies answer the question of whether or not morality ‘comes from within’, and the new dichotomy I am proposing may seem to bring nothing new. It is therefore important for me to explain how my proposed dichotomy differs from the more common one.

The main differences can be found in the *specific object* addressed by each dichotomy as well as in the *scope* characteristic of each. While the nativism debate deals with whether or not human beings are ‘equipped’ with morality at birth, and therefore investigates the moral phenomenon at the individual level, the debate on the origin of moral norms discusses whether and/or to what extent moral norms found in societies historically emerge from innate human dispositions, and therefore investigates the moral phenomenon at the population level.

Of course, the position one endorses in the nativism debate very often determines the position that one endorses in the origination debate – and vice versa. One who believes that humans are morally constituted at birth will generally believe that this innate constitution is the root of the moral norms found in societies, while one who believes that morality is not innate will generally believe that moral norms emerge from a different source. However, even though Input and Output models often combine respectively with anti-nativist and nativist theories, different combinations of the two dichotomies are perfectly conceivable, and the theory I will defend in Sect. 10.5 is just such a different combination.

Another debate that should not be confused with the origination debate is the one on the *universality* of moral norms. This distinction is clearly less controversial than the previous one, but it is nevertheless important to understand the specific role that the universalism debate plays in its relation to the origination debate. Essentially, the position one defends in the universalism debate usually serves to provide evidence for the position defended in the other debate. One who highlights the apparent universality of moral norms will often do so in order to support the claim that moral norms have their root in universal human dispositions. Chandra Sekhar Sripada nicely illustrates the kind of data used by advocates of the Output model to provide evidence of their model:

> Most societies have rules that prohibit killing and physical assault (Brown 1991). Most societies have rules promoting sharing, reciprocating, and helping, at least under some circumstances (Cashdan 1989). Most societies have rules regulating sexual behavior among various members of society, especially among adolescents (Bourguignon and Greenbaum 1973), and most societies have rules promoting egalitarianism and social equality (Bohem 1999). (Sripada 2008, p. 322)

The claim of moral universality is also reinforced by experiments in moral psychology which indicate that individuals the world over share similar judgments and intuitions when faced with various moral dilemmas. For instance, Hauser et al. tested the now famous ‘trolley problems’ on people from different countries and found that people offer similar answers to different versions of the dilemmas.[4] Even the Hadza, 'a small and remote group of hunter-gatherers living in Tanzania, show similar patterns of responses'(Hauser et al. 2008, p. 135). The apparent universality in moral judgments can be interpreted as supporting the idea that moral norms derive from innate dispositions (Output model), as well as the idea that humans are born 'equipped with' morality (nativist position).

On the other side of the debate, the anthropological record is used to identify exotic tribes or ancient civilizations that promote(d) what we would consider to be highly immoral or barbaric practices. Such examples are usually cited in order to show that moral norms are much more diverse than what the Output model or the nativist position can account for. In response to such anthropological and historical counterexamples, advocates of Output or nativist theories will often distinguish between *foundational* and *content* moral universalism, arguing that counterexamples only serve to refute the latter type of universalism. Counterexamples disprove *content* universalism by showing that moral norms can differ significantly in content from one society to another. Yet, such counterexamples do not disprove the possibility of a 'foundational similarity' between norms: perhaps even the most antagonistic norms ultimately rest on the same fundamental moral principles. Diversity in moral norms could result from combining differently the same basic moral principles, depending on various historical factors. Jonathan Haidt and Craig Joseph defend a form of foundational universalism when they say:

> humans come equipped with an intuitive ethics.... These intuitions undergird the moral systems that cultures develop, including their understandings of virtues and character. By recognizing that cultures build incommensurable moralities on top of a foundation of shared intuitions, we can develop new approaches to moral education and to the moral conflicts that divide our diverse society. (Haidt and Joseph 2004, p. 56)

For the rest of this article, I will take for granted that a plausible account of the origin of moral norms needs to accommodate *both* a degree of moral universality and a degree of moral diversity. If one does not endorse 'foundational universalism', one needs to offer an alternative explanation of why individuals and societies share such moral similarities – as described not only in Sripada (2008) and Hauser et al. (2008), but in other influential works such as Brown (1991). Likewise, a plausible theory will be one that recognizes the existence of moral diversity and will not endorse 'content universalism'. It is my hope that these will not seem like arbitrary or unfounded conditions.

With these distinctions and clarifications established, I can now turn to a contemporary version of the Input model: Jesse Prinz's 'Constructive Sentimentalism'.

[4] First introduced by Philippa Foot (1978), the 'trolley problems' are thought experiments used to test individuals' intuitions when faced with specific moral dilemmas. Although there have been different versions of the initial problem, all testing different intuitions, Foot's original dilemma was formulated as follows: 'A trolley is running out of control down a track. In its path are five people who have been tied to the track by a mad philosopher. Fortunately, you could flip a switch, which will lead the trolley down a different track to safety. Unfortunately, there is a single person tied to that track. Should you flip the switch or do nothing?' (Foot 1978, p. 20).

10.3 The Input Model

Jesse Prinz is one of the most prominent contemporary advocates of anti-nativism. In *The emotional construction of morals* (2007) and in subsequent papers (Prinz 2008a, b, c, 2009), Prinz cogently argues that moral values are social constructions with an essential foundation in human emotions.[5] His main contention is that whatever innate dispositions humans may have, none of them is *specifically moral*. Even the human disposition for vicarious distress, which is arguably the most likely candidate for a specifically moral disposition, is interpreted by Prinz as serving a primarily non-moral function:

> Doesn't vicarious distress show that we have an innate predisposition to oppose harm? Perhaps, but it's not a moral predisposition. Consider the communicative value of a conspecific's scream. The distress of others alerts us to danger . . . . It's an indication that trouble is near. It's totally unsurprising, then, that we find it stressful. (Prinz 2008b, p. 374)

Prinz has a very specific understanding of the concepts 'specifically moral' and 'innate'. According to Prinz, for a psychological phenotype P to be specifically moral is for it to be innate in the following sense: P is innate if and only if it is 'acquired by means of psychological mechanisms that are dedicated to P, as opposed to psychological mechanisms that evolved for some other purpose or for no purpose at all' (Prinz 2008a, p. 370). There are therefore two elements in Prinz's notion of innateness: a faculty or phenotype is innate only if (1) it is subserved by dedicated machinery – which will often involve specialized modules – and only if (2) it is an evolved adaptation that was directly selected for its fitness-enhancing qualities. For the sake of simplicity and to avoid any confusion, I will always have this definition in mind when using the concept 'innate' in this article.

There is no doubt that Prinz rejects the nativist picture of morality. According to Prinz, our moral sentiments, as well as our capacity to formulate moral judgments, are mere 'by-products' or 'spandrels'[6] of other capacities: 'Morality . . . is a by-product of capacities that were not themselves evolved for the acquisition of moral rules' (Prinz 2007, p. 270). Since no disposition ever evolved for the purpose of morality, morality does not meet Prinz's second criterion for innateness. Nor does it meet Prinz's first criterion; on multiple occasions, Prinz rejects the claim that there are specialized moral modules or any mechanisms specifically dedicated to morality (see for instance section 3.3 in Prinz 2008a).

[5] Prinz mainly refers to moral values, but most of his argumentation can apply to moral norms as well.

[6] A spandrel is a characteristic or trait that is not a direct product of adaptive selection, but which is instead a by-product of some other characteristic or trait that was specifically selected. A related concept is that of 'exaptation', which refers more specifically to a 'shift of function' in the process of evolution – such as bird feathers which evolved for the purpose of weather regulation, but which were eventually 'co-opted' for the act of flying. For a better description of these concepts and how they apply to morality, see Fraser (2010).

Although there is no doubt that Prinz endorses moral *anti*-nativism, it should be noted that one cannot *infer* from Prinz's anti-nativism an endorsement of what I call the Input model. This is simply because a disposition like vicarious distress, even if it primarily serves a non-moral function, could still be the source of moral norms found in human societies. If one construes moral norms as the natural extension of innate dispositions, one truly endorses the Output model – even if it turns out that those dispositions do not serve a primarily moral purpose. Parts of Prinz's argumentation seem to support such a view of the origin of moral norms: 'Natural selection has probably furnished us with a variety of behavioral and affective dispositions that contribute to the emergence of moral values .... There is a trivial sense in which every norm we have owes something to our biological makeup' (Prinz 2007, p. 255).

There is ample evidence however that Prinz really endorses the Input model. That is because, according to him, none of the innate dispositions favouring morality can ultimately outweigh the 'process of enculturation' (Prinz 2007, p. 257) or socialization. Culture is thus the real force to be reckoned with, and there is arguably no limit to what a society can come to endorse as a rule of conduct for its members. What is condemned in society A can very well be revered in society B:

> I tend to think, somewhat cynically, that the range of moral rules is relatively unconstrained .... I adamantly believe that we could teach people to value recreational torture of babies .... I'm sure a search of the anthropological record would uncover groups that tortured babies for fun – especially if the babies belonged to enemy groups that were defeated in battle. (Prinz 2008c, p. 429)

Despite his failure at finding groups that torture babies for fun, Prinz does provide examples of remote tribes that perpetuate shocking traditions, such as the Llongot tribe in the Philippines whose coming-of-age ritual for boys consists in bringing back the head of an innocent member of a neighbouring tribe (Rosaldo 1980, from Prinz 2008b, p. 373). Such examples clearly serve the purpose of showing that whatever innate dispositions human beings may have, none of them is strong enough to outweigh the pressures of socialization. And precisely because innate dispositions are *weak* or *non-pervasive* in that way, they cannot accurately be described as the source of moral norms. Therefore, the Output model is not the right model of the origin of moral norms.

However, as was discussed in Sect. 10.2, anthropological counterexamples cannot provide sufficient evidence against the Output model, because such counterexamples do not disprove the theory of 'foundational universalism', i.e. the claim that even radically different moral norms rest on a foundation of universally shared basic moral principles. For example, one could argue that the Llongots recognize the *pro tanto* wrongness of murder, but that this *pro tanto* wrongness is outweighed by metaphysical beliefs held by the tribe, such as the belief that neighbouring tribes are evil or somehow inhuman, or that such murders are necessary for the tribe's survival. Such beliefs could be seen by them as legitimizing an otherwise condemnable act. An advocate of the Output model could go on to argue that the recognition of the *pro tanto* wrongness of murder is a 'natural extension' of human dispositions rather than a social construction, and that the case of the Llongots only serves to illustrate that the same basic moral principles can find different expressions in different cultural contexts. Therefore, in order to provide a satisfactory defence of the Input model, Prinz would need to explain why basic moral principles, such as the *pro tanto* wrongness of murder, enjoy such universality.

Such an explanation is provided by Prinz. When considering the apparent universality of basic moral principles, he offers an explanation inspired in large part by game theory:

> There are some social pressures that all human beings face. In living together, we need to devise rules of conduct . . . . Cultures need to make sure that people feel badly about harming members of the in-group and taking possessions from their neighbors . . . . This is a universal problem, and given our psychological capacities (for emotion, reciprocation, mental state attribution, etc.), there is also a universal solution. (Prinz 2008c, p. 405)

According to Prinz, this ‘universal solution’ to this ‘universal problem’ is the main force that constrains the otherwise limitless range of moral norms. It is because there is a universal solution to a universal problem of coordination that there is such universality in basic moral principles. The fact that similar principles are held cross-culturally is therefore not so much the result of our sharing similar dispositions as it is the result of our facing similar problems and coming up with similar solutions.

Prinz’s argument about the importance of coordination pressures in understanding the origin of moral norms is in itself rather uncontroversial. Hardly anyone will deny that groups have their own dynamics, from which certain imperatives naturally arise, and that these dynamics can play a role in the emergence and prevalence of moral norms in human societies. The problem with Prinz’s thesis however is that he implies that coordination pressures are *sufficient* for explaining the universality of basic moral principles. It is not at all clear that coordination pressures *alone* can account for the extent of moral similarities across cultures. There are also good reasons to doubt that innate dispositions are as ‘weak’ as Prinz thinks they are, or that socialization and conditioning mechanisms are such irresistible forces. As will be discussed in the next section, many dispositions and ‘prepared emotional reactions’ actually seem quite robust and can become serious obstacles to unusual socialization projects.

If it turns out that innate dispositions are indeed robust, they may very well be an essential factor, alongside other factors such as coordination pressures, in explaining the origin of moral norms. What remains to be determined however is the *specific nature* of the relevant dispositions as well as the *specific role* that they play in bringing about moral norms. Different conceptions of the nature and role of dispositions will give us different versions of the Output model.

10.4 The Output Model

The Output model can take many different forms, and the innate dispositions construed as the sources of moral norms can range from a general ‘learning preparedness’ to a full-fledged moral sense. I will focus here on two contemporary theories which distinguish themselves in the amount of attention they receive from researchers. It should be noted that, just as with Jesse Prinz's Constructive Sentimentalism, the two theories I will discuss here are primarily associated with the *nativism* debate, and their current application to the *origination* debate involves an element of interpretation on my part. It is perfectly possible therefore that some advocates of each theory would be more sympathetic to the Input model, and I will briefly discuss in Sect. 10.5 how such a nativist version of the Input model is conceivable.

The two theories I wish to discuss are the 'Moral Grammar' theory, whose main advocates are Marc Hauser (2006, 2008), Susan Dwyer (2008), and John Mikhail (2008); and the theory of 'Social Intuitionism', developed and mainly defended by Jonathan Haidt (Haidt and Joseph 2004; Haidt and Bjorklund 2008). The two theories differ significantly in their understanding of the *nature* of the relevant innate dispositions, and I will focus in this section on their diverse conceptions of the nature of innate dispositions. In the next section, I will address the question of the *role* that innate dispositions play, according to the Output model, in bringing about moral norms in human societies.

I begin with the Moral Grammar theory (MG). Often referred to as the 'linguistic analogy', because of its roots in Noam Chomsky's linguistic model, MG describes the relevant innate disposition as a specialized moral faculty or competence. This faculty or competence is construed as an innate 'grammar', i.e. a set of abstract, general principles which unconsciously guide the individual's interpretation of social phenomena and facilitate the acquisition of a moral 'language' or system. This is certainly an unusual way of conceiving moral dispositions, and it is therefore important to understand how advocates of MG arrive at such a conception.

First, they note that a rule or principle such as 'it is wrong to inflict pain on others' appears to be 'endorsed' remarkably early by young children, long before socialization is able to leave its full imprint on them.[7] This is just one instance of the general phenomenon of the *precociousness* of human morality, of which John Mikhail offers an interesting overview which is worth presenting at length:

> Three- and four-year-old children use intent or purpose to distinguish two acts with same result (Baird 2001). They also distinguish 'genuine' moral violations (e.g. theft, battery) from violations of social conventions (e.g. wearing pajamas to school; Smetana 1983; Turiel 1983). Four- and five-year-olds use a proportionality principle to determine the appropriate level of punishment for principals and accessories (Finkel et al. 1997). Five-year-olds display a nuanced understanding of negligence and restitution (Shultz et al. 1986). One man shoots and kills his victim on the mistaken belief that he is aiming at a tree stump. A second man shoots and kills his victim on the mistaken belief that killing is not wrong. Five- and six-year-olds distinguish cases like these in conformity with the distinction between mistake of law and mistake of fact, recognizing that false actual beliefs may exculpate, but false moral beliefs do not (Chandler et al. 2000). Five- and six-year-olds also calibrate the level of punishment they assign to harmful acts on the basis of mitigating factors, such as provocation, necessity, and public duty (Darley et al. 1978). Six- and seven-years-olds exhibit a keen sense of procedural fairness, reacting negatively when punishment is inflicted without affording the parties notice and the right to be heard (Gold et al. 1984). (Mikhail 2008, p. 354)

[7] If not intellectually, at least *practically*, i.e. as reflected by their behaviours and actions.

And the list goes on. Using this list of claimed premature moral 'knowledge' as evidence, Mikhail, like other advocates of MG, argues that children must be born with a set of moral 'principles and parameters'. Without such innate principles and parameters, children would develop a moral system only through the stimuli received from their environment. The problem, argue MG advocates, is that 'moral stimuli' are so *poor* that they could hardly account for the level of moral knowledge possessed by children. *Ergo*, there has to be something resembling an innate moral 'grammar' guiding their experience – in the same way that an innate grammar is said to guide the acquisition of a natural language in the Chomskian linguistic paradigm.

The problem with this line of argument, however, as expressed by multiple authors, is that the moral stimulus appears 'poor' only if it is conceived entirely in terms of rational rules or principles that are expressed by parents or 'imbibed' by children through experience. But, as Kim Sterelny notes, 'children get more than verbal feedback. Audiences respond [to moral infractions] with emotional changes, and humans respond emotionally to their very own actions and to the effects of those actions on others' (Sterelny 2010, p. 293). Once emotional responses are considered to be moral stimuli, the claim regarding their poverty suddenly appears quite . . . poor.

The point here is not so much to reject the claim of moral precociousness in children as to reject MG's interpretation of the phenomenon. The fact that children exhibit moral behaviour and have a certain understanding of morality at a young age is generally recognized. For instance, Shaun Nichols, who is not himself an advocate of MG, comes to similar conclusions after reviewing a series of studies, noting that 'children have a strikingly early grasp of core moral judgment' (Nichols 2008, p. 261). He observes that some moral 'facts' grasped by children, such as the moral/conventional distinction, do seem quite precocious: 'These findings on the moral/conventional distinction are neither fragile nor superficial. They have been replicated numerous times using a wide variety of stimuli.' The point therefore is not to deny the phenomenon of 'early morality' but to understand it in light of the central role played by *emotions.*

One way of understanding the role of emotions is in terms of emotional *reinforcement*. This is the strategy adopted by Jesse Prinz to account for the phenomenon of early morality. Reinforcement strategies such as 'love-withdrawal', 'power assertion', and 'induction of empathic distress' (Prinz 2008b, p. 431) are all used by parents and other moral educators to bring children to assimilate their society's norms. The advantage of such an explanation is of course its great parsimony. The theory requires nothing more than a general responsiveness of children to conditioning to account for their early moral behaviours and judgments. The disadvantage of the explanation, however, is that it presupposes something that greatly resembles 'equipotentiality', a theory that is largely discredited nowadays in psychology: 'Garcia and Koelling (1966) demonstrated that equipotentiality – the equal ability of any response to get hooked up at any stimulus – was simply not true. It is now universally accepted in psychology that some things are easy to learn (e.g. fearing snakes), while others (fearing flowers or hating fairness) are difficult or impossible' (Haidt and Bjorklund 2008, p. 183).

The implication of a rejection of equipotentiality is that children are not as malleable as is sometimes claimed, and they will often resist attempts at emotional reinforcement:

> Children routinely resist parental efforts to get them to care about, value, or desire things. It is just not very easy to shape children, unless one is going with the flow of what they already like. It takes little or no work to get 8-year-old children to prefer candy to broccoli, to prefer being liked by their peers to being approved of by adults, or to prefer hitting back to loving their enemies. Socializing the reverse preferences would be difficult or impossible. (Haidt and Bjorklund 2008, p. 201)

This leads to a second way of understanding the role of emotions to account for the phenomenon of early morality, which is developed by Jonathan Haidt in the theory of Social Intuitionism. In addition to emotional reinforcement, one needs to take into consideration the emotional *preparedness* of children. According to Haidt, children are born equipped with 'an innate preparedness to feel flashes of approval or disapproval toward certain patterns of events involving other human beings' (Haidt and Joseph 2004, p. 56). More specifically, Haidt argues that these prepared emotional reactions correspond to five basic moral domains: harm/care, fairness/reciprocity, in-group/loyalty, authority/respect, and purity/sanctity. He essentially arrives at such a conclusion by reviewing important works detailing, among other things, commonalities in moral judgments across cultures as well as animal precursors of morality. According to Haidt, these five moral domains are the best way of capturing what is found in the literature, and he postulates the existence of a specialized module for each of these domains.

It is not clear however how Haidt arrives at the conclusion that one should see these five fundamental moral domains as being encoded in the human mind. The simple fact that we can successfully subsume moral phenomena into five categories is no proof that these categories are present in the brain at birth. As Ron Mallon puts it, referring to 'principles' rather than 'domains': 'The mere fact that we can describe principles that seem to capture intuitions about a set of moral cases gives us exactly no reason at all to think that those principles are themselves implemented directly in a computationally discrete way or by a computationally discrete faculty' (Mallon 2008, p. 151). The domains described by Haidt may very well turn out to be universal and the source of moral norms found cross-culturally, but these facts alone can hardly be seen as evidence, let alone proof, of the existence of five specialized modules.

It should be noted that the same criticism applies to many advocates of the Moral Grammar theory as well. The main weakness of theories that postulate the existence of specialized moral modules, especially in the Fodorian tradition, is that they are cognitively costly, which renders them dubious in the eyes of their anti-nativist critics. Why should one adopt a modular picture of morality when a less costly alternative is available? The answer offered by many nativists is that only this kind of nativist framework can efficiently account for the universality of basic moral norms and judgments.[8]

My personal contention, shared by anti-nativists, is that one does not need such a nativist framework to account for moral universality. Unlike most anti-nativists, however, I share the nativist's essential intuition, which is that moral universality speaks in favour of a conception of morality as emerging *from within* human individuals. I do not believe however that specialized moral modules or faculties are necessary to explain how morality comes from within, and for this reason I have introduced a distinction between nativist theories and the Output model. The Output model can provide an explanation of moral norms as being rooted in human constitution even in the absence of specialized modules or faculties. It is such an anti-nativist version of the Output model that I now wish to defend.

10.5 The Outgrowth Model, or Output Non-nativism

I have not yet provided a full picture of the different theoretical options in the debate on the origin of moral norms. Thus far, I have only presented one version of the Input and the Output model, namely, the anti-nativist version of the Input model and the nativist version of the Output model. Two theoretical options remain to be addressed.

First, one could defend a nativist version of the Input model. A theory of this type would essentially argue that humans are morally constituted at birth, *but* that moral dispositions are highly malleable and hold very little if any weight in shaping the content of a society's norms. One could call such a theory 'Weak Nativism', where 'weak' refers to the non-pervasive nature of innate moral dispositions. Because moral dispositions are non-pervasive, they cannot accurately be described as the source of moral norms found in human societies. Whether moral norms are seen as resulting mainly from contingent factors (geographical, political, religious, etc.), or from necessary ones (such as Prinz's coordination pressures), the relevant factors will be *external* to the individual's moral constitution.

Second, one could defend an anti-nativist version of the Output model – which is the kind of theory I endorse. The main idea is that moral norms have an essential source in innate dispositions but that none of these dispositions is specifically moral in nature. Not being specifically moral implies that these dispositions did not evolve for the purpose of morality and are not subserved by dedicated machinery – such as specialized modules. Yet, despite their not being specifically moral in such a

[8] Of course, not all moral nativists are committed to the existence of specialized moral modules. One can distinguish between 'strong' and 'weak' kinds of nativism, noting that only the former kind is committed to modules. Jesse Prinz offers a similar distinction between three kinds of nativism which he labels 'immodest', 'modest', and 'minimal' nativisms (Prinz 2009, p. 168). I am focusing here on 'strong nativism' simply because it is the type endorsed by most advocates of the Moral Grammar theory and Social Intuitionism.

Table 10.1 Four theoretical options for the origin of moral norms

	Nativism	Anti-nativism
Input model	Weak nativism	Constructive sentimentalism
Output model	Moral grammar theory; social intuitionism	The outgrowth model

way, these various dispositions are nevertheless the main cause of the development of moral norms in human societies. Morality should therefore be construed as an 'outgrowth' of those dispositions (Table 10.1).[9]

Moral norms can be seen as outgrowths of innate dispositions in at least two different ways, which I will refer to as the 'direct' and the 'indirect' versions of the Outgrowth model.

According to the *indirect* version of the Outgrowth model, moral norms grow indirectly out of innate dispositions in the sense that innate dispositions indirectly shape the content of norms. To indirectly shape norms is to play a *constraining* role in the determination of norms; when a norm is 'introduced' in a given society, innate dispositions are the main factor that determines whether or not the 'candidate norm' will successfully impose itself, i.e. be adopted by society and persist through time. Something like a Darwinian mechanism is in play which allows for only the *fittest* candidate norms to pass the test of time – and for a candidate norm to be *fit* it must be compatible with humans' natural constitution.

To illustrate this idea, one can take the example of cooperation norms. Imagine a society that would try to promote cheating, betrayal, and free-riding as moral ideals. Such a project would certainly be short-lived, and one obvious reason is the one raised by Jesse Prinz regarding coordination pressures: a society or any collective enterprise will sooner or later collapse if its members are unable to cooperate and trust each other. However, coordination pressures are only one part of the explanation. Indeed, if humans were simply unreceptive to cooperation imperatives and were inclined only to cheat, betray, and free-ride, cooperation norms could not successfully impose themselves, and society would simply collapse.[10] Imagine for instance a society composed entirely of psychopaths. Given psychopaths' natural selfishness and lack of empathy,[11] one can doubt that they would be able to

[9] It should be remembered that this kind of theory is described as 'anti-nativist' only insofar as it rests on Prinz's specific use of the concept 'innate'. If one were to adopt a different definition of the concept – for instance, if one were to say that a faculty can be called 'innate' even if it is a by-product or spandrel of other faculties – one could very well consider this kind of theory to be 'nativist'. This is why, using a different definition of 'innate', I presented this theory in a different article (Giroux 2011) as a form of 'moderate nativism' rather than as a form of anti-nativism.

[10] One could argue that social institutions could still be preserved if individuals were strongly constrained by external forces, such as in a police state. However, in this scenario, the individuals in charge of enforcing cooperation would themselves only be serving their personal interest. Therefore, in the absence of a genuine capacity for cooperation, society can only rest on very shaky grounds.

[11] See Blair et al. (2005) for a thorough description of psychopaths' unusual constitution.

truly assimilate cooperation norms; as a result, their society's institutions would eventually collapse, leaving them with something resembling a Hobbesian state of nature. This is but one example of how human dispositions render certain norms very likely, while others are rendered highly improbable, if not impossible.

Theorists who construe innate dispositions as 'indirectly shaping' the content of moral norms in this way are usually inspired by Dan Sperber's 'epidemiology of representations' (Sperber and Hirschfeld 2004). The main philosophers who adopt the 'Epidemiological model' are Shaun Nichols (2004, 2008), Chandra Sekhar Sripada (2008), and Steven Stich (2006).[12] Nichols defends a version of the Epidemiological model called 'affective resonance', which focuses on the constraining role of *emotional* dispositions: 'The affective resonance hypothesis predicts that, *ceteris paribus*, norms that prohibit actions that are independently likely to excite negative emotions should be more likely to survive than norms that are not connected to emotions' (Nichols 2008, p. 270). Sripada incorporates a wider range of dispositions, which he calls 'Sperberian biases': 'When their effects are summated over populations and over time, they generate a fairly strong population-level force which can have the effect of changing the distribution of norms in the direction favored by the Sperberian bias' (Sripada 2008, p. 333).

The idea that innate dispositions constrain the range of possible moral norms is a truly elegant explanation of why one finds so many similarities in the moral norms adopted by otherwise very different societies. I do not believe, however, that this indirect version of the Outgrowth model provides a complete account of the origin of moral norms. My claim is that innate dispositions play an even stronger role in shaping the content of moral norms: they provide the elementary moral principles, or 'building blocks' used by all societies in the creation of more complex moral norms. Because they actually *provide moral content*, as opposed to merely imposing general constraints, innate dispositions should be seen as *directly* shaping the content of moral norms.

According to the direct version of the Outgrowth model, innate dispositions directly shape the content of moral norms by helping every 'normally constituted' human individual to develop naturally the same basic moral principles.[13] Those basic principles provide the general structure on which human societies develop moral norms. Societies will diverge by giving more importance to certain basic principles rather than others, and by identifying different criteria for their application, but they will still incorporate a similar set of basic moral principles. Only exceptionally strong factors, such as extreme metaphysical beliefs, could potentially lead societies

[12] Jesse Prinz also defends a version of the epidemiological model, arguing that 'cultural transmission is a function of fitness' (Prinz 2007, p. 220). However, as was described in Sect. 10.3, Prinz does not assign a real constraining role to innate dispositions, and for that reason I did not include him as an advocate of the 'indirect Outgrowth' model: 'Biologically based behaviors are not quite a constraint on the genealogy of moral rules, because culture can override them, but they are often a central ingredient' (Prinz 2007, p. 274).

[13] Again, with the exception of individuals such as psychopaths.

to not incorporate some of the basic moral principles.[14] This view amounts to an endorsement of what was dubbed 'foundational universalism' in Sect. 10.2.

At this point, one may legitimately ask what those basic moral principles actually are. Since the present article only aims at clarifying the main theoretical options in the debate on the origin of moral norms, this is certainly not the place to defend an exhaustive list of basic moral principles. However, as a general indicator of what I have in mind, one can refer to W.D. Ross's list of *prima facie* moral duties, which includes principles of fidelity, reparation, gratitude, non-maleficence, justice, beneficence, and self-improvement (Ross 1930). I believe that individuals are naturally led to develop similar basic principles for all or most of Ross's domains. My specific contention is that individuals are neither 'born with' those principles nor do they merely 'internalize' them as a result of socialization. Rather, such principles naturally derive from human predispositions, even in the absence of specialized moral modules.

Putting aside pressures of socialization, at least two factors can be seen as leading individuals to develop similar basic moral principles. The first factor is what Shaun Nichols calls 'natural elicitors': certain events naturally elicit emotional reactions in individuals independently of culture. A good example is vicarious distress: 'even newborn infants respond aversely to some cues of suffering (e.g. Simner 1971) . . . . There is good reason to suppose that the emotional response to suffering in others is universal and innately specified' (Nichols 2008, p. 271). Innately prepared vicarious distress, even if it did not evolve for the purpose of morality, could lead individuals to develop intuitions about the 'negativity' of suffering in others. Other 'natural elicitors' could have a similar impact.

Second, humans naturally project on others the many rules and principles that they come to perceive as applying to themselves: being treated unfairly makes me feel outraged, therefore it is something that will make others (who are like me) feel outraged, and it is therefore something I have a *pro tanto* reason to avoid. Replace 'being treated unfairly' with 'being lied to' or 'having a promise made to me broken', and individuals can be brought to see that if it is true for themselves, it will usually be true for others, for the simple reason that others are *like them.* This is what Erik J. Wielenberg calls the 'likeness principle': 'If I believe that I am a bearer of certain moral barriers and that others are similar to me with respect to their known properties, I am disposed to form the belief that those others possess similar moral barriers' (Wielenberg 2010, p. 446). Steven Pinker defends more or less the same principle: 'No creature equipped with the circuitry to understand that it is immoral for you to hurt me could discover anything but that it is immoral for me to hurt you' (Pinker 2002, p. 193).

[14] It is quite hard to find actual examples of societies that do not incorporate the basic moral principles. An example that may come to mind is the practice of human sacrifice that was condoned by certain ancient religious traditions. Perhaps these killings were viewed by some societies as *wholly positive* (because they pleased the gods), but it is more likely that they were viewed as *necessary evils* that retained an element of *prima facie* wrongness which was simply outweighed by a greater good. For an example to be truly accurate here, it would have to show that a society does not even consider killing (or lying, not keeping a promise, etc.) as a *wrong-making feature* of an act.

The existence of natural elicitors and the likeness principle are only two examples of how individuals can develop similar basic moral principles even in the absence of moral modules. Of course, if it turns out that such modules actually exist, there will be an even stronger case for viewing dispositions as *directly* shaping norms.[15] My goal in this section has been simply to show that moral nativism, understood in a *Prinzian* fashion, is not a prerequisite for construing moral norms as emerging from within individuals.

10.6 Conclusion

With all the pieces in place, we have a preliminary picture of what could be a new theory of the origin of moral norms. According to the Direct Outgrowth model that I have presented, moral norms should be seen as the natural extension of human dispositions rather than as social constructions. In this perspective, moral norms are directly shaped by humans' emotional preparedness and by a variety of other innate dispositions. These innate dispositions are not specifically moral, which means that they did not evolve for the purpose of morality and that they are not subserved by dedicated machinery. Still, they lead every 'normally constituted' human individual to develop naturally certain basic moral principles, such as the seven principles described by W.D. Ross. These basic moral principles constitute the normative backbone of every human society, and moral diversity should be understood as the assignment of different weights to the same basic principles.

Of course, this model can and should incorporate other important factors outlined by other models. For instance, one should see 'coordination pressures' as emphasizing certain imperatives which are essential to any collective enterprise, and one should recognize the role of emotional reinforcement, alongside emotional preparedness, as a key factor accounting for the phenomenon of early morality. With the inclusion of these different elements, one arrives at a potentially more complete picture of the origin of moral norms.

A lot more needs to be said in defence of this model, however, before it can claim to be more plausible than alternative models. The model's main challenges will be to demonstrate that innate dispositions are as robust as it claims, and that basic moral principles that individuals naturally develop are truly pervasive. Those are challenges that I hope to address in the future.

Acknowledgments I am very grateful to Christine Tappolet, Benoît Dubreuil, Bruno Guindon, Jean-François Grégoire, and Joseph Owens for their helpful comments on this article. I also wish to thank participants at the international conference "What Makes Us Moral?" in Amsterdam, Netherlands, on June 23–24, 2011, for their comments on my presentation. I am especially grateful to Carel van Schaik and Carsten Nielsen for their comments.

[15] Output nativists generally seem to endorse the view that innate dispositions *directly* shape the content of a society's norms (Haidt and Joseph 2004, p. 56; Dwyer 2008, p. 414).

References

Baird, J.A. 2001. Motivations and morality: Do children use mental state information to evaluate identical actions differently? Paper presented to the biennial meeting of the Society for research in child development, Minneapolis.

Blair, J., D. Mitchel, and K. Blair. 2005. *The psychopath: Emotion and the brain*. Malden: Wiley-Blackwell.

Bohem, C. 1999. *Hierarchy in the forest*. Cambridge: Harvard University Press.

Bourguignon, E., and L. Greenbaum. 1973. *Diversity and homogeneity in world societies*. New Haven: HRAF Press.

Brown, D. 1991. *Human universals*. New York: McGraw-Hill.

Cashdan, E. 1989. Hunters and gatherers: Economic behavior in bands. In *Economic anthropology*, ed. S. Plattner. Palo Alto: Stanford University Press.

Chandler, M., B. Sokol, and C. Wainryb. 2000. Beliefs about truth and beliefs about rightness. *Child Development* 71(1): 91–97.

Darley, J., E. Klosson, and M. Zanna. 1978. Intentions and their contexts in the moral judgments of children and adults. *Child Development* 49: 66–74.

Dwyer, S. 2008. How not to argue that morality isn't innate: Comments on Prinz. In *Moral psychology*, vol. 2, ed. W. Sinnott-Armstrong, 407–418. Boston: MIT Press.

Finkel, N., M. Liss, and V. Moran. 1997. Equal or proportionate justice for accessories? Children's pearls of proportionate wisdom. *Journal of Applied Developmental Psychology* 18: 229–244.

Foot, P. 1978. *Virtues and vices and other essays in moral philosophy*. Berkeley/Los Angeles: University of California Press.

Fraser, B.J. 2010. Adaptation, exaptation, by-products, and spandrels in evolutionary explanations of morality. *Biological Theory* 5(3): 223–227.

Garcia, J., and R. Koelling. 1966. Relation of cue to consequence in avoidance learning. *Psychonomic Science* 4: 123–124.

Giroux, J. 2011. The origin of moral norms: A moderate nativist account. *Dialogue: Canadian Philosophical Review* 50(2): 381–406.

Gold, L., J. Darley, J. Hilton, and M. Zanna. 1984. Children's perceptions of procedural justice. *Child Development* 55: 1752–1759.

Haidt, J., and F. Bjorklund. 2008. Social intuitionists answer six questions about moral psychology. In *Moral psychology*, vol. 2, ed. W. Sinnott-Armstrong, 181–217. Boston: MIT Press.

Haidt, J., and C. Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus: On Human Nature* 133: 55–66.

Hauser, M.D. 2006. *Moral minds: How nature designed our universal sense of right and wrong*. New York: Ecco Press.

Hauser, M.D., L. Young, and F. Cushman. 2008. Reviving Rawls's linguistic analogy: Operative principles and the causal structure of moral actions. In *Moral psychology*, vol. 2, ed. W. Sinnott-Armstrong, 107–143. Boston: MIT Press.

Mallon, R. 2008. Reviving Rawls's linguistic analogy inside and out. In *Moral psychology*, vol. 2, ed. W. Sinnott-Armstrong, 145–155. Boston: MIT Press.

Mikhail, J. 2008. The poverty of the moral stimulus. In *Moral psychology*, vol. 1, ed. W. Sinnott-Armstrong, 353–359. Boston: MIT Press.

Nichols, S. 2004. *Sentimental rules: On the natural foundations of moral judgment*. Oxford/New York: Oxford University Press.

Nichols, S. 2008. Sentimentalism naturalized. In *Moral psychology*, vol. 2, ed. W. Sinnott-Armstrong, 255–274. Boston: MIT Press.

Pinker, S. 2002. *The blank slate: The modern denial of human nature*. New York: Viking Penguin.

Prinz, J.J. 2007. *The emotional construction of morals*. Oxford: Oxford University Press.

Prinz, J.J. 2008a. Is morality innate? In *Moral psychology*, vol. 1, ed. W. Sinnott-Armstrong, 367–406. Boston: MIT Press.

Prinz, J.J. 2008b. Reply to Dwyer and Tiberius. In *Moral psychology*, vol. 1, ed. W. Sinnott-Armstrong, 427–439. Boston: MIT Press.

Prinz, J.J. 2008c. Resisting the linguistic analogy: A commentary on Hauser, Young, and Cushman. In *Moral psychology*, vol. 2, ed. W. Sinnott-Armstrong, 157–170. Boston: MIT Press.

Prinz, J.J. 2009. Against moral nativism. In *Stich and his critics*, ed. M. Bishop and D. Murphy, 167–189. Malden: Miley-Blackwell.

Ross, W.D. 1930. *The right and the good*. Oxford: Oxford University Press.

Shultz, T., K. Wright, and M. Schleifer. 1986. Assignment of moral responsibility and punishment. *Child Development* 57: 177–184.

Simner, M. 1971. Newborn's response to the cry of another infant. *Developmental Psychology* 5: 136–150.

Smetana, J.G. 1983. Social cognitive development: Domain distinctions and coordinations. *Development Review* 52: 1333–1336.

Sperber, D., and L.A. Hirschfeld. 2004. The cognitive foundation of cultural stability and diversity. *Trends in Cognitive Sciences* 8: 40–46.

Sripada-Sekhar, C. 2008. Nativism and moral psychology: Three models of the innate structure that shapes the contents of moral norms. In *Moral psychology*, vol. 2, ed. W. Sinnott-Armstrong, 319–343. Boston: MIT Press.

Sterelny, K. 2010. Moral nativism: A skeptical response. *Mind and Language* 25(3): 279–297.

Stich, S. 2006. Is morality an elegant machine or a kludge? *Journal of Cognition and Culture* 6: 181–189.

Turiel, E. 1983. *The development of social knowledge: Morality and convention*. New York: Cambridge University Press.

Wielenberg, E.T. 2010. On the evolutionary debunking of morality. *Ethics* 120(3): 441–464.

Chapter 11
It's Complicated – Moral Nativism, Moral Input and Moral Development

Carsten Fogh Nielsen

11.1 Introduction

What makes us moral? This question has at least two distinct meanings. On the one hand emphasis can be put on the 'what'-part: the question of *what* makes us moral? On this interpretation the question means something like the following: What are the (necessary and sufficient) components of moral competence? What are the constituent elements of moral agency? What are the (cognitive and affective) capacities and competences which human beings possess that makes us capable of moral agency?

This however is not the only possible way to frame the question. Instead of focusing on the 'what', we can focus on the 'make'; the question of what *makes* us moral? Framed in this way the question's primary concern is not the problem of what the necessary and/or sufficient ingredients of moral competence are. The main concern is rather the question of moral development; the question of how we human beings develop and acquire the capacities and competences that enable us to navigate and orientate ourselves within the moral world.

Caveat: In several places this paper refers to Marc Hauser's writings. As I was finishing the paper Hauser was accused of and found solely responsible for eight counts of scientific misconduct. Furthermore Gilbert Harman has accused him of having plagiarized parts of his book *Moral Minds* from John Mikhail. To the best of my knowledge none of the problematic experiments are relevant for the argument of this paper. The general nativist position I discuss is shared by a number of people and is not necessarily dependent on Hauser's work. As for *Moral Minds* I still think that it provides a reasonable overview of much current empirical research on moral psychology.

C.F. Nielsen (✉)
Department of Culture and Society – The Study Committee for Philosophy,
Aarhus University, Denmark
e-mail: filcfn@hum.au.dk

B. Musschenga and A. van Harskamp (eds.), *What Makes Us Moral? On the capacities and conditions for being moral*, Library of Ethics and Applied Philosophy 31,
DOI 10.1007/978-94-007-6343-2_11, © Springer Science+Business Media Dordrecht 2013

These two ways of interpreting the ‘What makes us moral?’ question are obviously intimately related. Any plausible answer to the first question must include some account of how human beings *have* come to possess or *can* come to possess the capacities needed for competent moral agency. And in order to answer the second question, the question concerning moral development, we need some idea of what moral agency consist in and what abilities and faculties a human being needs to possess in order to be a competent moral agent.

Moral nativism provides an influential and important answer to both of these questions.[1] With regard to the first question moral nativism claims that (1) a capacity for moral judgment, a capacity to cognitively distinguish between (morally) right and (morally) wrong, is a necessary feature of moral agency and (2) that this capacity is structured by certain innate principles or mechanisms, mechanisms which are deeply embedded in human nature.

As for the second question, the question concerning moral development, nativism claims that these innate principles or mechanisms are what enable and structure the moral developmental process. In short: moral nativism claims that what makes us moral, and what makes it possible for us to eventually develop the judgmental capacities exhibited by mature moral agents, are principles, structures or mechanisms which are embedded in and hence an integral part of human nature.

This paper provides a critical discussion of certain limitations of current nativist approaches to the question of moral development. It does not dispute, nor does it intend to dispute, the basic nativist claim that moral development and functioning to some extent depend upon innate mechanisms or structures. The aim of the paper is more modest, namely to issue a warning against a lingering reductive tendency found among certain contemporary moral nativists: a tendency to greatly exaggerate the importance of such innate mechanisms for moral development while simultaneously downplaying the importance of other factors in this process. More precisely: the paper argues that the morally relevant input available in the social and cultural environment of human beings is much richer and more varied than typically acknowledged by moral nativists, and that by ignoring this richness the nativist runs the risk of seriously distorting her understanding of the very phenomenon she wants to explain.

[1] People sometimes distinguish between (moral) innatism and (moral) nativism, with innatism being the doctrine that the human mind is born with certain ideas and knowledge, and nativism being a specific modern version of this doctrine, which uses genetics, cognitive science and evolutionary biology to explain the existence of such innate ideas/knowledge. In this paper I use nativism indiscriminately to refer to all theories which take the human mind to be endowed with innate ideas/principles/mechanisms.

11.2 Setting the Stage

Historically moral nativism can be traced back at least as far as Plato. In *Meno* and *Phaido*, two of his most famous dialogues, Plato thus argues (or rather; has his protagonist Socrates argue) that certain (moral) principles or structures are imprinted upon the human soul prior to birth, and then subsequently brought to mind, remembered or recollected through a process of (moral) education. As Socrates puts it:

> The soul, then, as being immortal, and having been born again many times, and having seen all things that exist, whether in this world or in the world below, has knowledge of them all; and it is no wonder that she should be able to call to remembrance all that she ever knew about virtue, and about everything; for as all nature is akin, and the soul has learned all things; there is no difficulty in her eliciting or as men say learning, out of a single recollection – all the rest, if a man is strenuous and does not faint; for all enquiry and all learning is but recollection. (Plato. Meno, Trans. Benjamin Jowett 1871)

According to Plato the moral learning process thus does not start from scratch, but presupposes and draws upon principles or structures embedded in the human mind; principles which are needed to explain how human beings acquire the knowledge implicit in and necessary for moral virtue. And this, in a nutshell, is the basic claim of moral nativism: that the moral development and moral functioning of human beings cannot be properly explained or understood without appeal to innate principles, structures or mechanisms dedicated to fostering these processes.

Plato's version of moral nativism quite obviously relies on highly implausible, some might say wildly speculative, metaphysical assumptions such as the possibility of re-incarnation and the existence of a non-material, immortal soul. Many other classic nativist positions such as those of Descartes and Leibniz implicitly or explicitly appeal to God in order to explain the existence and importance of innate moral principles. Such claims do not sit well with modern science, nor do they cohere with or seem necessary for the most plausible current views on morality.

Contemporary forms of moral nativism however do not rely on or appeal to such controversial, metaphysical claims and concepts. Today moral nativists employ data from evolutionary biology, empirical psychology and cognitive science to argue that human beings are evolutionary adapted for and (presumably) genetically encoded with principles, faculties or mechanisms, which facilitate and structure the development of moral competence.[2]

One prominent argument in favour of moral nativism is the so-called *Poverty of the Stimulus (POS) argument*. The POS argument claims that the moral stimuli to which human beings are exposed are insufficient to generate and explain the richness and complexity of mature moral agency. More concretely: the POS argument claims

[2] See Hauser (2006) for a comprehensive overview of the biological, evolutionary and cognitive theories which inform contemporary moral nativism.

that the moral experience of children, the morally relevant input to which they are exposed, is too impoverished to generate and explain the complex moral distinctions and responses, which even young children appear capable of making.[3]

A prominent example of this apparent asymmetry between moral input and moral output often quoted by moral nativists such as Mikhail and Hauser is the Doctrine of Double Effect (DDE). The DDE is a rather sophisticated moral principle, formulated and defended by a number of moral philosophers as well as many Catholic theologians (Foot 1967; Aquinas 1988). Crudely put the DDE states that when an agent performs an action which will result in both good and bad effects, the action is morally permissible if and only if the action itself is morally good (or at least morally neutral) and the bad effects are not the primary intended effects of the action, nor a means to achieving the good effects.[4]

Interestingly, it turns out that when faced with hypothetical examples of moral dilemmas (typically trolley-examples where a runaway trolley has to be derailed or stopped in order to save a number of people by means of deliberately killing someone or letting someone die), many people respond in a way which coheres with the DDE. One might even say that in certain types of cases people's moral judgments seem to be determined by DDE or some analogous principle (Mikhail 2002, 2011, Appendices; Hauser et al. 2008).

However, with the exception of people acquainted with moral philosophy and practical theology, almost no one is capable of formulating the DDE on their own. Nor is the DDE usually part of the explicit moral instructions, which children receive from their parents and teachers. As Harman puts it: 'An ordinary person was never taught the principle of Double-Effect or the deflection principle, and it is unclear how such a principle might have been acquired from the examples available to the ordinary person' (Harman 2000, p. 225; see also Dwyer 2009).

The moral input available to the agent (explicit instructions and other socially transmitted forms of information) thus seems insufficient to explain the output

[3] Noam Chomsky is often credited as being the first to formulate (a version of) the poverty of the stimulus argument. Chomsky's argument (see Chomsky 1957, 1959) did not concern morals, but was directed against empiricist accounts of language acquisition such as B.F. Skinner's. Moral nativists such as Susan Dwyer, Marc Hauser and John Mikhail have later adopted Chomsky's argument and directed it against empiricist accounts of moral development. It is worth noting however that already Plato formulated and made use of (a version of) what is nowadays known as the POS argument. See *Meno* 82A-86C, where Socrates gradually elicits sophisticated mathematical reasoning from a slave who has no previous experience with mathematics, and uses this as proof that knowledge is not something taught or learned, but something which is always-already embedded in the human mind, albeit in inarticulate and implicit form.

[4] The DDE can be more formally stated as follows: 'A person may licitly perform an action that he foresees will produce a good and a bad effect provided that four conditions are verified at one and the same time: (1) that the action in itself from its very object be good or at least indifferent; (2) that the good effect and not the evil effect be intended; (3) that the good effect be not produced by means of the evil effect; (4) that there be a proportionately grave reason for permitting the evil effect' (Mangan 1949, p. 43).

(moral judgments which systematically cohere with DDE). So how are we to explain the fact that people's explicit moral judgments quite systematically seem to adhere to DDE?

There are other examples of apparent asymmetry between moral input and moral output. Even very young children thus seem able to distinguish between moral and conventional rules (Turiel 1983; Smetana 1983. See Machery and Mallon 2010 for a critical discussion of the experimental and conceptual evidence for this claim). And children at an early age seem to have very distinct notions of what constitutes a fair distribution of goods and resources (Olson and Spelke 2008; Rochat et al. 2009). How do children learn to draw such sophisticated moral distinctions? The POS argument claims that the conceptual resources and moral information needed to develop these distinctions cannot be found in or extrapolated from the moral environment of the child. So where do these cognitive capacities come from?

This is *the Acquisition Problem*: The problem of how human beings acquire and develop the sophisticated and complex cognitive apparatus which characterizes (competent) moral judgment, when the social and natural environment seems incapable of providing the necessary resources. How do human beings become mature and competent moral agents, capable of highly sophisticated moral deliberation and action, if the moral environment is as impoverished as the POS argument claims? How, in other words, do something as complex and sophisticated as (mature) moral competence arise from such a meagre starting point?

Moral nativism provides a solution to the Acquisition Problem. If, as the POS argument claims, the moral judgments made by human beings are too sophisticated and complex to be explained by the moral stimuli to which we are exposed, then the only other plausible solution is that these moral judgments rely and necessarily presupposes possession of certain innate moral principles or learning mechanisms. Such innate principles which from the very outset structure, refine and organize the moral development of human beings would bridge the apparent gap between moral input and moral output and provide an explanation for the (acquisition of) the sophisticated human capacity for moral judgment (Dwyer 1999; Hauser 2006, pp. 65–67).

As John Mikhail puts it:

> the moral competence of both adults and children exhibits many characteristics of a well-developed legal code, including abstract theories of crime, tort, contract and agency. Since the emergence of this knowledge cannot be explained by appeals to explicit instruction, or to any known processes of imitation, internalization, socialization and the like, there are grounds for concluding it may be innate. (Mikhail 2008, p. 354)

One can plausibly disagree with Mikhail's comparison of moral competence with a 'well-developed legal code'. But his basic nativist point, that innate principles or mechanisms are required to explain the development of moral competence, does not depend on his exact characterization of morality. All that is needed for the nativist argument to get going is the acknowledgement that the moral competence of human beings (children as well as mature moral agents) exhibits a complexity and sophistication that goes far beyond the input provided by the moral environment.

It is important to note that moral nativists typically do not deny that there is important and relevant information available in our environment, which influences moral development and moral cognitive competence. As Susan Dwyer puts it:

> The nativist claim is not that there is no information in the child's environment relevant to her acquisition of the capacity to distinguish between moral and conventional rules. The nativist's concern is whether that information is sufficient to explain the capacity the child possesses and whether it is available to all children. (Dwyer 2006, p. 241)

The claim that the moral input available to us in our environment is poor or impoverished (the basic premise of the POS argument) should thus be understood in its proper context. The claim is *not* that there is no morally relevant information available in the environment. The claim is rather that this information is not sufficient to explain how children acquire the moral competences (the capacity to distinguish between moral and conventional rules; the ability to judge in accordance with DDE), which they seem to possess. To explain how and why children everywhere, despite being raised in widely different environments, seem to acquire and express more or less the same complex moral capacities in more or less the same way, we must presuppose the existence some sort of innate learning mechanism or some innate structure of basic and implicit moral principles and rules.

The POS argument thus claims to establish that the moral input available to us is impoverished *relative to the perceived capacities of the developing child*. The argument does *not* establish that moral input *as such* is impoverished, i.e. extremely limited in either content (the actual information presented to us) or range (the types of input available). The claim is merely that the moral stimuli we receive and are exposed to are too limited to explain the complexity of the moral judgments we seem capable of. There is an asymmetry between moral input and moral output, and it is in light of this asymmetry that our moral input appears impoverished.

In practice however contemporary moral nativists such as Susan Dwyer, Marc Hauser and John Mikhail have a tendency to implicitly ignore, downplay and underestimate the actual complexity and breadth of the moral stimuli which children are exposed to and encounter during upbringing. Or that at least will be the argument in the rest of this paper. I will argue that the domain of morally relevant input and information provided by the environment extend far beyond what main stream proponents of nativism usually presuppose.

In the next section, I focus on explicit moral instructions and admonitions which nativists often take as the paradigmatic example of the sort of moral stimuli to which children are exposed. Indeed, some moral nativists seem to regard explicit moral instructions as the *only* morally relevant stimuli available in our social environment. Against this view I argue first that children are faced with a far wider and more complex array of explicit moral instructions than nativists usually assume. I then move on to other ways in which the social environment can influence moral development. Employing Kim Sterelny's and G.W.F. Hegel's accounts of how culture and social structures facilitate and shape moral development, I argue that many nativists seem to operate with a rather restricted understanding of what can and should count as ‘morally relevant input’. I then conclude with a brief discussion of whether and to what extent the arguments presented in the paper undermine the nativist position.

11.3 Explicit Moral Instructions

According to Susan Dwyer, Marc Hauser and Bryce Huebner (Dwyer 1999; Hauser 2006; Dwyer et al. 2010) parental admonitions and explicit moral instructions constitute the most important external stimuli to the moral developmental process. In her influential discussion of the acquisition problem Susan Dwyer thus focuses almost exclusively on ‘the explicit moral instructions children typically receive’. Such instructions, Dwyer contends, are either very detailed and context dependent (‘You should not hit Mary just because she took your ice cream!’) or very general and abstract (‘You must not hurt other people’). In both cases the explicit moral admonitions ‘are fairly coarse-grained, offering only limited guidance to children in their future actions.’ Dwyer therefore concludes that

> Absent a detailed account of how children extrapolate distinctly moral rules from the barrage of parental imperatives and evaluations, the appeal to explicit moral instruction will not provide anything like a satisfactory explanation of the emergence of mature moral competence. (Dwyer 1999, pp. 172–173. See also Hauser 2006, pp. 65–66)

The nativist argument is clear: explicit parental admonitions and reprimands are the primary moral input which children are exposed to during upbringing. Such admonitions are either too specific or too coarse-grained to explain the output of the developmental process: the complex and sophisticated capacity for moral judgment possessed by mature moral agents. In short: the moral input is too impoverished to explain the moral output.

Three things are worth noting here. First of all, so called ‘discipline encounters’ – situations in which parents and other authority figures attempt to change a child’s behaviour against its will through explicit moral instructions and admonitions – are a frequent and pervasive feature of upbringing. Collecting data from a number of different studies the development psychologist Martin Hoffman estimates that

> by the end of the second year fully two thirds of all parent–child interactions are discipline encounters. More specifically, children in the 2–10-year age range experience parental pressure to change their behavior every 6 to 9 minutes on average, which translates roughly into 50 discipline encounters a day or over 15.000 a year! (Hoffman 2001, pp. 140–141)

If Hoffman is right, then moral nativists would seem to be right when they claim that explicit verbal instructions are a central, pervasive and inescapable feature of moral upbringing. In fact, discipline encounters appear to be such a pervasive part of the moral developmental process that the somewhat abrupt nativist dismissal of parental admonitions as a potential source of knowledge concerning moral norms begins to look rather strange. The sheer number of discipline encounters provides good reason to think that they do provide an important structuring influence on the development of moral cognition – unless you start out from the assumption that parental interference is necessarily unstructured. And that is clearly not an assumption which can be simply taken for granted. It must be justified and defended. And I have not yet seen a systematic nativist discussion of this point.

This brings us to the second point. Are explicit verbal admonitions necessarily as coarse-grained as moral nativists seem to assume? Hoffman suggests that they are not. Parents do not simply issue commands and *post hoc* evaluations without any explanations. In fact parents do typically offer justifications for their disciplinary instructions and in the process employ and appeal to a variety of different reasons, norms and motivational pressures.

Hoffman has thus shown that so-called 'inductions' – disciplinary encounters where 'parents highlight the other's perspective, point up the other's distress, and make it clear that the child's actions caused it' (Hoffman 2001, p. 143) – are an important and highly effective way of getting children to internalize particular social and moral norms.

Inductions typically appeal to the discomfort and pain expressed by other people. For this reason this particular sort of disciplinary encounters are most efficient when the norms in question involve physical harm. When we are dealing with other forms of norms which do not necessarily involve physical harm, e.g. norms of social etiquette, then other sorts of justification and other means of internalization (explicit power assertions for instance, or threats of love-withdrawal) will have to be employed. (Hoffman 2001, pp. 146–148 and 148–150).

Explicit moral instructions, then, can take many shapes and forms and can be used in a variety of different ways depending on the context and the sort of transgression which is being berated. Explicit verbal admonitions are thus not nearly as coarse-grained and simplistic as moral nativists seem to assume, and provide a much richer source of moral information than the basic nativist line of argument seems to acknowledge.[5]

Finally: parental admonitions are not the only sort of explicit moral advice and instruction available in the child's moral environment. In a cohesive, well-functioning and well-integrated society morality is not regarded, nor does it function as an external system of principles and norms imposed on the society's members through specific disciplinary structures. Values and norms pervade our social

[5] One reviewer questioned the relevance of Hoffman's work for discussions of moral nativism, since the primary focus of Hoffman's research is on empathy and how to foster and develop pro-social motives whereas moral nativists are primarily concerned with moral cognition and the development of the capacity for moral judgment. However, the aspect of Hoffman's research appealed to in this paper concerns the pervasiveness and complexity of disciplinary encounters in moral development. As such it can be fairly clearly separated from Hoffman's research on pro-sociality (although the two are clearly linked). And research on disciplinary encounters is relevant for discussions of moral nativism because explicit parental admonitions, which nativists often take to be the most important moral input children are exposed to in moral upbringing, is a common and crucial form of disciplinary encounter.

environment and are transmitted, verbalized and made explicit in a number of different ways. As Kim Sterelny puts it:

> The narrative life of a community – the stock of stories, songs, myths and tales to which children are exposed – is full of information about the actions to be admired and to be deplored. Young children's stories include many moral fables: stories of virtue, of right action and motivation rewarded; of vice punished. So their narrative world is richly populated with moral examples. (Sterelny 2010, p. 289)

Stories, songs, myths, children's television shows and other narrative forms of entertainment present and articulate moral instructions and values, but do so in ways which are markedly different from e.g. parental scolding or disciplinary inductions. Information concerning morality and morally relevant phenomena can be transmitted in many forms, not merely or simply through explicit verbal admonitions. Such informal forms of moral knowledge transmissions might not be as strict, nor have as direct and distinct an influence as parental reprimands but that does not mean that they are irrelevant or non-existent. At the very least they should be taken into account when we discuss the question of what might count as morally relevant influence on children during moral upbringing.

So, moral nativists are clearly right when they take explicit moral instructions to be an important part of the moral experience of children. However, if discipline encounters are as frequent and widespread as Hoffman suggests, then it seems plausible to assume that their influence on moral development must be quite substantial. At the very least it seems somewhat disingenuous to dismiss the importance of parental discipline as quickly as certain moral nativists (in particular Susan Dwyer) seem to do (Dwyer 1999, p. 173 and n. 3).

Furthermore, the nativist focus on explicit parental admonitions as the primary way of transmitting moral norms to children seems too simplistic and limited. It is simplistic because it greatly underestimates the variety of ways in which parents and other authority figures can distribute, argue for and justify particular moral norms. And it is limited because it leaves out a wide variety of ways in which moral principles and values can be made explicitly known to children. Songs, jokes, movies, sitcoms, children's shows, good night stories, discussions and verbal fights are just a few of the many ways in which the implicit values and norms of a community can be made and become explicit.

Moral nativists thus seem to implicitly skew their understanding of how parents and society can pass on and transmit moral norms to children in ways which support their own position. This is quite understandable, but it is nonetheless problematic and something which it is important to keep in mind when discussing nativist claims and arguments.

11.4 Implicit Social Influence

As already mentioned, morality should probably not be thought of as a system of principles and norms separate from and independent of the rest of society, and imposed on members of society primarily through explicit instructions, disciplinary encounters and parental reproaches. Moral principles and values are a pervasive and intrinsic feature of the social life of all societies. Social institutions and associations embody, express and pursue a multitude of different principles and norms, and by actively participating in the activities of these associations people can gradually become aware of and learn to identify with these norms and principles. Or that at least is a claim made by many philosophers, sociologists and psychologists.

If this is true, then explicit moral instructions and admonitions *are* an important part of moral education, but such explicit verbalizations by no means exhaust the ways in which morally relevant input is passed on to children and adolescents. The implicit (and occasionally explicit) norms which structure the institutions and associations of our common social life provide a rich and varied field of morally significant experiences and stimuli that influence moral development. To illustrate this let us take a brief look at two different ways of understanding this idea.

11.4.1 Kim Sterelny: Human Culture as a Dedicated Learning Environment

Kim Sterelny has argued that human culture can plausibly be viewed as a highly sophisticated learning environment, which is organized so as to (a) secure cross-generational transmission of relevant information and (b) enable human beings to safely and efficiently acquire valuable (evolutionary adaptive) skills and information (Sterelny 2006), 'Parents', Sterelny explains,

> do not just wait and hope that children will acquire the information they need. They organize the informational world of their children. They make trial-and-error learning safer, by substituting social signals of failure for environmental ones. They provide informational resources: toys, tools and other props . . . . In short, the parental generation actively engineers the learning environment of the next generation: the reliability of the social transmission of normative information across the generation is much enhanced by parent's informational engineering. (Sterelny 2010, pp. 290–291)

Parents organize and structure the social environment so that children have access to and become acquainted with a far richer array of (morally) relevant information than they would have without such parental intervention. And they provide children with means (toys, tools, games) which augment the reliability of the cross-generational transmission of normative information. According to Sterelny the moral information available to children is thus far more varied and structured than usually acknowledged by moral nativists. And this, Sterelny believes, undermines at least part of the force of the POS argument, which explicitly regards the social environment as being either too unstructured to provide moral guidance or too information-poor to provide the necessary conceptual distinctions to explain the moral competence of children.

Sterelny also takes explicit issue with another tendency which he thinks characterizes at least some forms of nativism: the idea that children are primarily passive recipients of socially transmitted information and norms. At least in some places Dwyer, Huebner and Hauser thus seem to think that children primarily become aware of the values and norms embodied in their social surroundings by either observing the behaviour and actions of other people or by having these values directly imposed upon them by authority figures such as their parents (Dwyer 1999, pp. 171–172; Dwyer 2006, Section 1; Dwyer et al. 2010, Section 2.2).

As Sterelny points out, this is hardly an accurate description of how children engage with their social environment. Children actively explore their surroundings and experiment with their physical and social worlds. They learn by doing and (equally important) by doing wrong. They engage in cooperative ventures with friends, get involved in conflicts with their peers and parents, and in the process have to learn how to navigate the shores of social life. Active involvement with and participation in different social associations and contexts expands the range of possible moral experiences and hence provides the child with numerous new ways of acquiring different sorts of norms and values (Sterelny 2010, p. 291).

Furthermore, children are not simply passive observers, who have to figure out 'from the outside' the strange mores and customs of their peers and parents. They are not explorers passing through a strange and exotic country, where they have to learn and adapt to the local way of life. 'The task of the child,' so Sterelny explains, 'is not to *discover* the set of practices governing a particular community. Rather, her task is to join her community; to share rather than describe those norms' (Sterelny 2010, p. 291).

Children are prospective members of a particular society, and as such they are always-already engaged with and part of (at least some of) the social practices and institutions which define their community. They do not stand apart from and passively observe a system of norms and values, which they will eventually come to internalize and identify with. They are always-already busy learning, adopting, internalizing and identifying with the particular norms and moral outlook that distinguish their community from others.

Human culture thus, according to Sterelny, provides a teleologically structured, information-rich environment engineered so as to optimize the child's exposure to and acquisition of morally relevant principles, skills and norms. The more complex the culture in question, the more complex and varied the available learning environment will be. And this complexity is further enhanced when we factor in that children actively seek out and engage others in different kinds of cooperative ventures (games, plays), which extend the range of possible moral input far beyond those available to the 'passive moral observer' envisioned by at least some moral nativists.

Sterelny does not deny that the social transmission of norms, values and normative information relies upon a biological foundation, or that moral learning depends upon and presupposes evolutionary developed psychological mechanisms. But he denies the nativist claim that these mechanisms are specifically adapted for moral learning. 'Normative cognition could be genuinely universal, and develop robustly without that development relying on innate, tacit, abstract principles'

(Sterelny 2010, p. 294). Our basic psychological and biological makeup, combined with the learning processes embedded in our (informationally engineered) social environment is, so Sterelny claims, enough to explain the complex and sophisticated moral competence of children and mature adults.

11.4.2 Hegel: **Sittlichkeit** *and the Social Cultivation of Emotions and Needs*

Sterelny is not first to stress the importance of human culture for the transmission of moral norms and the influence of social structures on the moral development of human beings. The eighteenth-century German philosopher G.W.F. Hegel devoted much of his social, political and anthropological writings to precisely these questions. In fact, Hegel reaches many of the same conclusions as Sterelny, but from quite a different starting point. Hegel quite obviously did not know anything about evolutionary biology, the methodological starting point of Sterelny's arguments. Hegel's arguments were rather based on philosophical considerations concerning the inherent historicity of human beings and the social world.

In the final part of his *Outlines of the Philosophy of Right* (Hegel 2008),[6] in the section entitled 'Ethical Life' [*Sittlichkeit*], Hegel sets out to explicate what he takes to be the implicit normative structures inherent in and distinctive of the most important institutions of modern society: the nuclear family, civil society (basically the free market and the institutions associated with the production and distribution of goods) and the modern state.

Hegel's basic argument in this section is that participation in and membership of these particular institutions (and, one may plausibly assume, other institutions similar to these) transform the initially self-centred desires and wants of human beings in ways which enable individuals to gradually adopt, identify with and be motivated by less selfish, more general, perhaps even universal concerns.

Going into a detailed account of Hegel's notion of *Sittlichkeit* would take us too far afield from our present discussion, but a short overview of some of his main points can serve to illustrate his basic idea. To simplify things I shall here focus on Hegel's account of the family and of civil society.

According to Hegel the essential bond that unites the members of the modern nuclear family is love, and love, so Hegel believes, implies letting one's own particularity be defined through and constituted by others. 'Love', Hegel thus explains,

> means in general terms the consciousness of my unity with another, so that I am not in isolation by myself but win my self-consciousness only through the renunciation of my independence [*Fürsichsein*] and through knowing myself as the unity of myself with another and of the other with me. (Hegel 2008, §158 Z)

[6] All citations and references to Hegel's *Outlines of the Philosophy of Right* are to the numbered paragraphs in this work. Hegel's additions (*Zusatzen*) to these paragraphs are indicated by a 'Z'.

In love my individuality is necessarily entwined with and defined through my relation to particular others, paradigmatically the members of my family.

> 'Hence, in a family, one's disposition is to have self-consciousness of one's own individuality within this unity as the essentiality that has being in and for itself, with the result that one is in it not as an independent person but as a member' (Hegel 2008, §158).

In the family, Hegel thus claims, my individuality, my life and my happiness, is inextricably intertwined with the individuality, life and happiness of particular others: parents, children, spouse etc. Who I am, the particular individual I take myself to be, is defined by and constituted through my membership of my family.

In *civil society* on the other hand the primary social bond that ties people together is enlightened self-interest. In the modern world everyone must participate in the social production and distribution of goods and services in order to earn a wage and thereby acquire the means for satisfying their own particular needs and ends. As active participants in this 'system of needs' (Hegel 2008, §188), we thus all implicitly rely on and depend upon the active cooperation of countless others as a means of satisfying our own interests. As Hegel puts it:

> In the course of the actual attainment of selfish ends . . . there is formed a system of complete interdependence, wherein the livelihood, welfare, and rightful existence [*rechtliches Dasein*] of one individual are interwoven with the livelihood, welfare and rights of all. (Hegel 2008, §183)

However, being a participant in the social production and distribution of goods not merely implies that my desires become necessarily intertwined with the desires of others. It also changes the very meaning and content of these desires by *transforming* our immediate needs and *creating* new and different sorts of desires and wants:

> Understanding, with its grasp of distinctions, multiplies these human needs, and since taste and utility become criteria of judgment, even the needs themselves are affected thereby. Finally, it is no longer need but opinion which has to be satisfied . . . . (Hegel 2008, §190 Z)

Civil society, in particular the system of need, 'the infinitely complex, crisscross, movements of reciprocal production and exchange, and the equally infinite multiplicity of means therein employed' (Hegel 2008, §201), thus fundamentally alters the content of my desires and needs. And since the satisfaction of my desires depends upon my active participation in this system, my desires, needs and wants and thus very individuality becomes inextricably bound up with the individuality of others. As Hegel puts it:

> The fact that I must direct my conduct by reference to others introduces here the form of universality. It is from others that I acquire the means of satisfaction and I must accordingly accept their views. At the same time however, I am compelled to produce means for the satisfaction of others. We play into each other's hands and so hang together. To this extent everything particular becomes something social. (Hegel 2008, §193 Z)

We are, Hegel claims, necessarily compelled to participate in and be members of 'the system of needs'. This introduces an irreducibly social, potentially universal, dimension into our individual needs, wants and desires. Through our participation in the general patterns of self-interested behaviour constitutive of civil society our individuality thus becomes infused with commonality.

According to Hegel the family and civil society thus represent different ways in which our social environment can gradually transform our immediate self-interest and enable us to adopt a broader, more general view of our interests and the values and norms which structure modern social life. As a member of a family and as a participant in the system of needs our individuality, our subjectivity, is always always-already socially mediated, always-already intimately intertwined with the life of others. Reflection on what it means to be a member of a family and to participate in the system of needs thus reveals that our interests and desires are not necessarily opposed to and in conflict with the needs and desires of others, and that the satisfaction of our particular, individual needs actually depends upon and presupposes the active cooperation of other people.

Hegel's suggestions are admittedly highly speculative. But they are also very suggestive and point out different ways in which our social surroundings can influence and transform our immediate desires and wants in morally relevant directions. Morality almost certainly involves being willing and able to occasionally be motivated by general, perhaps even universal, concerns which extend beyond our own immediate self-interest. What Hegel proposes is that the process through which we become capable of this is, at least in part, a social process of affective and emotional cultivation.[7]

Nativists rarely (if ever) provide any explicit discussion of the sort of social cultivation of desires, needs and wants, which Hegel (and other political philosophers such as Marx and Rawls) take to be a fundamental part of moral development. One reason for this is that moral nativists are primarily interested in our capacity to make (complex) moral judgments. More precisely, they are interested in the capacity for moral judgment which expresses itself in and can be investigated through systematic applications of the kinds of hypothetical questions mentioned in Sect. 11.2 above. And this capacity is often presumed to be both conceptually and empirically distinct from our capacity for moral motivation and from the sort of affective developmental transformation discussed by Hegel and other political philosophers.

It has thus been argued, and empirical studies have partly confirmed, that psychopaths and people with 'acquired sociopathy' reliably make moral judgments that cohere with those of normal moral agents, but lack the appropriate motivation

[7] Contrary to what many people might think Hegel is perfectly willing to admit the importance of (uncultivated) human nature in explaining and understanding the moral development of human beings. Hegel's analysis of socially mediated affective and cognitive development is firmly grounded in a detailed account of human beings as part of and embedded within nature. For Hegel human beings are importantly (but not simply and not only) embodied and desiring creatures; creatures who want and need material things in order to survive and who utilize and develop their capacity for reason as a way of acquiring that which they desire. See e.g. Hegel's *Phenomenology of Spirit* (Hegel 1971) or his philosophical anthropology as presented in part three of his *Encyclopaedia of the Philosophical Sciences* (Hegel 2007).

to act on and be moved by these judgments (see e.g. Roskies 2003). Psychopaths' capacity for moral judgment thus appears intact whereas their motivational and affective systems show signs of being severely damaged. This would seem to indicate that the cognitive capacities involved in moral judgment are distinct from and can be investigated and discussed in separation from our affective capacities. If this is true, then moral nativists are quite right in not including the social cultivation of desires and emotions in their discussion of how we acquire and develop the capacity for moral judgment.

There are, however, reasons to be somewhat sceptical of the claim that motivation and the capacity for moral judgment can be neatly separated. For one thing, the empirical studies are not as clear-cut as those who subscribe to the separation thesis might wish. In a well-known paper from 1995 Blair thus showed that psychopaths (people with psychopathic tendencies) seem less able to make the moral/convention distinction than people from a control group (Blair 1995). This would seem to indicate some form of impairment to the psychopaths' capacity for moral judgment. (See Levy 2007; Vargas and Nichols 2007; Maxwell and Le Sage 2009 for critical discussions of the validity and possible implications of Blair's findings.)

Whether and to what extent this diminished capacity is related to the affective impairments usually associated with psychopathy – shallow emotions; lack of remorse and guilt; lack of (certain forms of) empathy – is unclear. At the very least these findings point toward the rather modest conclusion that the question of how to understand the relation between affective and cognitive (judgmental) structures in moral functioning is still open for discussion, and that we need more studies before any clear conclusions can be drawn (see e.g. Prinz 2008). Perhaps Hegel's analysis of the development of extended forms of other-regarding concern is not simply irrelevant for an understanding of moral judgment. Only time (and further studies and experiments) will tell.

Another reason why moral nativists typically do not discuss the sorts of social cultivation of desires and motivations which Hegel is interested in, might be purely pragmatic. They have chosen to focus on another aspect of moral development; the fact that at least part of the developmental process is structured by innate, psychological mechanisms, which are the same for all human beings across societies and cultures. The nativist bias might thus be nothing but a methodologically motivated 'division of research'. If so then there is in principle no reason why ideas such as Hegel's cannot eventually be integrated with nativist ideas and concerns, when we have developed a more adequate and comprehensive account of moral development and moral functioning, which can embrace both lines of research.

However, when you read at least some moral nativists there seem to be more at stake than simply a pragmatic decision to focus on a particular aspect of human moral development. The very fact that people such as Mikhail, Dwyer and Hauser adopt and endorse the POS argument as their starting point seems to indicate a deep-seated bias against the relevance and importance of social and cultural influence on moral learning. And this implicit bias is further strengthened when you factor in that the socialization theories and theories of social learning, which they usually oppose and criticize, are typically not the best, most well-developed and interesting theories on the market, but rather simplistic versions or interpretations of these theories. (For a clear example of this, see Dwyer's discussion of Bandura's social learning theory in Dwyer 1999, pp. 170–173.)

This might lead an uncharitable reader to suspect that the reason why moral nativists do not engage with theories such as Hegel's, is that they simply do not believe in the kind of socially mediated cultivation of emotions and affects which he advocates. Now, there might be very good nativist reasons for dismissing Hegel's ideas. There might even be good nativist arguments which conclusively prove that Hegel's (and Sterelny's) views on the importance of the social context for moral development are either wrong or irrelevant for an adequate understanding of moral learning. But, and this is important: this is clearly not something which we can simply presuppose for the sake of argument. In particular not when the very question we are discussing is how impoverished our social and moral environment really is.

The moral nativists thus owe us an extended explanation or argument for why, precisely, the sort of implicit socialization and enculturation proposed by Hegel (and in different terms by Sterelny) cannot and does not count as part of the basic moral input which influences and structures moral development. Or, to put the same point slightly less provocatively: they need to explain in greater detail why they usually chose to disregard the sort of questions which Hegel and Sterelny take to be important and interesting.

11.5 Nativism or Not?

The aim of the previous sections has been to point out ways in which our social environment can plausibly be said to influence and shape moral development, perhaps even influence the acquisition of a capacity for moral judgment. These socially mediated forms of influences are, so I have argued, rarely discussed by moral nativists. Insofar as they *do* discuss these influences nativists typically adopt a very restrictive and quite uncharitable view of the relevant phenomena, as their discussion of explicit moral instructions and admonitions clearly shows.

Why is this important? It is important because from the very outset it skews the discussion of the POS-argument and the acquisition problem so as to favour a (strongly) nativist solution. The more restrictive and limited your views on what can plausibly count as moral input is, the stronger the POS argument will appear. And the stronger the POS argument appears to be, the more difficult and intractable the acquisition problem will seem. And the more difficult this problem seems, the more plausible an appeal to innate principles or dedicated learning mechanisms will look. In short: If you start out with implicitly reductive and simplistic accounts of what kinds of moral output are available to us in our environment, then the discussion of moral nativism is biased in favour of nativism from the very beginning. And this might blind us to important and interesting ideas and viewpoints which should be taken into consideration when discussing the plausibility and scope of moral nativism. If on the other hand you start out with a more generous, complex and sophisticated account of the moral input provided by and made available through our environment, then the POS argument becomes less compelling and the acquisition problem loses at least some of its sting.

That being said, do any of the arguments presented above imply that moral nativism is false? No, they do not, for a number of reasons.

First of all, the most that this paper can possibly hope to show is that the POS argument and the acquisition problem in and of themselves do not necessarily provide sufficient evidence for nativism. The strength of both the POS argument and the acquisition problem depend on how you frame the notion of ‘moral input’. And as we have seen moral nativists typically adopt a restrictive and (in my opinion) rather implausible definition of what counts as morally relevant input, which we should not necessarily accept – at least not without further discussion.

Secondly, the POS argument and the acquisition problem are not the only tools in the nativists’ argumentative repertoire. The apparent universality of certain types of moral distinctions and beliefs for instance is thus quite often used to argue for the existence of an innate moral sense or set of dedicated moral learning mechanisms. So, even if this paper’s critique of the POS argument is valid this does not suffice to positively establish that moral nativism is false, nor does it rule out the possibility of other arguments for nativism.

Thirdly, this paper does in fact *not* undermine the POS-argument. The argument in the preceding sections has merely been that nativists typically underestimate the complexity, variety and pervasiveness of the morally relevant input provided by our (social) environment. Nativists could, at least in principle, accept this; they could admit that the information and stimuli available in moral experience is more structured and complex than they have usually acknowledged. Having accepted this, the moral nativist could then go on to claim that no matter how complex, sophisticated and information-rich our social and moral environment is, it is *still* an open question whether moral experience in all its complexity and variability is sufficient to adequately explain the development of mature moral competence.

Richard Joyce, who himself has certain problems with contemporary versions of moral nativism, seems to adopt such a position. After reviewing the current debate on the POS argument (including Sterelny’s criticisms of the nativist view) he issues the following challenge in a recent paper:

> The ingredients offered by Sterelny suffice for a social creature who is sensitive to harm situations, who feels empathy for his fellows, who generalizes from exemplars, for whom departures from the cooperative order are memorable and salient, and who, as a consequence, operates extremely well in his social world. But where is the morality? (Joyce 2013, p. 564)

Joyce does not view this as merely a rhetorical question. He does not suppose that his readers will necessarily and immediately agree that the correct answer to this challenge is: ‘It is not there’. His point is rather that this is the right sort of question to ask in relation to the debate concerning moral nativism and that we need to rid ourselves of the simplistic terms and distinctions which up until now have marred and distorted this debate.

This would also be my basic point and take-home message. I agree with the nativists that some sort of dedicated learning mechanism is in all likelihood necessary to fully explain the development of our rather complex (mature) moral competence. But I am not convinced by the typical nativist account of the moral stimuli available to us in our environment. By adopting accounts of the process of moral development, which are either blatantly simplistic or implicitly reductive, moral nativists quite often seem to beg the question concerning the very possibility of a non-nativist position. And this is highly problematic, not simply because it weakens the nativists' argument, but also, and more importantly, because it flattens and deflates the discussion of moral development.

Let us return to the question from the beginning of the paper: What makes us moral? Nativists claim that innate principles or mechanisms are a necessary, but not sufficient, part of the answer to this question. Innate structures are needed to bridge the apparent asymmetry between moral input and moral output and explain the acquisition and development of our complex capacity for moral judgment. This seems quite reasonable, and there are a number of empirical studies which seems to back up at least some of the nativist claims.

However, in their attempts to explain and investigate these mechanisms nativists occasionally lose sight of the fact that there is more to moral development and moral cognition than simply activating or bringing online certain innate moral structures. In this paper I have thus focused on locating and criticizing some of the blind spots of contemporary moral nativists. More precisely, I have argued that the moral input available to us in our social environment is more complex and sophisticated than moral nativists typically seem willing to acknowledge.

To paraphrase Hamlet: There are more things between birth and mature moral competence than are dreamt of in the theories of contemporary moral nativists. My hope is that nativist psychologists in the future will pay more attention to the (inherent) complexity of our moral environment and the moral stimuli it makes available to us. More broadly, I hope that the debate between nativists and non-nativists will soon move beyond the stereotypical caricatures which both sides at the moment seem all too happy to accept and indulge in. Morality, moral development and the interaction between social context and human nature are, not surprisingly, phenomena which are more complicated than our current theories seem able to explain and account for.

Acknowledgments Thanks to Emily Hartz for comments and corrections on an earlier version of this paper. Also thanks to the anonymous reviewer who gave helpful suggestions and criticism. Finally thanks to the Centre for Subjectivity Research, University of Copenhagen, for the postdoc funding that enabled me to write this paper.

References

Aquinas, Thomas. 1988. Summa Theologica II-II, Q. 64, art. 7, 'Of killing'. In *On law, morality, and politics*, ed. W.P. Baumgarth and R.J. Regan, 226–227. Indianapolis/Cambridge: Hackett Publishing.

Blair, R.J.R. 1995. A cognitive-developmental approach to morality: Investigating the psychopath. *Cognition* 37(1): 1–29.

Chomsky, N. 1957. *Syntactic structures*. London: Mouton.

Chomsky, N. 1959. A review of B.F. Skinner's verbal behavior. *Language* 35(1): 26–58.

Dwyer, S. 1999. Moral competence. In *Philosophy and linguistics*, ed. K. Murasugi and R. Stainton, 169–190. Boulder: Westview Press.

Dwyer, S. 2006. How good is the linguistic analogy? In *The innate mind, Vol. 2: Culture and cognition*, ed. P. Carruthers, S. Laurence, and S. Stich, 237–256. Oxford: Oxford University Press.

Dwyer, S. 2009. Moral dumbfounding and the linguistic analogy: Methodological implications for the study of moral judgment. *Mind and Language* 24(3): 274–296.

Dwyer, S., B. Huebner, and M. Hauser. 2010. The linguistic analogy: Motivations, results, and speculations. *Topics in Cognitive Science* 2(3): 486–510.

Foot, P. 1967. The problem of abortion and the doctrine of double effect. *Oxford Review* 5: 5–15.

Hauser, M. 2006. *Moral minds. How nature designed our universal sense of right and wrong*. New York: Ecco Press.

Hauser, M., L. Young, and F. Cushman. 2008. Reviving Rawls's linguistic analogy. In *Moral psychology, Vol. 2: The cognitive science of morality: Intuition and diversity*, ed. W. Sinnott-Armstrong, 107–143. Cambridge: MIT Press.

Harman, G. 2000. Moral philosophy and linguistics. In *Explaining value and other essays in moral philosophy*, 217–226. Oxford: Oxford University Press.

Hegel, G.W.F. 1971. *Phenomenology of spirit.* Trans. A.V. Miller. Oxford: Oxford University Press.

Hegel, G.W.F. 2007. *Hegel's philosophy of mind. Part three of Hegel's encyclopaedia of the philosophical sciences (1830).* Trans. W. Wallace and A.V. Miller, rev. with introduction and commentary by Michael Inwood. Oxford: Oxford University Press.

Hegel, G.W.F. 2008. *Outlines of the philosophy of right.* Trans. and ed. T.M. Knox, rev. with introduction by Stephen Houlgate. Oxford: Oxford University Press.

Hoffman, M. 2001. *Empathy and moral development: Implications for caring and justice*. Cambridge: Cambridge University Press.

Joyce, R. 2013. The many moral nativisms. In *Cooperation and its evolution*, eds. K. Sterelny, R. Joyce, B. Calcott, and B. Fraser, 549–572. Cambridge: MIT Press.

Levy, N. 2007. The responsibility of the psychopath revisited. *Philosophy, Psychiatry & Psychology* 14(2): 129–138.

Machery, E., and R. Mallon. 2010. Evolution of morality. In *The moral psychology handbook*, ed. J. Doris and Moral Psychology Research Group. Oxford: Oxford University Press.

Mangan, J. 1949. A historical analysis of the principle of double effect. *Theological Studies* 10: 41–61.

Maxwell, B., and L. Le Sage. 2009. Are psychopaths morally sensitive? *Journal of Moral Education* 38(1): 75–91.

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. Social Science Research Network. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=762385. Accessed 28 June 2012.

Mikhail, J. 2007. Universal moral grammar: Theory, evidence, and the future. *Trends in Cognitive Sciences* 11(4): 143–152.

Mikhail, J. 2008. The poverty of the moral stimulus. In *Moral psychology, Vol. 1: The evolution of morality: Innateness and adaptation*, ed. W. Sinnott-Armstrong, 353–360. Cambridge: MIT Press.

Mikhail, J. 2011. *Elements of moral cognition: Rawls's linguistic analogy and the cognitive science of moral and legal judgment*. Cambridge: Cambridge University Press.

Olson, K.R., and E.S. Spelke. 2008. Foundations of cooperation in preschool children. *Cognition* 108(1): 222–231.

Plato. Meno, Trans. Benjamin Jowett (1871). Project Gutenberg: http://www.gutenberg.org/files/1643/1643-h/1643-h.htm

Prinz, J. 2008. Resisting the linguistic analogy: A commentary on Hauser, Young, and Cushman. In *Moral psychology, Vol. 2: The cognitive science of morality: Intuition and diversity*, ed. W. Sinnott-Armstrong, 157–170. Cambridge: MIT Press.

Rochat, P., M.D.G. Dias, L. Guo, T. Broesch, C. Passos-Ferreira, A. Winning, and B. Berg. 2009. Fairness in distributive justice by 3- and 5-year-olds across 7 cultures. *Journal of Cross-Cultural Psychology* 40(3): 416–442.

Roskies, A. 2003. Are ethical judgments intrinsically motivational? Lessons from 'acquired sociopathy'. *Philosophical Psychology* 16(1): 51–66.

Smetana, J. 1983. Social cognitive development: Domain distinctions and co-ordinations. *Developmental Review* 3(2): 131–147.

Sterelny, K. 2006. The evolution and evolvability of culture. *Mind and Language* 21(2): 137–165.

Sterelny, K. 2010. Moral nativism: A skeptical response. *Mind and Language* 25(3): 279–297.

Turiel, E. 1983. *The development of social knowledge: Morality and convention*. Cambridge: Cambridge University Press.

Vargas, M., and S. Nichols. 2007. Psychopaths and moral knowledge. *Philosophy, Psychiatry & Psychology* 14(2): 157–162.

Chapter 12
Learning to Be Moral

Julia Hermann

12.1 Introduction

How we respond to the question 'What makes us moral?' depends on what we take to be involved in being moral. Does it consist of acting in accordance with moral norms? Of not just acting in accordance with these norms but also because this is what morality commands? Of being virtuous, i.e. prudent, just, courageous, temperate, and so forth? On what basis do we judge that someone is moral?

In this paper, I will argue that being moral involves the exercise of numerous interrelated rational and emotional capacities, which are developed in practice. We determine that a person is moral on the basis of her actions, judgments, and feelings. The knowledge characteristic of moral agents should not be conceived of as systematic theoretical moral knowledge or propositional knowledge of moral principles, i.e. not as a form of *knowledge that*, but as first and foremost a form of *knowledge how*.[1] However, Gilbert Ryle, on whose distinction between *knowing how* and *knowing that* I draw, was reluctant to classify moral knowledge as *knowing how*. He refers to it sometimes as 'our knowledge of the difference between right and wrong' and sometimes as 'virtue', and comes to the conclusion that it is ultimately not a form of knowledge at all.[2] While Ryle is eager to point out dissimilarities

[1] That moral knowledge is systematic and theoretical is claimed for instance by Michael DePaul (1993, p. 110).

[2] I am grateful to Thomas Schramme for pointing out to me that Ryle has himself considered and rejected the view that moral knowledge is a form of *knowing how*. The two papers in which Ryle does so are relatively unknown, and I had not come across them prior to my encounter with Thomas Schramme at the conference which gave rise to the present book. Due to that encounter, this paper goes far beyond what I presented at the conference.

J. Hermann (✉)
Social Philosophy, Maastricht University, Maastricht, The Netherlands
e-mail: Julia.hermann@maastrichtuniversity.nl

B. Musschenga and A. van Harskamp (eds.), *What Makes Us Moral? On the capacities and conditions for being moral*, Library of Ethics and Applied Philosophy 31,
DOI 10.1007/978-94-007-6343-2_12, © Springer Science+Business Media Dordrecht 2013

between knowing the difference between right and wrong and knowing how to play chess or how to speak Spanish, my aim is to highlight their similarities. As I intend to show, Ryle (1963) provides us with a number of arguments against taking the knowledge characteristic of a moral agent to be *knowledge that*, and in favour of a 'skill model of virtue' (Stichter 2007, p. 183).

In Sect. 12.2 follows my conception of a morally competent agent, which is distinguished from conceptions of the virtuous agent. I will then, in Sect. 12.3, discuss Ryle's distinction between *knowing how* and *knowing that* and argue that the knowledge of moral persons should be understood as first and foremost a form of *knowing how*. In Sect. 12.4, I examine Ryle's reasons against the classification of moral knowledge as *knowledge how*, and look at different models of virtues as skills. Having thus clarified what it means to *be* moral, I discuss, in Sect. 12.5, what *makes* people moral. I will argue that moral competence, which is a condition for being moral, must not only prove itself in practice, but is also acquired through practice. I discuss Ryle's claim that virtue cannot be taught by training and argue that a significant amount of training is involved in the way we learn to be moral.[3] Finally, I show, in Sect. 12.6, that just as the classical understanding of virtue rules out the possibility of a person who is truly virtuous not acting virtuously, it is likewise impossible that an agent who is fully morally competent is not moved by any moral considerations. Therefore, the philosophical conception of the amoralist is implausible and the attempt to convince him futile.

It is widely assumed that whether someone develops moral competence and to what degree depends to a great extent on the environment in which he grows up. Therefore, the adults who create this environment have a significant influence on the moral development of children. I believe that greater efforts are needed to create an environment in which children are able to develop the capacities necessary for moral agency. In this paper, I will not discuss how to improve moral learning, but my focus on moral capacities highlights the importance of that debate.

12.2 Moral Competence

How are we to determine whether someone is moral or not? Where do we have to look? Can we tell this by looking at his actions alone, or do we also need to try looking 'into his head' in order to establish whether he has good intentions? Do we have to assess the person's character, and how do we do that? In my view, we characterize someone as moral if his actions (including verbal actions) and reactions (including emotional reactions), in a number of situations over a period of time, reveal moral competence. Here I seem to disagree with Aristotle, who held that,

[3] He makes this point in Ryle (1972). In Ryle (1958) he talks about people being trained to be virtuous.

in order to act truly virtuously, an agent ‘must have knowledge’, must ‘choose the acts, and choose them for their own sakes’, and ‘his action must proceed from a firm and unchangeable character’ (Aristotle 1998, 1105a17–1105b5). Gerard Hughes interprets Aristotle as believing that ‘if we are to conclude that someone acted virtuously, we need to see *not only what she did or said*; we need to know *how she saw what she was doing or saying*’ (Hughes 2001, p. 55, my italics).

Ryle argued that what distinguishes an intelligent action from a non-intelligent action is not that it is preceded by a secret act in the mind of the agent (Ryle 1963, p. 25). Hughes’s reading of Aristotle suggests that this is what is required to distinguish a moral action from an action that merely conforms to morality. The relevant passage of the *Nicomachean ethics* reads: ‘Actions, then, are called just and temperate when they are such as the just or the temperate man would do; but it is not the man who does these that is just and temperate, but the man who also does them *as* just and temperate men do them’ (Aristotle 1998, 1105b5–1105b28). In my view, this formulation is compatible with the view that an action is just or temperate in virtue of being performed ‘in a certain manner or with a certain style or procedure’ (Ryle 1963, p. 48), not in virtue of being preceded by a secret act in the mind of the agent. To rule out that someone acts the way he does merely to avoid indignation, blame or punishment, we do not have to attempt to look into his head. Moral actions are not distinguished from actions that merely conform to morality by what Ryle calls a secret act in the mind.

Although my morally competent agent shares a number of features with what is called a virtuous agent, my account is not a version of virtue ethics. I start from an understanding of morality or ethics that focuses on how we ought to behave towards others, not from the view that ethics is primarily concerned with how to live a fulfilling life. Moreover, I do not conceive of being moral as consisting mainly or even exclusively of psychological states, as seems to be the view of many virtue ethicists. In addition, in my account conscious deliberation does not have the crucial role it has in ancient virtue ethics. In virtue ethics, spontaneous actions – actions that do not result from a process of deliberation – cannot qualify as truly virtuous actions, while in my view they can show that the actor possesses moral competence. While I am concerned with what are often called ‘moral feelings’ or ‘moral emotions’ (shame, guilt, remorse, indignation, etc.), virtue ethics is concerned with a broad range of human feelings, which the virtuous agent has ‘at the right times, with reference to the right objects, towards the right people, with the right motive, and in the right way’ (Aristotle 1998, 1106b9–1106b35). Furthermore, I oppose the strict division of the soul into rational and non-rational parts characteristic of ancient virtue ethics. My position shares with virtue ethics the criticism of the focus on rules and principles that is characteristic of the dominant moral theories.

In my conception, a morally competent agent has a number of capacities, including a capacity for moral judgment, empathy, and feeling remorse when regarded as appropriate.[4] He is able to take the interests of others into account

[4] For this account of moral competence, see also Hermann (2011, chapter 6), Hermann (2012, 2013).

for their own sake, to imagine how they feel, and to take into consideration the consequences his actions can have for others. In short: he is able to take a moral point of view. Taking such a point of view involves not only thinking, but also feeling in moral terms. It involves feeling pity for someone who is suffering, feeling guilty for having acted selfishly, being ashamed of having disappointed a good friend, and feeling indignation in the face of someone else's bad deeds. It also involves being motivated to do a number of things, such as helping, reducing someone else's suffering, keeping a promise, telling the truth, and so forth.

The competence concerned can be understood as operating at different levels. Deliberation about moral problems takes place at a different level than more spontaneous exercises of our moral capacities, where we don't engage in any kind of reasoning. In many situations there would not even be time for reasoning. While for Aristotle only actions resulting from deliberation can be truly virtuous, I take unreflective and spontaneous acts of helping others in distress etc. as no less moral than actions based on deliberations. In fact, the latter rest on unreflective and spontaneous exercises of moral capacities.[5]

12.3 Knowing How and Knowing That

Let us now consider Ryle's distinction between *knowing how* and *knowing that*. Is it illuminating to think of moral knowledge primarily in terms of *knowing how*? Ryle himself comes to reject talk of moral knowledge altogether, arguing that the learning of virtue does not terminate in any kind of knowledge, but in 'being so-and-so' (Ryle 1972, p. 330).

Ryle introduces the distinction between the two kinds of knowledge in the context of his critique of the 'intellectualist legend' (Ryle 1963, p. 29). According to the legend, any intelligent action is preceded by a mental act such as the contemplation of a rule. In Ryle's positive account of intelligent actions, such actions are exercises of skills, or *knowledge how* (p. 33). A skill or *knowledge how*, in turn, is 'a disposition, or a complex of dispositions' (p. 33), whose 'exercises are observances of rules or canons or the applications of criteria, but ... not tandem operations of theoretically avowing maxims and then putting them into practice (p. 46, my italics).

What does Ryle mean by dispositions? There is no agreement amongst philosophers about what dispositions are. Andreas Kemmerling suggests that Ryle's 'intelligent capacities' should be seen as 'dispositions to correct behaviour' (Kemmerling 1975, p. 163, my translation). They are not 'single-track dispositions' like the brittleness of glass or the smoking habit, but 'higher-grade dispositions' or – a term not used by Ryle himself but common in the literature – 'multiple-track dispositions' (Ryle 1963, pp. 44 ff.). Single-track dispositions are always

[5] I defend this view in Hermann (2011, chapter 6).

manifested in the same reaction, whereas multiple-track dispositions are manifested in multiple ways. It is impossible to indicate all of their specific manifestations. In addition, we can provide neither necessary nor sufficient conditions for their existence (Kemmerling 1975, p. 161). Although intelligent capacities are, like habits, a sort of ‘second nature’, they differ from habits in being inculcated through training (as opposed to pure drill) and being manifested in numerous ways (Ryle 1963, pp. 42 ff.). A further important difference is that someone who exercises a capacity keeps learning and improving his skill. Analogous to this conception of ‘intelligent capacities’, emotional capacities can be seen as ‘dispositions to proper affection’. An agent who has the emotional capacities required for being moral is disposed to feel guilt, shame, remorse, etc. in the appropriate situations.

A possible objection to this view is that the emotional dispositions of a moral person could hardly be taken as capacities.[6] According to the objection, it is possible that someone has the capacity to be properly affected without actually being properly affected. In my view, however, this objection is not plausible. Since we do not experience moral feelings deliberately, someone who is capable of, for instance, feeling guilt, will feel guilt in situations where this is appropriate. While I do not deny that some people manage to suppress moral feelings such as guilt and get rid of them almost completely, I want to stress that those people lose the capacity to have these feelings.

One of Ryle’s examples of *knowing how* is the knowledge of a chess player. Ryle emphasises that knowing the content of the catalogue of chess rules is not equivalent to knowing how to play the game, i.e. to being a competent chess player. Someone who knows all the rules of chess by heart may not be capable of actually playing the game. Moreover, the chess player does not have to recite the chess rules to himself before he makes a move. What the chess player needs is not propositional knowledge, at least not primarily, but *knowledge how*.

Being ‘second natures or acquired dispositions’, competences like knowing how to calculate or how to play chess are both acquired and revealed through practice. The chess player’s competence is revealed in ‘the moves that he makes, or concedes, and in the moves that he avoids or vetoes’ (Ryle 1963, pp. 41 ff.). Only by looking at what he does on the board, over a period of time, can we judge whether or not he knows how to play chess. There is no need to try looking into his head.

It seems that moral competence is in many respects similar to the supposedly less complex competences Ryle discusses. Just like the competences he mentions, moral competence is revealed through what an agent says and, more importantly, through what he does. No one will deny that we cannot conclude from the fact that someone says that he shares our moral views on, say, the evilness of murder that he is morally competent. Such a judgment requires that we take other things into account, such as further judgments he makes, the way he reacts to cases of what are usually conceived of as cases of wrongdoing, and his behaviour towards other people.

[6] I thank an anonymous reviewer for raising this objection.

Just as knowing the chess rules by heart is neither necessary nor sufficient to be a competent chess player, being able to recite a long list of moral precepts is neither necessary nor sufficient to be able to think, feel, and act morally. Knowing the respective rules by heart is insufficient in both cases because having propositional knowledge of a rule and being able to apply that rule in a particular case are two entirely different things. A rule considered by itself does not tell us how it is to be followed. Ryle talks about knowing rules ‘in the executive way of being able to apply them’. This kind of knowledge has to be acquired through practice. It is the result of training, not of learning any rules by heart. Thinking of what the moral or virtuous person knows in terms of *knowing how* is illuminating in that it highlights that being moral requires capacities that are developed in practice. As Matt Stichter points out, it is a benefit of the analogy between skills and virtues that ‘our familiarity with practical skills can give us insight into the development of virtue’ (Stichter 2007, p. 184).[7]

Regarding the question as to whether propositional knowledge of moral principles is needed for being moral, a distinction has to be made between the development of moral competence and the exercise of it once it is fully developed. A morally competent agent does not need propositional knowledge of moral principles to make moral judgments and act morally, and quite a few philosophers would agree with me on this point. The Stoics for example characterize the expert knowledge of the truly virtuous person as ‘the expertise to choose wisely what will be a good move without relying on the instructions’ (Devettere 2002, p. 133). More recently, Hubert and Stuart Dreyfus have argued that at the last stage of ethical development, the stage of ethical expertise, a person is able to judge ‘without appeal to rules and maxims’ what the right thing to do is in a concrete situation (Dreyfus and Dreyfus 1991, p. 237). What is more contested is whether propositional knowledge of moral principles is necessary for *learning* to be moral. Here I disagree with Dreyfus and Dreyfus (see below, Sect. 12.5).

Research in psychopathology supports the view that propositional knowledge of moral principles is not a central characteristic of moral agents. People who suffer from an antisocial personality disorder might not verbally reject the proposition ‘Murder is evil’ and yet show through their actions that they find nothing wrong with murder. As empirical studies suggest, the general reasoning abilities of these people are not impaired, but they lack the capacity to respond with feelings like empathy and guilt and thus do not care about the harmful consequences their actions have for others (see Haidt 2001, p. 824; Blair et al. 2006, p. 18). They ‘primarily lack certain emotions: sympathetic pleasure at another’s happiness, dismay at another’s sorrow, remorse at having brought trouble to another’ (Morton 2004, p. 48).

[7] The comparison between a moral person and a chess player might be objected to insomuch as the former is not only capable of having proper feelings, but actually has proper feelings. However, the same point can be made with regard to the chess player: a competent chess player is not only capable of making the right moves, but actually makes the right moves. Moreover, if to respond in concrete situations with the appropriate moral feelings requires more than just having a capacity, to make the correct moves in a game of chess requires more than just that.

12.4 Virtues and Skills

Ryle acknowledges that in many important respects, 'our knowledge of the difference between right and wrong' is 'much more like a mastery than like the retention of a piece of information'. Thus, both are 'inculcated by upbringing rather than imparted by dictation' (Ryle 1958, p. 149). However, he explicitly rejects the view that the knowledge concerned should be conceived of as a form of *knowing how*, thus apparently dismissing a skill model of virtue.

Before I address Ryle's reasons for rejecting the view that knowledge of the difference between right and wrong is *knowledge how*, a remark on his terminology is required. He speaks both of the knowledge of the difference between right and wrong and of virtue. His use of these terms suggests that he identifies a virtuous person as a person who knows the difference between right and wrong. He thus uses 'knowing the difference between right and wrong' in the broad sense of 'being virtuous', which might explain, at least partly, why he emphasises the differences between knowledge of right and wrong and the knowledge of, for instance, a chess player.[8] In many places, he refers to more than the capacities necessary for being moral. As I will argue below, his emphasis of the differences might also be due to the particular view of skills he endorses in Ryle (1958, 1972).

Let me now address his reasons for thinking that knowing the difference between right and wrong/being virtuous is not a kind of skill. These are:

(i) Our knowledge of the difference between right and wrong 'does not get rusty'. The notions of being out of practice, forgetting and decay are inappropriate to virtues. (Ryle 1958, p. 150.)
(ii) 'For any skill or proficiency it is always possible that a particular exercise of it be both technically first rate and unscrupulous' (Ryle 1972, p. 326).
(iii) To have acquired a virtue 'is not a matter of having come to know how to do anything' (Ryle 1972, p. 330). The virtuous person is not 'good at doing anything' (Ryle 1958, p. 151).
(iv) Virtue is taught differently than skills are; we do not become virtuous through training (Ryle 1972, p. 327).
(v) Learning of the difference between right and wrong does not terminate in knowing, but in 'being so-and-so' (Ryle 1972, p. 330).
(vi) Someone who becomes, for example, inconsiderate does not just lose knowledge, but considerateness (Ryle 1972, p. 331).

Is his first point, which is at the centre of Ryle (1958), plausible? He argues that while our knowledge of how to play chess can get rusty due to a lack of practice, our knowledge of the difference between right and wrong cannot. The notion of being out of practice is said to be inappropriate to virtues. It would be odd to excuse my selfish behaviour by saying that I haven't practised being unselfish for a while,

[8] I thank the anonymous reviewer for pointing this out.

or to say that I forgot the difference between right and wrong. ‘Forgetting’ is not applicable to what we know when we know that difference. According to Ryle, this is because ‘to have been taught the difference is to have been brought to appreciate the difference’. This appreciation includes an ‘inculcated caring, a habit of taking certain sorts of things seriously’ (Ryle 1958, p. 156).

Ryle’s point here is subtle. He argues that ‘we do not keep up our honesty by giving ourselves regular exercises in it’, in order to emphasise that ‘drill[ing] ourselves into good habits and out of bad habits’ should not be assimilated to ‘the exercises by which we prevent our Latin or our tennis from getting rusty’. He does not deny that moral deterioration occurs, but denies that ‘such deteriorations are to be assimilated to declines in expertness, i.e. to getting rusty’ (Ryle 1958, pp. 150 ff.).

I think that Ryle overstates the difference. Firstly, also in the case of many skills it would be odd to speak about forgetting. Think of cases such as knowing how to speak Spanish or knowing how to ride a bicycle. While we *forget* certain words or grammatical rules, we ‘*unlearn’* how to speak Spanish or ride a bicycle (the latter, however, seeming very difficult to ‘unlearn’). Secondly, I would argue that ‘moral deterioration’ is in many respects similar to getting rusty. When someone becomes, let’s say, less courageous it can be due to the fact that he hasn’t been in challenging situations for a long time. It makes perfect sense to say that a lack of practice subsequent to the acquisition of moral or character virtues can lead to a person becoming less virtuous. As Aristotle writes, ‘men become builders by building and lyre-players by playing the lyre; so too we become just by doing just acts, temperate by doing temperate acts, brave by doing brave acts’ (Aristotle 1998, 1103a33–1103b3). It requires ‘some level of routine practice’ to maintain the expertise that has been achieved, and even the fully virtuous person has ‘to work to maintain her virtue’ (Stichter 2011, p. 82). I will return to this point in Sect. 12.6 when I address Ryle’s assertions (v) and (vi). There I will discuss whether learning the difference between right and wrong is *eo ipso* coming to appreciate the difference.

Let me now address (ii) and (iv) together. Ryle rightly points out that the idea of exercising one’s moral competence by acting unscrupulously is self-contradictory. He argues that we have to distinguish between the skills someone has, for example a surgeon, and the end for which he exercises these skills. A surgeon can use his medical skills in order to save a person’s life, but he can also use it to harm that person or even to take a life (see Ryle 1972, p. 326). By acquiring medical skills, we do not *eo ipso* acquire the motivation to use these skills for morally good purposes. From this Ryle concludes that we must learn to be moral in a way that differs from the way we acquire (technical) skills.

My point is that the development of moral competence is in many respects similar to the development of such skills. As pointed out by Aristotle, both involve ‘learning by doing’. They include learning by example, being corrected, being praised and blamed, and so forth. Moral competence differs from other skills in that it is impossible that a particular exercise of it be both technically first rate and unscrupulous. The claim that a particular action is both an exercise of certain moral capacities and unscrupulous is contradictory. Moreover, saying that an exercise of moral competence is technically first rate is misleading, given that being morally competent does not amount to having mastered a particular technique. This, however, does not compel us to sharply distinguish the knowledge characteristic of the moral agent from skills or *knowing how*. Every skill has its peculiarities, and to conceive of moral knowledge as primarily knowledge *how* does not mean to deny the differences between learning to think, feel, and act morally and, for instance, learning to swim. There certainly are a number of differences. Learning to be moral takes much longer than learning to swim or to ride a bicycle. Unlike swimming, cycling, etc., it is nothing we learn in isolation from other things. It is embedded in the broader process of socialization. To a much greater extent than the acquisition of other capacities, the development of moral capacities requires love and care.

What about Ryle's claim (iii) that by having acquired a virtue we have not come to know how *to do* anything, and that the virtuous person is not good *at doing* anything? The claim seems to conflict with Ryle's description of what 'a person who has learned the difference between right and wrong has learned' (Ryle 1958, p. 155), which closely resembles my understanding of moral competence. Ryle describes the virtuous person as having learned to say, do, and feel certain things, and emphasises that all of this is learned together. As I see it, acquiring virtue or moral competence implies that we come to know how to react in certain situations and how to judge them. The virtuous or morally competent agent is good at doing a number of things: at perceiving the morally relevant features of a situation; at judging what the morally right thing to do is in a given situation; at imagining how other people feel; at treating other people with respect and consideration; at responding appropriately to someone in distress; at keeping secrets; at deliberating in situations of moral conflict, and so forth.

In his article 'Can virtue be taught?' Ryle comes to the conclusion that what is at stake is not a form of knowledge at all (v). He criticises Socrates for having thought that 'if virtue can be learned, then … the learning terminates in knowing' (Ryle 1972, p. 330). Ryle is worried that talking about knowledge as he himself did in his earlier paper (Ryle 1958) distorts what we have actually acquired once we have acquired virtue. Is he right and should we therefore dispense with the notion of knowledge altogether? I still believe that his account of *knowing how* is of help in thinking about the moral agent. Ryle claims that learning to be honest results in *being* honest, not in knowing how to do anything (Ryle 1972, p. 330). In my view, however, it results in both. Since telling the truth is not always the right thing to do, learning to be honest involves learning *how to judge* when telling the truth would be wrong. The honest person *is* honest, but being honest involves *knowing how to judge*, in a particular situation, whether it is moral to lie or to tell the truth. I do not claim that being moral is exclusively a matter of skills. In addition to particular skills, a moral person has certain cares and commitments, which, I believe, are acquired in the course of acquiring moral capacities.[9] Learning to be moral results in being moral, which involves knowing how to think, act, and feel morally.

[9] I am grateful to the anonymous reviewer for emphasizing this point.

Finally, Ryle argues that someone who becomes inconsiderate, cowardly, or unjust does not just lose knowledge, but becomes a different kind of person (vi). He says that the heart of that person hardens (Ryle 1972, p. 331). I do not want to deny that Ryle has a point which should be taken seriously. Yet I think that once we consider how someone's heart might harden,[10] it doesn't seem distorting to conceive of that person as also losing certain capacities. A person's emotional capacities might be impaired by negative experiences, deep personal disappointments, or a loss she cannot cope with. She might also, for instance, work in an environment that strongly encourages selfish and inconsiderate behaviour and thus prevents her from using her capacities for empathy, being considerate, etc.

In ancient Greece, it was a common view among virtue ethicists to hold that virtues are similar to skills. For the Stoics, for instance, ethics were 'the most important skill of all, the skill of living' (Devettere 2002, p. 129). As far as Aristotle is concerned, interpreters disagree about whether he endorsed or rejected a skill model of virtue.[11] His frequent use of analogies between skills and virtues supports the view that he merely rejected a skill model of virtue based on an 'intellectualist view of skills' (Stichter 2007, p. 188).[12] According to that intellectualist view, acquiring skills is 'largely a process of acquiring a body of knowledge about how to make a product', which 'can be reduced in large measure to a set of instructions ... that the beginner can learn from the expert' (Devettere 2002, p. 129). Genuine skills are conceived of as having a strong intellectual component, a 'theory behind the skill', which 'the expert is able to teach' (Stichter 2007, p. 185).

Ryle seems to endorse the intellectualist view of skills in Ryle (1958), and to reject it in Ryle (1963). In Ryle (1958), he conceives of the training appropriate to skills as 'technical instruction' (Ryle 1958, p. 153). In Ryle (1963), however, he seems to have adopted the alternative view of skills that Stichter calls the 'empiricist view' (Stichter 2007, p. 188). In this account of skills, the source of expert knowledge is experience. Stichter argues convincingly that Aristotle endorsed a skill model of virtue based on an empiricist conception of skills (Stichter 2007, pp. 189 ff.). Ryle certainly does not endorse such a model, but his position seems compatible with it.

Let me recapitulate the main points. Although we do not say that someone forgot the difference between right and wrong or that he acted selfishly because he is out of practice, being moral involves numerous capacities that can get rusty through a lack of practice. We don't become moral simply by acquiring non-moral capacities, but learning to be moral is in many respects similar to learning to play chess, to calculate, etc. Someone who has learnt to be moral has thereby learnt how to do a

[10] Unfortunately Ryle doesn't say anything about this.

[11] Annas (1995), Bloomfield (2000), and Devettere (2002) hold that Aristotle was the only ancient virtue ethicist who rejected the skill model of virtue. Hutchinson (1988) and Stichter (2007), by contrast, interpret Aristotle as rejecting such a model only if it is based on an intellectualist view of skills.

[12] Stichter calls this view the 'Socratic' view of skills (see Stichter 2007, p. 184).

number of things. This knowledge is closely connected to the kind of person he is. The person who becomes less virtuous or moral does thereby also lose *knowing how*. Finally, thinking of virtue in terms of skills requires an ‘empiricist’ view of skills.

12.5 Learning to Be Moral

Let us now consider the way in which moral capacities are developed. Ryle emphasizes that virtue is taught in a way that differs both from the teaching of facts (‘teaching by dictating’) and from that of skills (‘teaching by training’) (Ryle 1972, p. 326). Both training and teaching virtue involve some conditioning, but the teaching of virtue is said to be in many respects peculiar: ‘our parents reprimanded certain sorts of conduct in quite a different tone of voice from that in which they criticized or lamented our forgetfulness or our blunders’; the serious punishment for moral wrongdoing has a completely different quality from the penalties associated with the violations of conventional rules; we are taught to treat ‘certain sorts of things as of overwhelming importance’ (Ryle 1972, p. 328).

Ryle compares the teaching of virtue with the teaching of tastes (Ryle 1958, pp. 153 ff.). In both cases, we do not only learn a certain difference, for example the difference between virtuous and non-virtuous actions and the difference between good and bad wine, but also to appreciate that difference, i.e. to enjoy good wine, to admire virtuous persons, and to want to drink good wine or to act virtuously. Ryle suggests thinking of virtue in terms of ‘educated tastes’ or ‘cultivated preferences’ rather than in terms of skills (Ryle 1958, p. 151).

Yet there does not seem to be a sharp distinction between skills and educated tastes. One of Ryle’s (1958) examples of the latter is a bridge player. The bridge player ‘had to learn both to play and to enjoy playing bridge’. However, in Ryle (1963), one of the examples for someone who is mainly characterized by skills is the chess player. He doesn’t address the chess player’s ‘keenness’ as opposed to his skills, but he describes the competent chess player’s way of playing the game as exhibiting an appreciation of the difference between good and bad moves as well as a desire to play the game well (Ryle 1963, p. 41). In Ryle (1972), he writes that ‘to become good at bridge or cricket is not the same thing as to become keen on the game, though there is a natural connexion between the two’ (Ryle 1972, p. 331).

As I see it, not only learning to distinguish good wine from bad wine and good literature from bad literature, but also learning to play a particular game, to ride a bicycle, or to swim involves learning to appreciate what is recognized as good or correct. The way in which any person capable of chess wants to play well and shows admiration for particularly good players may not quite be the same as a passion for good wine or literature. Yet the idea that someone learns to play chess but does not aim to play it well or show admiration for very good players, seems rather odd.

A possible objection to this could be that it only shows that *if* someone is motivated to exercise his chess skills, he will be motivated to play well, and not that he is actually motivated to exercise these skills.[13] However, the recourse to educated tastes is of no help here, since educated tastes do not imply motivation in this sense either. A connoisseur, for instance, might, for some reason, not be motivated to drink wine. As far as motivation is concerned, there is a difference between the moral person on the one hand and the chess player and the connoisseur on the other.

Let us now look more closely at how we learn to be moral. Moral education starts in early childhood. It is undertaken not only by one's parents or other primary caregivers, but also by teachers, friends, and ultimately by society as a whole. To a large extent it is not provided by any one particular person. Rather, interactions with others provide a 'training-ground' for the development of moral competence.

I will concretise this picture by contrasting it with the model of moral learning developed by Dreyfus and Dreyfus. Thereby, I will address the role propositional knowledge of rules has in moral teaching and learning. While I generally endorse Dreyfus and Dreyfus's account of ethical expertise, I am not convinced by their view of how that expertise is acquired. According to Dreyfus and Dreyfus, the first stage of moral learning involves learning moral rules. A child is thought to learn these rules prior to being able to judge in a concrete situation whether violating them would be justified. In the example they give, which, as they admit, is 'greatly oversimplified and dramatic', the child initially follows the rule 'never lie' by not lying in any circumstances whatsoever (Dreyfus and Dreyfus 1991, p. 237). Only upon confrontation with a serious dilemma does the child realize that moral rules need to be contextualized.

In the example, the child faces a dilemma we know from Kant: the choice between telling the truth, thereby extraditing a friend to a killer, or lying about where the friend is. It is the experience of regret and guilt after having told the truth and extradited the friend which makes the child realize that whether telling the truth is morally right or not depends on the context. According to Dreyfus and Dreyfus, the next step for the child is to seek maxims like 'Never lie except when someone might be seriously hurt by telling the truth' (Dreyfus and Dreyfus 1991, p. 237). By acting according to such maxims, it will again experience negative moral feelings, and it is only with presumably a lot of experience that the child (which might by then be an adolescent or an adult)[14] would be able to judge, *without appeal to either rules or maxims*, in a concrete situation whether lying or telling the truth is the morally right thing to do. This is the stage of ethical expertise.

What strikes me is that the authors assume that at the very beginning of its moral development, the child is already able to react with the appropriate moral feelings. The authors do not comment on this, but they obviously presuppose that either we are already born with these emotional capacities or these capacities develop somehow prior to the initial stage of moral development (which is an odd assumption). Negative moral feelings play a crucial role in their account of the

[13] Thanks to the anonymous reviewer for raising this point.

[14] Dreyfus and Dreyfus do not say anything about this.

development of ethical expertise, but they do not discuss how the capacity to have these feelings when they are appropriate is itself developed.

How do the relevant emotional capacities develop? It seems plausible to conceive of moral learning as often starting with concrete situations in which our actions meet resistance. The experience of conflict and of being blamed for one's behaviour, as well as the experience of the emotional reactions of others to one's own actions, seems to be important for developing the capacity to feel guilt, shame, remorse, etc. in appropriate situations. The same kind of experiences, however, seems to teach us not to do certain things, such as lying, harming others, cheating, stealing, and so forth. It seems as if we could learn to judge in a particular situation whether lying, cheating, etc. is morally permitted by experiencing numerous situations in which our behaviour is sometimes corrected, we are sometimes praised or blamed, or our behaviour affects others in certain ways. But if this is so, learning the propositional content of a moral rule is unnecessary. As the sociologist Michael Rustin emphasizes, the 'essence of moral learning is not intellectual subscription to abstract precepts, but a process of learning-within-a-situation, from experience and example, in which the implications and effects of feelings and actions can be reflected on with others' (Rustin 1997, p. 87).

Martin Hoffman (2000) describes a form of moral education suitable for the development of emotional capacities: 'induction'. Parents who use this form of moral education attempt, for instance in a case where their child hurts another child, to make him imagine what it would feel like to experience similar harm. As Hoffman points out, induction 'highlights both the victim's distress and the child's action that caused it and has been found to contribute to the development of guilt and moral internalization in children' (Hoffman 2000, p. 10). According to moral sentimentalist and virtue ethicist Michael Slote (2010),[15] inductive training involves even more than 'a parent's deliberately making a child more empathically sensitive to the welfare or feelings of others'. The parents also demonstrate an empathic concern, in this case for the child that got hurt, which can be directly understood by the child. Slote calls this 'a kind of empathic osmosis'. It is a kind of modelling which differs from that advocated by most educational theorists by being non-deliberate and possibly unconscious, and which can also be expected to occur at times where induction is not used and a child simply notices a parent's empathic concern and attitude of care (Slote 2010, p. 20).

Unlike Dreyfus and Dreyfus, I think that the development of emotional moral capacities goes hand in hand with the development of the ability to judge what is morally right in particular situations. Both are developed in the course of concrete experiences, and in this process feelings and judgments mutually affect each other. We should not even conceive of the capacity for moral judgment as a distinct faculty, since it seems to involve, for instance, a capacity for empathy.

[15] Moral sentimentalism is a meta-ethical view according to which morality is somehow grounded in emotions ('moral sentiments' in the terminology of philosophers such as David Hume and Adam Smith). It is an alternative to rationalism, which is the dominant position.

Propositional knowledge of moral rules does not provide any explanatory work in my account. In the course of correcting the child's behaviour, adults presumably also give it general rules such as 'Do not lie' or 'Do not steal'. However, if a child was only told these general rules, we could not expect it to develop a capacity for moral judgment, or to be motivated to refrain from the prohibited types of action. I expect everyone to agree with me that much more is involved in teaching a child that lying and stealing are morally wrong. Yet once we think of these other aspects of moral education, formulating general rules appears unnecessary.

In the answer to the question 'What makes us moral?' induction and modelling play a far greater role than the learning of moral principles. They are ways of advancing, for instance, the development of the capacity for empathy. While modelling is probably involved in the acquisition of all sorts of capacities, induction is peculiar to moral learning. It goes beyond what is involved in the training through which people become competent chess players, speakers of a foreign language, or good writers.[16] Similarly, however, becoming a good writer involves other ways of teaching and learning than becoming a good chess player.

Moral teaching and learning involve a significant amount of training, such as conditioning, i.e. pure drill in addition to arguments, teaching and learning by example, being corrected, being praised or blamed, learning in concrete situations, and gradual improvement through practice. It shares these features with the teaching and learning of languages, games, sports, and so forth. By comparing moral learning with learning how to play chess or how to swim, I do not mean to overlook the peculiarities of each practice and each way of learning. Highlighting the training involved in moral development, I reject the view that learning rules is sufficient, that becoming moral is merely a matter of understanding rationally that some things are right and others wrong, and that we become moral by argument. This leads me to my final point.

12.6 The Incongruity

A philosophical consequence of the view of moral competence presented in this paper is the futility of attempting to convince the amoralist by argument.[17] The amoralist, also called 'moral sceptic', is a philosophical construct that serves to pose what seems to be the biggest challenge a moral theory has to meet. As an ideal type, he is by definition not susceptible to any moral considerations. He shares none of our moral assumptions and his actions reveal no signs of moral competence.

The attempt to convince the amoralist of the authority of moral norms by providing an argument for the rationality of acting morally – an argument that does

[16] I am grateful to the anonymous reviewer for stressing this point.

[17] For the following discussion see also Hermann (2013).

not rely on any moral assumptions – presupposes that the amoralist does not lack any capacities necessary for being moral. Capacities cannot result from the acceptance of an argument. However, a conception of an amoralist who has the necessary capacities for being moral and at the same time refuses to accept moral norms is implausible. I agree with Ryle that 'to have been taught the difference between right and wrong *is* to have been brought to appreciate the difference' (Ryle 1958, p. 156, my italics). Ryle rightly points to the incongruity 'in the idea of a man's knowing the difference between right and wrong but not caring a bit whether he lies, say, or tells the truth' (Ryle 1958, p. 158). Similarly, Aristotle holds that 'the virtue of man . . . will be the state of character which makes a man good and which makes him do his own work well' (Aristotle 1998, 1106a17). It would be impossible to consider a person virtuous who had developed the character states that are the virtues but never behaved in the appropriate way (see Devettere 2002, p. 72).

Likewise, we would not ascribe moral competence to someone who was not moved by any moral concerns. Acquiring moral competence goes hand in hand with the acceptance of moral requirements and the motivation to act morally. As I argue elsewhere, for a morally competent agent, some moral judgments are beyond doubt (Hermann 2011, chapter 4). In addition, it seems that a person who entirely disregarded moral considerations in her judgments and actions would lack certain emotional capacities. Thus, the idea of the amoralist only appears to be consistent if we conceive of him as a sociopath, who surely cannot be cured by argument.

12.7 Conclusion

If we conceive of being moral as involving the exercise of a number of interrelated capacities relevant for moral agency, what makes us moral involves a large portion of training. It is through practice that we learn to be moral. What makes us moral is *not* our desire for happiness, the fear of sanctions, the insight that being moral serves our self-interest or is the only rational thing to do, or any other kind of argument. We become moral by growing up in an environment that enables us to develop the rational and emotional capacities necessary. What characterizes the moral person and enables her to be – and to want to be – moral is not systematic, theoretical moral knowledge or propositional knowledge of moral principles, but a number of interacting capacities that are developed in practice. The dominant rationalistic views overemphasize the role of moral principles, conscious deliberation, and theoretical moral knowledge. As moral persons, we have mastered moral practices.

Ryle highlights the differences between the knowledge that characterizes the moral agent and skills like that of a chess player, but he also overstates some of them. Thus, a lack of practice of our moral capacities, for example of the capacity for empathy, affects our moral competence, as does continuous practice of these capacities. While we keep on refining our moral competence as adults – a feature which moral competence shares with other skills – we can also become less morally competent through a lack of practice as a consequence of negative experiences, certain working environments, or (voluntary) retreat from society. If we conceive skills as acquired mainly through experience, a comparison between skills and virtues is illuminating.

Let me conclude with the normative claim that ideally, a growing awareness of the importance of moral capacities – and of emotional capacities in particular – will lead to greater efforts in creating an environment that enables children to become morally competent agents.

References

Annas, J. 1995. Virtue as a skill. *International Journal of Philosophical Studies* 3(2): 227–243.

Aristotle. 1998. *The nicomachean ethics.* Trans. D. Ross. Oxford/New York: Oxford University Press.

Blair, J., A.A. Marsh, E. Finger, K.S. Blair, and J. Luo. 2006. Neuro-cognitive systems involved in morality. *Philosophical Explorations* 9(1): 13–27.

Bloomfield, P. 2000. Virtue epistemology and the epistemology of virtue. *Philosophy and Phenomenological Research* 60(1): 23–43.

DePaul, M. 1993. *Balance and refinement: Beyond coherence methods of moral inquiry*. London/New York: Routledge.

Devettere, R.J. 2002. *Introduction to virtue ethics. Insights of the ancient Greeks*. Washington, DC: Georgetown University Press.

Dreyfus, H.L., and S.E. Dreyfus. 1991. Towards a phenomenology of ethical expertise. *Human Studies* 14(4): 229–250.

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

Hermann, J. 2011 *Being moral: Moral competence and the limits of reasonable doubt*. Dissertation, European University Institute, Florence.

Hermann, J. 2012. Man as a moral animal: Moral language-games, certainty, and the emotions. In *Language, ethics and animal life: Wittgenstein and beyond*, ed. M. Burley, N. Forsberg, and N. Hamalainen, 111–123. New York: Bloomsbury (formerly Continuum Publishing).

Hermann, J. 2013. Die Praxis als Quelle moralischer Normativität. In *Moral und Sanktion: Eine Kontroverse über die Autorität moralischer Normen*, ed. E. Buddeberg and A. Vesper, 137–166. Frankfurt/New York: Campus.

Hoffman, M.L. 2000. *Empathy and moral development. Implications for caring and justice*. Cambridge: Cambridge University Press.

Hughes, G.J. 2001. *Aristotle on ethics*. London/New York: Routledge.

Hutchinson, D.S. 1988. Doctrines of the mean and the debate concerning skills in fourth-century medicine, rhetoric, and ethics. *Apeiron* 21: 17–52.

Kemmerling, A. 1975. Gilbert Ryle: Können und Wissen. In *Grundprobleme der großen Philosophen – Philosophie der Gegenwart*, ed. J. Speck, 126–166. Göttingen: Vandenhoeck & Ruprecht.

Morton, A. 2004. *On evil*. New York/Oxon: Routledge.

Rustin, M. 1997. Innate morality: A psychoanalytic approach to moral education. In *Teaching right and wrong: Moral education in the balance*, ed. R. Smith and P. Standish, 75–91. Staffordshire: Trentham Books.

Ryle, G. 1958. On forgetting the difference between right and wrong. In *Essays in moral philosophy*, ed. A.I. Melden, 147–159. Seattle: University of Washington Press.

Ryle, G. 1963. *The concept of mind*. Harmondsworth: Penguin.

Ryle, G. 1972. Can virtue be taught? In *Education and the development of reason*, ed. R.F. Dearden, P.H. Hirst, and R.S. Peters, 323–332. London/Boston: Routledge and Kegan Paul.

Slote, M. 2010. *Moral sentimentalism*. Oxford/New York: Oxford University Press.

Stichter, M. 2007. Ethical expertise: The skill model of virtue. *Ethical Theory and Moral Practice* 10: 183–194.

Stichter, M. 2011. Virtues, skills, and right action. *Ethical Theory and Moral Practice* 14: 73–86.

Chapter 13
Why Mental Disorders Can Diminish Responsibility: Proposing a Theoretical Framework

Gerben Meynen

13.1 Introduction

The question ‘What makes us moral?’ might suggest that we can take our moral responsibility for granted, and that we merely have to answer the question: *what* is it that makes us moral? Yet, although human beings are usually considered morally responsible agents, we do not take this responsibility for granted. For example, we take a person’s age into account: (young) children are usually not considered truly morally responsible for their actions (Widerker and McKenna 2003). But in adults too, there may be reasons not to consider them morally responsible agents. This chapter considers one of such reasons: the influence of a mental disorder on a particular action.[1]

The significance of a mental disorder for ascribing moral agency is generally accepted (Fields 1987; Mele 1995; Elliott 1996; Edwards 2009). For instance, consider an elderly patient who has been admitted to a general hospital because of a hip fracture. On the first night in the hospital he suffers from a delirium and he rudely accuses his doctor and several nurses of conspiring against him. He even punches some of the staff members. As long as the patient’s behaviour is understood within the context of a mental disorder (a delirium in this case), the patient is unlikely to be blamed for his unjustified accusations and even for hitting personnel. The view that mental disorders can excuse is also reflected in criminal law via the insanity defence

[1] In this chapter, I will not address the question whether or why, in general, it is justified to hold people responsible for their actions; I will assume that human beings are generally responsible for their actions, and that at least in some cases mental disorders can undermine such responsibility.

G. Meynen (✉)
Faculty of Philosophy, VU University Amsterdam, Amsterdam, The Netherlands
e-mail: g.meynen@vu.nl

B. Musschenga and A. van Harskamp (eds.), *What Makes Us Moral? On the capacities and conditions for being moral*, Library of Ethics and Applied Philosophy 31,
DOI 10.1007/978-94-007-6343-2_13, © Springer Science+Business Media Dordrecht 2013

(Elliott 1996). Yet, the example of this elderly patient makes clear that the relevance of this idea extends beyond criminal scenarios and courtrooms.

Although the view that mental disorders can excuse is widespread, there is an on-going debate about why it is that, in certain circumstances, mental disorders diminish responsibility (Elliott 1996; Kinscherff 2010; Perring 2010; Pouncey and Lukens 2010). Interestingly, some people do not accept that mental disorders can affect responsibility at all. Recently, Pouncey and Lukens (2010) pointed to the tension between the recovery movement in psychiatry on the one hand and ethicists, psychiatrists and lawyers on the other regarding the widespread idea that mental disorders may exculpate. Although ethicists, psychiatrists, and lawyers agree that mental disorders in some instances excuse, the recovery movement's message has been 'that persons with severe mental illness can and should be responsible for their own life choices' (Pouncey and Lukens 2010). This powerful message was helpful to diminish the stigma of mental illness. Yet, by 'deliberately emphasizing the capabilities of persons with mental illness for self-determination, recovery advocates leave unaddressed important questions about how, when, and to what extent mental illness can limit a person's capacity to make sound choices, or even her moral accountability' (Pouncey and Lukens 2010, p. 94). In fact, Pouncey and Lukens (2010, p. 103) state that 'the question of agency in severe mental illness is fundamental, but neither the recovery movement nor bioethics has devoted much attention to it, either severally or jointly', and they suggest that efforts should be made to clarify and articulate why exactly it is that mental disorders sometimes affect moral agency.

The purpose of this paper is to develop a theoretical framework that explains why, in certain instances, a mental disorder excuses a person for a harmful action. Since it is meant to facilitate actual discussions within society, e.g. between psychiatrists, lawyers and clients – thereby reaching beyond the borders of academic philosophy and ethics – the model should consist of components that have a *prima facie* relevance and plausibility with respect to psychiatric features and symptoms. In other words, the model should neither require several argumentative steps nor commitments to certain philosophical positions before one can explain why a person is exculpated.[2] The framework should enable straightforward translations from particular psychopathological symptoms to the model's components. In principle, holding a person responsible for a harmful action can result in either praise or blame; in this paper, however, I focus on blame (or rather excuse).

One could suppose that what excuses a person is that the mental disorder influenced the action. But this answer is not very helpful. For instance, a crime may be influenced by a person's hate, but the mere fact that hate influences the act does not excuse that person. So, the basic question is: What is so special about the influence of a mental disorder on human behaviour and agency that explains why

[2] Given this approach, in this paper I am not committed to a particular account of moral responsibility, like the one suggested by Fischer and Ravizza (1998). Such commitment would render the model only applicable to those supportive of – and familiar with – this view.

we excuse that person? Some have answered this question by referring to 'free will'. In their view, the reason that mental disorders can diminish legal responsibility is by affecting a person's free will (Reich 2005; Stone 2008).[3] In a previous paper I explored the relationship between the concept of 'free will' and mental disorder (Meynen 2010). Three senses of 'free will' were considered and I examined how they can be related to (various kinds of) psychopathology.[4] In the present paper I will acknowledge the relevance of 'free will' in this context (Sect. 13.2), but I will conclude that although 'free will'-related elements are helpful in explaining why mental disorders excuse, they are *not sufficient*. Additional factors, namely urges (Sect. 13.3), false beliefs (Sect. 13.4), and moral insensitivity (Sect. 13.5) are needed in order to straightforwardly explain and communicate why mental disorders sometimes excuse.[5]

The model in fact *combines* several concepts that have been identified in the literature as responsibility-relevant features undermined by specific mental disorders. In that sense, the model's four components – free will, urge, false belief, moral insensitivity – are not meant to be new or original.[6] This paper intends, however, to show that in so far as one aims at straightforward explanations with respect to the full range of mental disorders, firstly, *each* of these factors (free will, urge, false belief, and moral sensitivity) is relevant; secondly, the relevance of each factor is *limited*; therefore, a theoretical model comprising *less* than these four elements is deficient. To my knowledge, so far no account integrates these four elements. The model could facilitate the dialogue between those who do and do not think that mental disorders can excuse – either in general or in particular cases.

The four factors – free will, urges, false beliefs, and moral insensitivity – can have different meanings. For instance, with respect to free will, there is no consensus about its definition. Therefore, depending on one's position on free will, the factor free will could (partially) overlap with, for example, the factor 'urge'. So, depending on the specific meaning attached to each of the four factors, there may be some redundancy in the proposed model. This will be addressed in some more detail in Sect. 13.6.

[3] Morse (2007) for instance, criticizes this view.

[4] Derived from Walter (2001), see Sect. 13.2 of the present chapter.

[5] Whereas the focus of the paper is to explain our moral responses, the analysis is informed by discussions on moral responsibility as well as legal responsibility. This is in line with, e.g., the accounts by Pouncey and Lukens (2010) and Elliott (1996) who do not make a strict distinction between the moral and legal domain as far as relationship between responsibility and mental disorder is concerned. Notably though, jurisdictions vary considerably worldwide in the exact way they approach 'criminal responsibility'. Still, there appears to be a profound and shared view reflected by these different legal systems: that somehow mental disorders can excuse. The present paper tries to build a framework from an ethical perspective that could also inform the legal/forensic debate.

[6] They feature in various forms in legal and ethical debates on criminal and moral responsibility of people suffering from a mental disorder (see the next sections for references). Often an attempt has been made to understand exculpation using (only) one overarching concept, like free will, or irresistible impulse. For an overview within the context of the insanity defence, see Elliott (1996).

13.2 Factors Related to Free Will

Often, the idea that mental disorders can excuse a defendant is thought to be crucially related to free will (Stone 2008; Juth and Lorentzon 2010). This might not seem surprising since many moral philosophers take free will to be central in ascribing responsibility (Kane 2002; McKenna 2009). So, in principle, given the strong relationship between the concept of responsibility and free will, free will could be helpful in explaining why the responsibility of a defendant can be suspended by mental disorder.

Still, the debate in forensic psychiatry on the conceptual grounds for the insanity defence has been troubled by the problematic nature of free will; often, in philosophy as well as in (neuro)science, the very existence of free will is contested (Kane 2002). Within the context of the present paper, however, I intend not to take a specific position on the metaphysical issues surrounding free will.[7] The concept of free will and notions attached to this concept are used only as far as they don't imply a commitment to one particular metaphysical position.[8] Basically, I will take 'free will' to refer to certain agency-related capacities.

What concept of free will could be of use for this paper? Various views of free will are present in the current philosophical debate (Kane 2002). My approach is to discuss three senses or aspects of free will based on an account suggested by the philosopher and psychiatrist Henrik Walter (2001).[9] He proposes to take three elements as key players in the philosophical discussion on free will: (1) acting for reasons, (2) being able to do otherwise, and (3) being the source of an action. In this section I will indicate to what extent these three elements can be helpful to explain why certain features of mental disorders are sometimes considered grounds for excuse.

Firstly, let us consider the element of *acting for reasons*. I use this 'acting for reasons' in a broad sense: the action should have been sensitive to (or based on) reasons (Muller and Walter 2010).[10] In the neuropsychiatric Tourette syndrome, people may do things – e.g., say nasty words – without any reason for that behaviour:

[7] Such as compatibilism and (hard) incompatibilism, and the specific notions of free will that have been developed within the context of these positions (Kane 2002).

[8] It is virtually impossible to talk about free will in a way that would not conflict with any of the many metaphysical positions on free will (Kane 2002). Since free will is central to many ethical as well as forensic psychiatric considerations of moral responsibility and to the effects of mental disorder on responsibility, I take the concept of free will – in spite of the many metaphysical complexities – as the first factor to examine in order to find out why mental disorders sometimes excuse.

[9] See Meynen (2010) for a more detailed exploration of the relationship based on the account suggested by Walter (2001). With some variations, in this paper, I will follow the same line of thought as in Meynen (2010).

[10] I do not take these reasons to be, for instance, 'reasonable', or 'rational', or 'the right reasons', or the result of a perfect reflective process or of an infallible perception apparatus. Of course, the reasons *may* be the result of such perfect processes, and the reasons may be 'reasonable', or

the tics may come without any motivation in terms of reasons for that utterance or movement. If a person diagnosed with Tourette's insults another person due to a tic, we might say: 'He didn't do it for *any reason* at all, *it just happened*', and therefore we may not consider this an act for which the person can be justifiably held responsible.[11] So, not acting for reasons appears to be relevant to excuse due to mental disorders.

Secondly, mental disorders may undermine a person's *ability to act or choose otherwise*. This notion has been central in the metaphysical debates on free will: is free will possible in a deterministic world, i.e. a world without alternative possibilities? In this paper I aim to avoid strict metaphysical notions and discussions, which is especially challenging with respect to the notion of 'alternative possibilities' (see Meynen 2010). Still, some psychiatric phenomena may provide compelling examples of conditions in which alternative scenarios are blocked. For instance, consider certain voices (auditory hallucinations) as they may occur in schizophrenia. A particular kind of voices that these patients may hear are 'commanding' voices (Braham et al. 2004). Now, usually patients have the freedom not to obey these voices, but there are instances where the nature of these voices is such, that the patient cannot but obey them; no 'alternative possibilities' are left to the patient. If a patient commits a crime because voices of such a commanding nature occurred, we may well excuse the patient, because, apparently, he could not but act on these voices. So, the notion of 'alternative possibilities' enables one to explain some cases in which mental disorders excuse as well.[12]

Thirdly, people suffering from a mental disorder may not be considered to be the *real source* of the action. For instance, Peter Strawson, in his seminal paper on reactive attitudes, paid attention to the moral reactive attitudes we may find ourselves to have in response to the actions influenced by mental conditions (Strawson 2003). In response to a person who performed such an action we may say: 'He wasn't himself'. For instance, a person who has involuntarily taken a drug (like cocaine) might not be considered to be himself: he still does a lot of things, contemplates a lot of things, makes decisions for reasons etcetera, but it is just not *him* (at least not the way people have known him for the last 15 years). Now, in case this person performs a harmful action, we might excuse him because he himself wasn't the *genuine source* of the action. So, the concept 'being the source of an action', also helps to explain some cases in which mental disorders excuse.

The gist of the discussion so far is that (each of the three senses of) free will can be helpful in explaining and communicating why it is that we excuse a

'rational/right', but the aspect of free will that is considered here merely brings forward that the action was responsive to or based on reasons (no matter their specific nature).

[11] Surely, the tic may be influenced by – or a response to – another person's behaviour or certain features of the situation. But the tic itself is not generated 'for a reason' in the sense our behaviour usually is. Meanwhile, not all tics are performed involuntarily according to patients themselves, see Lang (1991), Verdellen et al. (2008).

[12] Note that in the Tourette case, people might also want to refer to the lack of alternative options with respect to tics in order to explain why they excuse a Tourette patient for a particular action.

particular person who suffers from a mental disorder when violating a moral or legal obligation. In fact, we might disagree about which element of free will helps *best* to explain exculpation in individual cases. Probably more than one of them can be applied (this should not be surprising, since the three elements are not unrelated). Still, although these three factors appear to provide us with a substantial conceptual toolkit for explaining excuse due to mental disorder, we can ask ourselves: Does this differentiated approach to free will help to clearly and straightforwardly explain *all* cases in which we excuse a person with a mental disorder who violated a moral or legal norm? My answer is that other factors should be considered as well. In fact, given the diversity of mental disorders and their symptomatology, three other, separate factors seem to be required for the framework: urge (next section), false belief (Sect. 13.4), and moral insensitivity (Sect. 13.5).

13.3 Urges

Mental disorders may result in extreme urges (I take urges to be on a par with impulses in this analysis). As it seems, responsibility can be undermined by such extreme urges – at least partially. Notably, *irresistible* impulses are not the topic of this section, because such irresistible impulses, in my understanding of the concept, undermine – or bypass – free will: if an impulse is so strong that it is irresistible it appears that there are no 'alternative possibilities' open to the patient other than to act according to the impulse. Therefore, such irresistible urges have to do with the subject matter of the previous section (second sense of free will, above).[13] The phenomena referred to in the present section, however, are extreme but not irresistible urges. In extreme but resistible impulses, the capacity for decision making itself is uncompromised. For even an (extreme) urge by itself doesn't suspend the fact that a person can act for a reason, or still has alternatives, or is himself the source of the action. In other words, based on the three senses of free will, these people can be considered *free not* to give in to such urges and impulses.

Kleptomania might be a good example of a disorder accompanied by an extreme urge. At least, the DSM-IV criteria are compatible with the fact that the urge to steal is in principle resistible, though very strong.[14] Within the context of this paper, let us

[13] Irresistible impulses have been considered relevant for the insanity defence, see Elliott (1996), the Irresistible impulse test in criminal law.

[14] DSM-IV criteria for kleptomania are: A. Recurrent failure to resist impulses to steal objects that are not needed for personal use or for their monetary value. B. Increasing sense of tension immediately before committing the theft. C. Pleasure, gratification, or relief at the time of committing the theft. D. The stealing is not committed to express anger or vengeance and is not in response to a delusion or a hallucination. E. The stealing is not better accounted for by Conduct Disorder, a Manic Episode, or Antisocial Personality Disorder (American Psychiatric Association 1994). Notably, these criteria do not state that the impulses are irresistible; they merely state that there is recurrent *failure* to resist.

consider kleptomania to be a condition in which people experience a massive urge to steal. Notably, as it appears, the conditions of ‘free will’ described in the previous section are, basically, still intact. Yet, should the person concerned give in to the urge to steal, for example, sunglasses, then we might (partially) withhold blame, because we know it is *so much harder* for such a person not to steal the sunglasses than it would be for us. (Still we might blame her for going to the shop in the first place, at least if she knows that she is particularly vulnerable to such an urge in this shop.) Often, mental disorders may result in quite extreme urges, like the urge to escape from a fearful situation in anxiety disorders. We may take these urges into account when assessing a person’s responsibility after that person violated a moral or legal norm. So, (massive) urges can be a component of the conceptual framework developed in the present paper.

Meanwhile, some people might still want to explain our intuitions in these cases of extreme urges in terms of *lack of alternative possibilities* (so in terms of a free will-related factor). Consider, however, cognitive behavioural treatment of an anxiety patient. Such patients are likely to experience extreme fear and a massive urge to flee from a frightening situation. Yet, the fact that, in principle, alternative possibilities *are* open to this patient is the cornerstone of behavioural therapy for anxiety disorders. In (cognitive) behavioural therapy of anxiety disorders, patients explore and pursue *other* behavioural responses to their fears than giving in to the *almost* overwhelming urge to flee from that particular fearful situation (for instance, in case a person experiences a panic attack in a supermarket) (Olatunji et al. 2010). Behavioural therapy, in general, will encourage patients *not* to give in to anxiety-based urges but to expose themselves to what they fear. Denying that patients suffering from such disorders have alternative possibilities open to them would also deprive them of an (effective) form of therapy.

13.4 False Belief

The relevance of epistemic factors in addition to free will in ascribing responsibility is well established in moral philosophy, like *not knowing* about a certain state of affairs or certain alternative options (O’Connor 2010). In order to bring forward the relevance of this epistemic factor in general, let us consider Jim, a person not suffering from any mental disorder. He is convinced that someone in the shop has a gun hidden on his person and from this person’s behaviour it is evident to Jim that he is about to perform an armed robbery. Jim, in a heroic mood, throws this person forcefully to the floor; the person involved being rather unfortunate in breaking his arm. Yet, Jim turns out to be completely mistaken – the person was not at all planning to rob the shop. Jim is accused of a crime (assault). Yet, Jim might be *excused* because he acted on a *false* belief. For as long as we think that he wasn’t reckless or that he wasn’t to blame for having this particular false belief, he may be excused. In this case a false belief exculpates Jim. Just to be sure, we can ask: Did Jim act of his own ‘free will’? I think there is little reason to doubt such a thing in the sense that Jim acted as he had been doing all day – ‘freely’ – and his action in the shop was not different from how he reached other decisions about behavioural options that day. In fact, he acted for a reason, was able to consider different options, he was the genuine source of the action (and he didn’t act on a pathological urge either). However, in this particular case he acted on a false belief (and given this belief, his action was not inappropriate).

In Jim’s case his false belief wasn’t the result of a mental disorder. But mental disorders, especially psychotic disorders, may induce false beliefs in patients. In fact, with respect to the question why mental disorders may excuse a person such false beliefs appear to be highly relevant (see below), and therefore epistemic factors are the third element of the model.[15] These false beliefs may be due to delusions (Bentall et al. 2009; Bortolotti 2010) or hallucinations (Sadock and Sadock 2005), or both. For instance, due to a delusion a person may come to believe that someone is on the verge of attacking him. In ‘self-defence’, he attacks the other person.[16] If someone else, who does not suffer from a mental disorder, would have the *same belief* he might well have acted similarly out of self-defence. In fact, anyone could act in such a way in case a mental disorder produced such a false belief. Therefore, if we indeed excuse this person, referring to the false belief is likely to provide us with the clearest explanation of *why* we excuse the person in this particular situation. We withhold blame, then, not on the grounds of the absence of ‘free will’, nor because of some extreme urge, but primarily because of the presence of a *false belief* on which he acted (and which was the result of a mental disorder).

That epistemic issues are important is also reflected by the most influential legal insanity rule, the M’Naghten Rule. This is how it is formulated: ‘At the time of committing the act, the party accused was laboring under such a defect of reason, from disease of the mind, as not to know the nature and quality of the act he was doing; or if he did know it, that he did not know what he was doing was wrong’ (Elliott 1996, p. 11). It is clear that the rule is not primarily about free will-related issues, like doing things for reasons, having alternative possibilities, or being the genuine source of the action. And it isn’t about extreme urges either – none of the concepts discussed in Sects. 13.2 and 13.3 can be found (directly) in this rule. The rule meanwhile brings forward that something else matters: knowledge or rather, lack thereof (of course, one can ask further questions about the nature of the knowledge, but it is nevertheless first of all about ‘knowing’).[17]

Let us take a closer look at the historical figure Daniel M’Naghten whose case led, in the end, to the M’Naghten Rule. Daniel M’Naghten, a Scotsman, believed

[15] Within the context of mental disorder, however, it is not primarily the *absence of correct* knowledge but the *presence of false* belief that appears to be most relevant with respect to diminished responsibility.

[16] Notably, a person is not responsible for this delusional belief; it is the result of a mental disorder he could not have avoided, and for which he is not held responsible (Edwards 2009).

[17] However, as indicated at the beginning of this section, mental disorders are more likely to lead to *distorted* knowledge than to a mere *lack* of knowledge (like in amnesia).

that the Tories were persecuting him and he planned to kill the Tory prime minister because of that. Yet, in his attempt to kill the prime minister, he killed a secretary instead. Eventually, Daniel M'Naghten was acquitted because of insanity. Given his case, the reason Daniel M'Naghten was excused did *not primarily* have to do with any of the free will-related topics or with specific urges. They may be relevant, of course, but something else seems to be patently wrong with Daniel M'Naghten: his beliefs.[18]

13.5 Moral Insensitivity

So far, we have taken for granted that people have moral sensitivity. In other words, we took for granted that people have the capacity to be open to moral feelings, responses, etcetera. If a person lacked such moral responses in the first place, it is difficult to see how she could be considered morally responsible, for the moral domain does not appear to be accessible to such a person. Discussions about such moral insensitivity or amorality have evolved and revolved around the moral status of the psychopath (Fine and Kennett 2004; Haji 2010; Kinscherff 2010; Litton 2010; Malatesti and McMillan 2010). The nature and moral status of the psychopath remains controversial. Is he or she the prototype of an evil being, or in fact an amoral creature, not at all responsible for the many terrible things he or she may do? Nowadays, a considerable part of moral philosophers and scientists appear to be in favour of the latter view (Malatesti and McMillan 2010). Since the present paper aims at providing a conceptual framework that enables us to communicate our moral views about people suffering from a mental disorder, a category accommodating the alleged incapacity of the psychopath cannot be left out. Within the present framework, a psychopath could be interpreted as someone who acts for reasons, is perfectly able to contemplate different options, acts authentically (the three factors related to free will), is not subject to extreme urges that put pressure on his decisions, and finally, is not the victim of false beliefs. Yet, he lacks moral sensitivity, and, therefore we may withhold blame. Notably, also after brain damage, a person's moral sensitivity might be diminished (Cf. the famous case of Phineas Gage). In other words, other conditions than psychopathy, at least in principle, could lead to diminished 'moral sensitivity'.[19]

[18] Of course, more information about the case would be necessary in order to reach a final decision on these matters. Meanwhile, we might still think that M'Naghten's response was blameworthy: one should not kill the prime minister, even not in case one is convinced that he causes one nothing but trouble.

[19] Still, a vivid debate is going on about how to understand the specific capacity that is lacking in psychopaths. According to some, the problem psychopaths face has to do with certain 'moral beliefs' (Maibom 2008). Then, this fourth element of the proposed framework may in fact merge into the third element: (false) beliefs.

13.6 A Fourfold Framework

Based on the analysis so far, we have the global picture:

1. A person violated a legal or moral norm.
2. Looking more carefully at what happened, we may ask: Was a *mental disorder* present such that:
    (a) one or more of the three senses of free will were affected, and/or that
    (b) an extreme urge or impulse was present, and/or that
    (c) the person's beliefs were significantly distorted, e.g., via delusions or hallucinations or both, and/or that
    (d) the person's moral sensitivity is suspended or absent?

Answering 2(a)–(d) should provide us with the element(s) needed to explain why it is that we (partially) withhold blame in certain cases when a mental disorder was present at the moment a person violated a moral norm. To be sure, I do not claim that referring to the factors of the conceptual framework will result in agreement about a person's moral responsibility. It should, however, help to clarify and straightforwardly communicate why we excuse a person in a particular case, even if that means that differences of opinion become more visible.

The fact that the framework consists of four domains or areas implies that accounts aiming to explain why mental disorders excuse that focus on *one* of these (like, apparently, the M'Naghten Rule) are bound to be too narrow to accommodate our moral intuitions.[20]

[20] Christian Perring has provided an insightful account of the concept of mental disorder in which the idea of 'involuntariness' is central. He writes (Perring 2004, p. 489): 'I will propose and test the thesis ... that all behavioral symptoms of mental disorders must be involuntary.' He understands involuntariness to come in three different forms (Perring 2004, p. 496): 'To summarize, I have found three ways in which we can count a form of behavior as involuntary: (a) It is the result of an irresistible craving or overpowering fear. (b) It is the result of an aberrant and temporary desire external to a person's true personality. (c) It is the result of a delusion. I am proposing involuntariness of all symptoms as a necessary condition of mental disorder, not a sufficient one.' Although there are several significant differences between his approach/proposal and mine, there are similarities as well, mainly with respect to distinguishing between delusions (although I prefer false beliefs, which includes false beliefs arising from hallucinations as well as delusions) and not being the source (b) of the action. The proposed model also contains more (and different) elements than the one proposed by Steve Matthews (2004), who basically addresses the legal insanity defence. He proposes to leave out any reference to mental disorders, 'referring instead to a defence of failed agency'. According to Matthews (2004, p. 452), 'the test of responsible agency ... is failed if any one of the following three conditions is satisfied: (a) the person lacked the capacity to understand the nature of what he or she was doing; or (b) the person lacked the capacity to understand that what he or she was doing was wrong (that is, the person's conduct was insufficiently reasons-responsive, constitutively speaking, to conventional, moral or legal codes of behaviour); or (c) the person was unable to control his or her conduct.' As is clear from this quotation, Matthews conceives of the insanity defence in terms of circumscribed incapacities.

Symptomatology of mental disorders is extremely diverse. For instance, during a delirium, the person's behaviour may simultaneously be influenced by paranoid delusions, hallucinations, incoherence, and extreme urges. In other words, several aspects of the disorder *together* may have to be taken into account if one tries to explain why a person is excused in a particular situation. Meanwhile, in practical cases it is likely that one of them will be most helpful.[21]

Although it is impossible, within the context of this paper, to show that the proposed model covers *all* mental disorders, I suggest that the four factors are indeed sufficient to explain instances of excuse with respect to the entire spectrum of psychiatric illnesses. But, of course, it might be that other factors are relevant as well. In principle, the model should be flexible enough to be extended with other components.

As already mentioned, the framework is, depending on the interpretation of the distinctive factors, probably somewhat redundant. Based on certain interpretations of the components, it should be possible, therefore, to conceptually reduce the number of components. For instance, an attempt can be made to understand the component of false beliefs in terms of alternative possibilities (so, a sense of free will). This can be done as follows: the false belief (decisively) influences the options that a person considers open to him. In case of the delusionary belief that your neighbour carries a dangerous weapon, the options one perceives open to one are very much influenced. In that sense the *nature* of the options perceived is changed. However, the person's ability to choose between options in itself doesn't seem to be affected (cf. Jim's case). There are still options available, although they are distorted (so, this case is different from a commanding voice which cannot be disobeyed). In Sect. 13.3, I pointed to the *relevance of acknowledging* such behavioural alternatives in mental conditions, e.g. within the context of cognitive behavioural therapy for anxiety disorders. So, on a conceptual level, it is likely that arguments can be developed showing that some of the framework's components are related, or that they can be reduced to one another. Meanwhile, from a psychiatric or clinical point of view, it is helpful to *distinguish* between different factors. As stated in the introduction, I aim at developing a framework consisting of components that have a *prima facie* relevance and plausibility that directly reflects (clusters of) psychiatric symptomatology; it should not require various argumentative steps or commitments to certain philosophical positions or views to explain one's moral intuition. Therefore, although the conceptual framework offered in this paper may consist of elements that are not unrelated, distinguishing between them could still help given the purpose of explaining and communicating our intuitions.

[21] One can argue that it is relevant these that false delusional beliefs were acquired 'unfreely'. For if the person had *freely* chosen to believe the contents of the delusion then the person would be considered responsible. Yet, my response to this objection is that the fact that the delusional false belief has been acquired involuntarily is already entailed by the fact that we say that it was due to a mental disorder, because, as Edwards points out, people are not considered responsible for the disorder (Edwards 2009).

13.7 Conclusion

The purpose of this paper was to develop a conceptual framework, guided by actual psychiatric symptomatology, which helps to explain why we excuse a person who is suffering from a mental disorder while performing a particular harmful action.

In discussions on why it is that mental disorders exculpate, 'free will' has been mentioned as the crucial factor. In this paper I have been open to the relevance of free will-related concepts to explaining our moral intuitions with respect to mental disorder – leaving aside metaphysical issues and controversies. In fact, I took 'free will' as a starting point for the proposed model. Free will has different meanings to different people, therefore, like before (Meynen 2010), I articulated three senses or aspects of free will. Distinguishing between these three elements – instead of using the general concept of free will – enables one to be specific about what it is about the disorder that is morally salient in a particular case. Although these three elements appear to be helpful in explaining part of the cases in which we excuse a person due to the influence of a mental disorder, further factors are needed as well: mental disorders may result in extreme urges, false beliefs, and, finally, a lack of what I call 'moral sensitivity'. Apart from free will, these three factors may be helpful to explain our intuitions in concrete cases. This implies that no single factor (free will, urge, false belief, or moral insensitivity) has sufficient explanatory power to communicate and articulate why we exculpate persons with a mental disorder who violated a moral norm. Hopefully, the distinctions arrived at can facilitate the dialogue between those who do and do not think that a mental disorder excuses a person in a particular case. Such a dialogue could, in turn, lead to adjustments or additions to the proposed framework.

What makes us morally responsible agents? Within the context of the present chapter we can answer this question as follows: at least, the absence of a mental disorder influencing a person's behaviour in one of the four ways identified by the framework.

Acknowledgment I thank Alan Ralston, MD, for his helpful comments.

References

American Psychiatric Association. 1994. *Diagnostic and statistical manual of mental disorders-IV*. Washington, DC: American Psychiatric Association.

Bentall, R.P., G. Rowse, N. Shryane, P. Kinderman, R. Howard, N. Blackwood, R. Moore, and R. Corcoran. 2009. The cognitive and affective structure of paranoid delusions: A transdiagnostic investigation of patients with schizophrenia spectrum disorders and depression. *Archives of General Psychiatry* 66: 236–247.

Bortolotti, L. 2010. *Delusions and other irrational beliefs*. Oxford: Oxford University Press.

Braham, L.G., P. Trower, and M. Birchwood. 2004. Acting on command hallucinations and dangerous behavior: A critique of the major findings in the last decade. *Clinical Psychology Review* 24: 513–528.

Edwards, C. 2009. Ethical decisions in the classification of mental conditions as mental illness. *Philosophy, Psychiatry, and Psychology* 16: 73–90.

Elliott, C. 1996. *The rules of insanity. Moral responsibility and the mentally ill offender*. Albany: State University of New York.

Fields, L. 1987. Exoneration of the mentally ill. *Journal of Medical Ethics* 13: 201–205.

Fine, C., and J. Kennett. 2004. Mental impairment, moral understanding and criminal responsibility: Psychopathy and the purposes of punishment. *International Journal of Law and Psychiatry* 27: 425–443.

Fischer, J.M., and M. Ravizza. 1998. *Responsibility and control. A theory of moral responsibility*. Cambridge: Cambridge University Press.

Haji, I. 2010. Psychopathy, ethical perception, and moral culpability. *Neuroethics* 3: 135–150.

Juth, N., and F. Lorentzon. 2010. The concept of free will and forensic psychiatry. *International Journal of Law and Psychiatry* 33: 1–6.

Kane, R. 2002. *The Oxford handbook of free will*. Oxford: Oxford University Press.

Kinscherff, R. 2010. Proposition: A personality disorder may nullify responsibility for a criminal act. *The Journal of Law, Medicine & Ethics* 38: 745–759.

Lang, A. 1991. Patient perception of tics and other movement disorders. *Neurology* 41: 223–228.

Litton, P. 2010. Psychopathy and responsibility theory. *Philosophy Compass* 5: 676–688.

Maibom, H.L. 2008. The mad, the bad, and the psychopath. *Neuroethics* 1: 167–184.

Malatesti, L., and J. McMillan (eds.). 2010. *Responsibility and psychopathy. Interfacing law, psychiatry, and philosophy*. Oxford: Oxford University Press.

Matthews, S. 2004. Failed agency and the insanity defence. *International Journal of Law and Psychiatry* 27: 413–424.

McKenna, M. 2009. *Compatibilism*. Retrieved March 23, 2011, from http://plato.stanford.edu/entries/compatibilism.

Mele, A. 1995. *Autonomous agents. From self-control to autonomy*. New York: Oxford University Press.

Meynen, G. 2010. Free will and mental disorder: Exploring the relationship. *Theoretical Medicine and Bioethics* 31: 429–443.

Morse, S.J. 2007. The non-problem of free will in forensic psychiatry and psychology. *Behavioral Sciences & the Law* 25: 203–220.

Muller, S., and H. Walter. 2010. Reviewing autonomy: Implications of the neurosciences and the free will debate for the principle of respect for the patient's autonomy. *Cambridge Quarterly of Healthcare Ethics* 19: 205–217.

O'Connor, T. 2010. *Free will*. Retrieved February 2, 2011, from http://plato.stanford.edu/entries/freewill/.

Olatunji, B.O., J.M. Cisler, and B.J. Deacon. 2010. Efficacy of cognitive behavioral therapy for anxiety disorders: A review of meta-analytic findings. *The Psychiatric Clinics of North America* 33: 557–577.

Perring, C. 2004. Conceptual issues in assessing responsibility for actions symptomatic of mental illness. *International Journal of Law and Psychiatry* 27: 489–503.

Perring, C. 2010. *Mental illness*. Retrieved September 21, 2011, from http://plato.stanford.edu/entries/mental-illness/.

Pouncey, C.L., and J.M. Lukens. 2010. Madness versus badness: The ethical tension between the recovery movement and forensic psychiatry. *Theoretical Medicine and Bioethics* 31: 93–105.

Reich, W. 2005. Psychiatric diagnosis as an ethical problem. In *Psychiatric ethics*, 3rd ed, ed. S. Bloch, P. Chodoff, and S. Green, 193–224. Oxford: Oxford University Press.

Sadock, B.J., and V.A. Sadock (eds.). 2005. *Kaplan & Sadock's comprehensive textbook of psychiatry*, 8th ed. Philadelphia: Lippincott Williams & Wilkins.

Stone, A.A. 2008. The ethical boundaries of forensic psychiatry: A view from the ivory tower. *The Journal of the American Academy of Psychiatry and the Law* 36: 167–174.

Strawson, P.F. 2003. Freedom and resentment. In *Free will*, ed. G. Watson, 72–93. Oxford: Oxford University Press.

Verdellen, C.W., C.A. Hoogduin, B.S. Kato, G.P. Keijsers, D.C. Cath, and H.B. Hoijtink. 2008. Habituation of premonitory sensations during exposure and response prevention treatment in Tourette's syndrome. *Behavior Modification* 32: 215–227.

Walter, H. 2001. *Neurophilosophy of free will. From libertarian illusions to a concept of natural autonomy*. Cambridge, MA: MIT Press.

Widerker, D., and M. McKenna (eds.). 2003. *Moral responsibility and alternative possibilities: Essays on the importance of alternative possibilities*. Aldershot: Ashgate.

Chapter 14
Natural Morality, Moral Natures and Human Flourishing

Darcia Narvaez

14.1 Introduction

Morality and moral self-knowledge have never been more important as humanity faces unprecedented challenges like ecosystem degradation and climate instability (Millennium Ecosystem Assessment 2005; Intergovernmental Panel on Climate Change 2007). Yet human moral capacities (in the USA) seem to be diminishing. Among college students, narcissism is on the rise (Twenge and Campbell 2009) and empathy has been dropping over decades, significantly so in the last decade (Konrath et al. 2011). College student moral judgment scores are falling from an emphasis on post conventional reasoning not to conventional but down to a preference for personal interests (Thoma and Bebeau 2008). Adult misbehaviour is more widespread than in the past and has become normative (Callahan 2004). Even young children are at a greater likelihood of being expelled from preschool for aggression and are increasingly being prescribed psychotropic drugs for misbehaviour and dysphoria (Gilliam 2005; Powell et al. 2003). These are just a few of the signals that moral functioning may be in trouble.

What is the source of these declines? Is it a lack of moral reasoning development? Although moral reasoning offers some guidance for categorizing and constructing argumentation and forming and formulating social agreements (Rest et al. 1999, 2000), reasoning is insufficient for instigating much moral behaviour. Although individuals develop in the sophistication of their reasoning based on age, amount of education and expertise built from experience (Narvaez and Gleason 2007) and those with post conventional reasoning capacities show better clinical performance

D. Narvaez (✉)
Department of Psychology, University of Notre Dame, South Bend, IN, USA
e-mail: dnarvaez@nd.edu

B. Musschenga and A. van Harskamp (eds.), *What Makes Us Moral? On the capacities and conditions for being moral*, Library of Ethics and Applied Philosophy 31,
DOI 10.1007/978-94-007-6343-2_14, © Springer Science+Business Media Dordrecht 2013

beyond intelligence in multiple professions (Rest and Narvaez 1994), the predictive power of moral judgment for moral action is weak. Some people with sophisticated reasoning capacities take moral action but many do not.

Is it a lack of punishment? – only if you believe that humans are naturally evil and must be punished to learn be good, a Puritanical view running deeply in America's social practice, including in schools where 'zero-tolerance' policies are normative (Fowler 2010). Psychological, social and anthropological sciences would certainly foment substantial global evidence against this view.

Is it a lack of moral formation? Yes, goes the argument here. Neurobiological and psychological sciences show that the roots of individual and cultural moral expression lie in the body and in the emotions, aspects of the self that are shaped in early life (Narvaez 2008, 2010). As embodied creatures, human morality is linked to neurohormonal, emotional and other biological functioning. When these are deficient, virtue is harder to come by. Moreover, for any dynamic system, initial conditions are critical. Earlier development is built upon by later development. The source of moral difficulties may emerge from the deficient developmental environments of children whose psychobiological decrements sometimes only become apparent in adolescence or adulthood. Because humans are born extremely helpless and needy (with only 25 % of the brain developed at full-term birth), their brains and bodies evolved to expect *intensive* responsive parenting (Trevathan 2011), the kind of care that seems to be diminishing in advanced nations, particularly in the USA.

14.2 What Is Evolved Responsive Care?

Most adults would agree that babies have needs that must be met by responsive parenting. In fact, responsiveness to the emotional needs of the child is one of the best predictors of positive child outcomes (e.g. Kochanska 2002). However, responsivity alone may not be sufficient for optimal development; additional caregiving practices may be required (Narvaez and Gleason 2013).

Humans have evolved to be born extremely early to accommodate a woman's small pelvis and the child's large head (18 months early compared to other primates; Trevathan 2011). The result is that the child requires post-natal care that resembles an external womb (Montagu 1978). The ancestral human mammalian milieu (AHMM; Narvaez and Gleason 2013) represents human practices documented from extant nomadic gatherer-hunter societies (SBGH), whose lifestyle represents over 99 % of human genus history (Fry 2006). Thus far scientists have focused primarily on summarizing the characteristics of early life for infants and young children among SBGH (Hewlett and Lamb 2005). See Table 14.1. In this chapter I also include some of the social environmental characteristics noted by anthropologists (Ingold 1999; see Narvaez 2013, in press for more). Apart from the human adaptation of multiple caregivers and variable breastfeeding length, AHMM parenting practices are over 30 million years old, having emerged with the catarrhine mammals (Konner 2010).

Table 14.1 Ancestral human mammalian milieu for babies and young children

Natural Childbirth (no interference with timing, no drugs, no separation of mom and baby)

Breastfeeding: Nursed frequently (2–3 times/h initially) for 2–5 years (average weaning at 4 years)

Touch: Held or kept near others constantly

Response: Prompt responses to fusses and cries

Extensive Maternal Support and Alloparents: Shared care by adults other than mothers; extensive social embedment of mother–child dyad

Play: Enjoy free play in natural world with multiage playmates

Positive Social Climate and Positive Emotions: Daily life filled with laughter, singing, dancing

We can see that many of these practices hardly exist today among the majority of Western parents, especially in the USA. Does it matter? What are the consequences for abandoning evolutionarily-normative childbirth and child rearing practices? The contributors to the volume, *Evolution, early experience and human development: From research to practice and policy* (Narvaez et al. 2013b), suggest that when these principles are violated there are lifetime consequences on health and well-being (see also Karr-Morse and Wiley 2012).

Early experience sets the baseline for multiple functions, psychologically as well as physiologically. One of the constructs most studied in developmental psychology is attachment and caregiving (Ainsworth and Bowlby 1991; Bowlby 1969/1982). The *attachment behavioural system* is a species-universal programme that bonds child to mother. It collaborates with the adult's *caregiving behavioural system*, which under normal conditions is a species-wide programme that guides caregiving. Both programmes are shaped during sensitive periods. Early life experience with caregivers sets up the attachment behavioural system in infants and young children and the caregiving behavioural system in the parent is influenced by the quality of attachment the caregiver developed during early life and other sensitive times (George and Solomon 2008). For the mother, the caregiving behavioural system is also influenced by what happens during pregnancy and birth (Trevathan 2011). Each of these species-wide programmes guides the activation and selection of behaviours to reach particular goals in social relationships.

Secure attachment. When caregiving is warm and responsive, comforting the child's easily-distressed immature reflexive systems, these systems learn to maintain calm. The child learns the communicative value of interpersonal signals, both cognitive and affective, through caregiver emotional signalling – verbal communication and behaviour that are synchronized, supportive and reliable. The child's needs are satisfied through the attachment figure. As a result, a *secure* attachment develops, accompanied by a repertoire of social communication and social behaviours appropriate for the cultural context – healthy baselines for the life to come – including the ability to self-soothe with mental representations of the caregiver.

Insecure attachment. Crittenden (1998) and Fosha (2003) describe three *psychological* ways in which infants adapt over time to habitual non-responsive caregivers developing forms of insecure attachment: ‘feeling but not dealing’, ‘dealing but not feeling’, and ‘neither feeling nor dealing’. When the caregiver is rejecting or inconsistent, an insecure attachment develops with that caregiver. When caregivers are *rejecting* of a young child’s overtures for affection or need satisfaction, the child learns to suppress emotion and develops an *avoidant* attachment (dealing but not feeling). When caregivers are *inconsistent* in response, the child learns to use affect as a way to get attention and needs met, shutting down cognition (since it is unreliable predictor of caregiver behaviour), developing an *ambivalent* or *anxious* attachment (feeling but not dealing). When the child is abused, the child does not develop in feeling or thinking and ends up with a *disorganized* attachment (neither feeling nor dealing). Those with insecure attachment become inflexible and self-centred in social situations – they emotionally withdraw, attack or manipulate.

These attachment styles indicate not only the quality of the caregiver–child relationship but also how well the brain works. Across animal and human studies, it is apparent that much damage to basic systems can be done in early life from perinatal trauma or poor early care (Lanius et al. 2010). The damage undermines the functioning of cell repair and development, neuronal communication and stability levels of neurotransmitters, receptors and thresholds for molecular action (Grosjean and Tsai 2007). When functioning is compromised at this level, it can be impossible to repair and can lead to chronic disease (Barker 1998). As a result of poor care, the brain and body are less well functioning, less physiologically integrated, and rely on systems that are minimally competent. If the body and brain can barely maintain homeostasis in the face of stress, virtue will be harder to come by. For example, self-regulation becomes a challenge when neurotransmitters run out of glucose rapidly (Galliot 2008). In fact, those with low glucose are functionally less prosocial (De Wall et al. 2010). Consequently, one may never have the sharp moral intelligence or prosocial orientation that characterizes those with good care during sensitive periods (Kochanska 2002).

Insecure attachment means that social development has gone awry for that individual (although it may be locally adaptive to avoid a rejecting parent), affecting social life for the long term, if intervention does not ensue. Only with a brain that displays secure attachment can we be sure that moral functioning is developing in a promising direction.

Children with secure attachment live in a different moral universe from those with insecure attachment. Their brains and bodies work better from the good care received (Narvaez et al. 2013b). They are more socially competent and confident, friendly and trusting of others (Hazan and Shaver 1987). They are more compliant with adult wishes and are more likely to exhibit empathy, conscience, and prosocial behaviour (Kochanska 2002). These are characteristics of adult moral exemplars (Walker and Frimer 2009). In contrast, children with insecure attachment are less empathic, more aggressive, and have difficulty getting along with peers (Sroufe et al. 2008). Those with avoidant attachment can be emotionally distant and exhibit compulsive self-reliance (Bartholomew and Horowitz 1991; Hazan and Shaver 1987). In short, personalities that emerge from insecure attachment make cooperation with others difficult.

Attachment is a measurable indicator of how well early life went. However, it is a blunt indicator that does not measure the underlying physiological scars of intermittent poor care. Remember that humans are born (at full term – 40–42 weeks gestation) with only 25 % of the brain developed, with next to no immune system, and with all sorts of other brain and body systems to be established and interconnected. Most of this growth occurs rapidly in the first 6 years of life, co-constructed by caregivers and a built-in biological timetable that co-evolved with particular parenting practices (Trevathan 2011).

Early life experience shapes the dynamic psychosoma system that is a child. For any dynamic system, initial conditions are critical for development and later parameters of possibility. With the epigenetic revolution, we are finding out that most of who we become is formed by our experiences after conception (although the experiences of our parents and grandparents also affect our health; see Gluckman and Hanson 2005, for a review).

Here is one example of how one physiological system is shaped by early caregiving that has long-term consequences on health and well-being (and there are many systems that could be described). The vagus nerve is the primary nerve of the parasympathetic nervous system, the majority of whose fibres (80 %) are afferent, carrying information from the viscera to the brainstem; the other 20 % of fibres are efferent, facilitating the modulation of arousal and body state, when properly developed (Porges and Carter 2010). The modulation capability is vital for social functioning, inhibiting flight or fight responses. How well the vagus functions (vagal tone) is calculated from measuring heart rate variability, a rough measure of how well the vagus nerve controls resting heart rate. Vagal tone function is highly affected by early-life caregiving, particularly touch (Porges 2011). The degree of variability in heart rate between breath inspiration and expiration is higher among those with good vagal tone, and is more likely to be found among those with secure attachment. Those with little variability have more difficulty handling negative emotions and socializing, as exhibited through behavioural inhibition to novel situations. Babies left to cry develop poor vagal tone and learn to employ self-protective responses in social situations. Moreover, vagal function is related to most systems in the body and when poorly functioning can lead to numerous detrimental health outcomes (digestive – e.g. irritable bowel; neuronal communication – e.g., seizures; mental health-depression; see Ghanem and Early 2006; Groves and Brown 2005). The vagus nerve can even prevent inflammation, an instigator of many diseases.

Vagal tone is one of many mechanisms related to homeostatic self-control. At a very fundamental level, self-control involves mechanisms that manage stress reactivity and negative emotions. When a person's basic biology has been neglected in terms of evolved expected care in early life, he or she is set up to be stress reactive, including epigenetic effects such as not turning on genes that regulate anxiety (e.g. Meaney 2001; Weaver et al. 2002). Stress reactivity is hard to remedy later.

Physiological effects, such as stress reactivity, have moral implications. If one does not feel at ease, how can one face new people, think outside the box and socially succeed? Instead, new things (e.g. people who look or act differently) are threatening and lead to immobilized thinking, feeling, relating, or worse, reactive aggression. The contention here is that current cultural practices in the treatment of children, beginning prior to birth and throughout development, are having an influence on the health and well-being of US citizens (and similar effects may be obtained in countries that have adopted US childbirthing and childrearing practices). These practices may be undermining humanity's evolved principles for human growth and development, affecting individuals, families and societies. I believe that early life caregiving also affects a person's *moral* nature (Narvaez 2008; Narvaez and Gleason 2007). Multi-ethics theory and research addresses these issues.

14.3 Moral Natures and Moral Development

Multi-ethics theory – formerly called 'triune ethics theory' (Narvaez 2008) – is a descriptive psychobiological theory of moral development and moral functioning. It offers a neurobiologically based explanation for different 'moral mind-sets' or moral natures. An individual may learn to habitually use one mind-set or another based on experience during sensitive periods. Certain situations also may evoke a particular mind-set.

Three distinctive moral systems, rooted in basic brain strata that create global brain states (MacLean 1990), propel human moral action on an individual and group level: safety, engagement, and imagination. Briefly, the safety ethic is oriented to feeling safe through following a set of conditioned prescriptions, including maintaining loyalty and purity. The engagement ethic is oriented to relational attunement with others, facilitating flexible empathic response. The imagination ethic attends to transcendent ideals and life narratives, often fuelled by one of the other ethics. See Table 14.2 for a summary.

The safety ethic is rooted in physiological systems for *self-protection* which are available at birth and shared with all animals. When an instinctive survival mind-set is used in making decisions and taking action, it becomes a safety ethic. The safety ethic gets triggered by threat, physical or psychological. We all have this ethic within us – when we are motivated to withdraw, emotionally or physically, from a relationship or lash out in self-defence. This sense of on-going immediate threat can become a dispositional social orientation if trauma, abuse or neglect was experienced during a sensitive period in life (sensitive periods include the first 5 years, early adolescence, early adulthood, therapy).

Table 14.2 Types of moral mind-sets

Type	Bunker safety	Wallflower safety	Engagement	Communal imagination	Detached imagination	Vicious imagination
Focus	Self-protection		Harmony or relational attunement	Egalitarian social problem solving	Manipulation and control	Dominance or revenge
Response to stress	Fight, flight	Freeze	Tend and befriend	Higher consciousness	Disassociation	Hostility, control

When early childhood is distressful, from poor caregiving or trauma, the safety ethic can become a favoured disposition when face-to-face with others. When caregiving has not provided expected companionship care (Trevarthen 2005) and a mutually responsive orientation (Kochanska 2002), individuals become distrustful and learn to structure relationships as a shift between dominance, submission and withdrawal.

There are two basic forms of safety ethic, one is proactively aggressive (*bunker* safety) where the individual feels enough strength and power to take action against the threat and maintain dominance. In the other mind-set the individual feels too weak or paralysed to take action and so withdraws physically and/or emotionally (*wallflower* safety). Dispositionally, individuals can favour one or the other or flip between them depending on the situation (a bully in one situation but a doormat in another). Because self-protective mechanisms can take over the mind/brain, poor early experience that leads to stress reactivity limits genuine autonomy in the present moment. Instead, the conditioned past determines which emotions and perceptual filters are triggered and influence perceptions and action possibilities. Neuroses, as explored by Freud, demonstrate that traumatic preverbal sensory experiences can be evoked by current circumstances, leading to subjectively rational but objectively 'irrational' behaviour for the situation. The moral natures represented by the safety ethic are harmful to self and others in the long term.

The second moral mind-set is engagement; it is about *relationship*, specifically, relational attunement. It draws on right brain experiencing, the awareness of energy and connection to all life that neuroscientist Jill Bolke Taylor (2008) described as she experienced a stroke on the left side of her brain. The engagement ethic focuses on connecting and bonding in the moment, on an equal basis, person to person, life form to life form. How much an individual is able to do this is based in brain systems and intuitions formed before speech, in the earliest years, and during other sensitive periods. What is occurring during relational attunement is limbic resonance (Lewis et al. 2000), or intersubjectivity with others (Trevarthen 2005). Deep engagement is required for full moral capabilities. The moral nature represented by the engagement ethic appears to match up well with the human nature exhibited by SBGH, where children are well cared for, as reported by anthropologists (Ingold 1999). In these societies, for example, individuals are cooperative and caring, value generosity and don't put up with cheating, aggression or deception. Humans share the engagement orientation with other primates but have a further moral capacity that is largely ours alone, the imagination ethic.

The third moral mind-set, the imagination ethic, is rooted in our capabilities for *abstraction*, or pulling away from the present moment – the calculating intelligence identified by Taylor (2008) or what is called intellect (MacMurray 1992). Intellect, rooted in the neocortex and prefrontal cortex systems, plays a central role in imagining the future, forming and executing goals, and inhibiting and guiding action. These areas are also dependent on nurturing and can be damaged throughout life. The imagination ethic can take different forms. Ideally, it links with the prosocial orientation of the engagement ethic, becoming a *communal* imagination, allowing imagination to be applied to the common good. However, if it is tied up with the safety ethic, with self-protective concerns, it becomes a *vicious* imagination, focused on plans for self-aggrandizement. Alternatively if one dissociates from emotions entirely, it can be a *detached* imagination, allowing one to create and innovate without attending to consequences. These represent different moral natures with different benefits to self and others. The SBGH appear to have a broad communal imagination, with a sense of relational responsibility to all other entities (e.g. trees, mountains, animals). They lived sustainably, for example, moving from a gathering place before they depleted the area or changing a harvested resource if it appeared distressed (Gowdy 1999).

Multi-ethics theory suggests that early experience shapes moral functioning, fostering different moral natures. That is, one's physiology and psyche built by early experience underlie one's capacities as an adult (although one is not necessarily trapped forever in the level of functioning promoted by early experience – there are other sensitive periods and one can transform oneself somewhat with extensive effort using therapy, meditation and similar practices). As noted earlier, moral capacities in the USA appear to be shrinking, emphasizing stress-reactive *safety ethic*, as well as *vicious* and *detached* imaginations. Current social environments in the USA foster general distrust, self-protection, self-aggrandizement, detachment from soft emotions, ultra rationalism, inability to gather empirical evidence from personal experience, and coercive and dysphoric personalities.

My laboratory has begun to study these ethics. We find that secure attachment (formed by responsive parenting) is related to prosocial personality, both of which predict having engagement ethic as a moral goal (or identity) – all of which predict moral outcomes such as helping the less fortunate (Narvaez et al. 2013a). Safety as a moral goal or identity is linked to insecure attachment, distrust and dishonesty. Of course it is obvious that a society will not flourish if its citizens habitually adopt safety ethics or are dispositionally detached or vicious. Such moral natures do not lead to flourishing.

14.4 Human Nature and the Path to Moral Wisdom

Pinker (2011) argues that humanity has become much more peaceful, especially in the twentieth century, due to progress in societal and institutional structures. He reaches this mistaken conclusion by ignoring prehistory and collapsing types of hunter-gatherer societies. The SBGH, representing 99 % of human genus history, were largely peaceful and did not engage in war (Fry 2006). Further, Pinker narrows his definition of violence to physical aggression. The world has far more emotional and institutional violence today, perhaps more than ever if one counts the number of countries that are not meeting the needs of their people and the increasing undercare of children which represents forms of relational violence, from trauma at birth to thwarting of needs for breast milk and physical closeness.

At the same time that Pinker touts human progress, he, like social moral psychology and Hobbesian evolutionary psychology today, emphasizes the selfcentredness of humanity. Experiments are conducted to find such outcomes but also based on samples from populations raised in the aforementioned 'undercaring' environments that violate evolved needs. For example, moral foundations theory (Haidt and Joseph 2007) emphasizes values that relate to safety ethic concerns (hierarchy, purity, ingroup, tit-for-tat fairness). If these are central concerns for one's morality, it reflects a deficient human nature – one that has been socialized away from the pinnacle of humanity's moral sense, namely communal imagination. These are not the concerns of SBGH, who were fiercely egalitarian, had porous boundaries and expected universal generosity (Ingold 1999). Instead, such self-protective concerns emerge from the degradation of social environments, especially in early life, and the downward spiral of believing in human evil, punishing children for their mammalian needs, and considering troubled outcomes to be normal rather than formed by poor environments.

Environments guide intuition development. Most of what we know and who we are resides in implicit systems that learn automatically and guide our behaviour and choices most of the time (Kahneman 2011). These systems learn from whole-body experience, especially in the early years when right brain emotion networks are established from responsive care. When children are put in front of books and screens, their right brains are minimally developed and they end up with a deficient storehouse of knowledge from which to draw intuitions the rest of their lives (Schore 2003a, b). This has moral implications. In the left brain conscious knowledge becomes the source of the moral enterprise, leading to an emphasis on detached imagination (left-brain objectivity), and then with poor early care, the use of the safety ethic in face-to-face situations.

Moral wisdom comes from understanding the nature of the world, requiring well-educated whole-brain intuitions and reliable 'personal' knowledge, both of which are primarily tacit, saturated with intelligence, experience and perspective (Polanyi 1958) – the opposite of left-brain emotional detachment. Whole-brain 'reasoning is primarily an affair of emotion' and 'none of our activities, not even the activities of thinking, can express our reason unless the emotions that produce and sustain them are rational emotions' (MacMurray 1962/1999, pp. 10–11). When we suppress and ignore emotion and emotional development, we necessarily stay in egocentric thought and cannot know the world as it is – we cannot know reality. Indeed, as mammals, emotions are central to intelligence and social functioning (Panksepp 1998). Humans with damage in emotional brain circuitry make poor decisions, including poor moral decisions (Damasio 1999).

The narrow, evaluative uneducated intuitions often studied in psychology experiments where novices are asked to evaluate the actions or character of someone doing something outrageous (e.g. Haidt 2001) represent the kind of evaluative activity that fits into the deficient skill level of persons who have been raised in isolation watching others act on screens – it does not comprise moral behaviour, which of course is much more complicated. This type of veridical decision making is relatively easy and can use the primitive systems of decision making that underlie the safety ethic. The real work of decision making, including moral decision making, is adaptive decision making – what is required in the real world where nothing is pre-packaged, when one has to figure out how to sort out the stimuli and figure out what is going on, what might be done, what role one might take, and so on and so forth. This is the real work of moral functioning and it can be called moral imagination.

Moral imagination is a sophisticated type of deliberation. As Dewey says, it includes ‘the capacity to concretely perceive what is before us in light of what could be’ (Dewey in Fesmire 2003, p. 2). It works with thought experiments, through the dramatic rehearsal of alternative courses of action and possible outcomes. This form of internalized social action requires tacit knowledge that includes the wisdom of emotion and the understanding of the nature of the world. Moral imagination allows humans the ‘ability to choose ends and goals’, ends that, unlike for other animals, are not dictated by simple biology. Humans also have the ability to calculate the means to reach their goals (Neiman 2008, pp. 189, 191). Moral imagination involves a variety of higher-order thinking skills considered to be key factors in astute thinking (Perkins 1995), such as the ability to decentre away from one’s current view and to generate alternative scenarios and look for counterevidence to first reactions, impulses, and preconclusions.

Real-life social experience builds the storehouse of knowledge from which moral wisdom can emerge. The more practiced and refined one’s imagination is, the richer the bank of possibilities and the more reliable one’s evaluations are (Dewey 1922/2000). But what Dewey may not have understood is that early life experience is critical for the bank of knowledge from which one draws examples and intuitions about imagined possibilities. Fluency in generating alternative viewpoints, particularly the perspective of others, is a skill that develops not only from prefrontal cortex development, highly influenced in the first years of life and fully mature by the third decade of life (Goldberg 2002), but also through life experience generally and within particular domains (Feshbach 1989; Selman 2003).

Deliberation generally facilitates ‘self-authorship’ (Baxter Magolda 2001). When we feel a sense of injustice or upset, we can step back from the response and ask whether it is an appropriate response and whether it is a reliable guide for action (Sen 2009). School-based programmes in social and emotional learning are documented to help students stop the rapid emotional response and think more carefully about action (e.g. Elias et al. 2008) and increase cognitive competencies in decision making (see Catalano et al. 2008, for a review), allowing the individual to monitor intuitions. Reason allows us to select the environments that ‘tune up’ our intuitions (Hogarth 2001), a means to self-cultivate virtue. But babies and young children cannot select their environment, despite its critical importance for long-term outcomes. Adults must take on the responsibility for ensuring that babies and young children are provided their basic needs. Human capacities and human nature are both better represented by the moral personalities of small-band hunter-gatherer groups where children are well cared for and adults behave virtuously as a matter of course (Ingold 1999).

If we take human flourishing and virtue as aims of development, then the SBGH social context offers a framework for both the aims and the means. See

Table 14.3 Comparison of social life in small-band gatherer-hunter communities and twenty-first-century USA

Topic	SBGH social life	Twenty-first-century USA social life
Maternal support	Complete	Undermining of mothers
Babies and children	Needs fully met	Minimal support
Child behaviour	No coercion	Coercion and punishment
Autonomy	Individuals do what they want	Limited mostly to consumer choices
Community	High commitment and cooperation	Minimal connection to others
Source of pleasure	Social activities	Individual activities (television, internet)
Virtues	Communal generosity and sharing expected	Self-interest
Vices	Cheating, deception, aggression not tolerated	Cheating, deception, aggression expected
Relation to natural world	Partnership; treatment of other entities as subjects not objects; deep connection to Life	Nature objectified, manipulated; everything else assumed to serve humanity's aims; little awareness of Life

Table 14.3. The means include meeting the needs of babies and young children fully, when bodies, brains and personalities are being formed. Humans evolved for intense social community and become depressed and psychotic in isolation, much like Harlow's (Harlow et al. 1965) monkeys. But this social embedment does not mean coercive conformity. Individuals have autonomy from a young age and even adults could choose to go on hunts or gathering trips (Ingold 1999). Social activities such as singing, dancing, joking, make communal life very enjoyable and attractive. The SBGH have a good sense about human egos and what can cause problems, so they were 'fiercely egalitarian', discouraging any boasting or self-aggrandizement. Principles of generosity and sharing were implicit, along with expulsion for aggression, coercion or deception. The social lifestyle from birth and early childhood through adulthood represents a moral fabric or flow of experience that facilitates virtue, including ecological virtue.

Should we be worried about the naturalistic fallacy? Some might accuse me of mixing fact and value, making ethical prescriptions based on descriptions of natural facts. This is a misunderstanding. As several philosophers have pointed out (e.g. MacIntyre 1959), the naturalistic fallacy is a false problem. As Owen Flanagan (2009) notes, Hume (1739) was warning readers to beware of *clerics* who made is-to-ought prescriptions, exhibiting deductive logical errors. Flanagan (2012) emphasizes that ethics represents a 'biopsychosocial technology involving sensible determination of ends and means.' I adopt Flanagan's (2012) view that 'eudemonics is a descriptive-normative theory of the causes and constituents of human flourishing, including ethics.' As Murdoch (1989) points out, philosophy should be addressing eudemonic questions, such as 'how do we make ourselves morally better?' That is the aim here.

14.5 The Future Is the Past

One might argue that the USA has learned to value and nurture the wrong things. With its emphasis on reason and scientific detachment instead of appreciative presence common among SBGH (Narvaez 2012), the USA has larded up self-protective and detached thinking, starving deep relational attunement and engagement. USA personalities generally are inadequate for facing humanity's moral problems today or for bringing about well-being across life forms. Somehow we have forgotten that the good life requires well-cultivated emotion. Humans apprehend life with emotional-cognitive capacities. How well these are developed by early experience influences subsequent socialization into a culture and by chosen experiences throughout life.

Through childrearing practices that violate the evolved practices described earlier and through other practices like sensory-deadening schooling, the USA may be fostering a misshapen morality, one that doesn't live up to human potential. By misguiding emotional development in early life (through the use of punishment, neglect of children's needs), it seems to have emphasized a narrow intelligence, a left-brain dominant version of functioning (McGilchrist 2009). This seems to be exemplified in the USA in that EQ (emotional intelligence) has been decreasing while 'IQ' scores have been rising throughout the twentieth century. But IQ scores have risen only because of increased scores on subtests representing abstract thinking, the ability to emotionally detach and answer hypothetical questions (i.e. scientific thinking; Flynn 2007). Thus IQ represents mere 'intellect', a narrow type of knowledge based in conscious, linear thinking, that current childrearing and schooling practices seem to emphasize. Schooling smothers emotion development, emphasizing left-brain detached thinking, not wisdom of any kind.

14.6 Conclusion

Current moral psychology findings are misleading. Human nature is not so selfish, except that we have adopted a culture that shapes us this way (speaking of the USA). The range of psychopathology has been made enormously wide due to miscultivation of children. As Turnbull (1983) noted, the evisceration of human emotion development leaves children empty by adolescence. It leaves them vulnerable to the ideologies of their surround, whether the Protestant work ethic, religious fundamentalism, or the hot ideology today, egoism. In contrast, wholistic knowledge encompasses well-trained emotions and intuitions built from coached experience, starting from responsive caregiving at birth. Infants expect companionship care and are ready for mutual responsive communication (Trevarthen 2005) – which builds a fully social and intelligent brain. In the societies of our 99 %, morality was integral to survival and was deep in personality and a relational psychology (see Narvaez 2013, in press for more). I agree with MacMurray that the key element of morality is the capacity for empathic, joyful coordination of a web of personal relationships, a personal emotional wisdom that one can apply to more complex problems. It involves small egos, the ability to give and take in playful interaction, a type of ego-detachment that allows one to truly see reality without purpose except to give full attention (Murdoch 1989). Moral wisdom emerges from this immersion in deep, extensive engagement with others.

References

Ainsworth, M.D.S., and J. Bowlby. 1991. An ethological approach to personality development. *American Psychologist* 46: 333–341.

Barker, D.J. 1998. In utero programming of chronic disease. *Clinical Science* 95: 115–128.

Bartholomew, K., and L.M. Horowitz. 1991. Attachment styles among young adults: A test of a four-category model. *Journal of Personality and Social Psychology* 61: 226–244.

Baxter Magolda, M.B. 2001. *Making their own way: Narratives for transforming higher education to promote self-development*. Sterling: Stylus.

Bowlby, J. (1982) [1969]. *Attachment and loss: Vol. 1 Attachment*, 2nd ed. New York: Basic Books.

Callahan, D. 2004. *The cheating culture: Why more Americans are doing wrong to get ahead*. New York: Harcourt Harvest.

Catalano, R.F., J.D. Hawkins, and J.W. Toumbourou. 2008. Positive youth development in the United States: History efficacy and links to moral and character education. In *Handbook of moral and character education*, ed. L.P. Nucci and D. Narvaez, 459–483. New York: Routledge.

Crittenden, P.M. 1998. The developmental consequences of childhood sexual abuse. In *Violence against children in the family and the community*, ed. P. Trickett and D. Schellenback, 11–38. Washington, DC: American Psychological Association.

Damasio, A. 1999. *The feeling of what happens*. London: Heineman.

DeWall, C.N., R.S. Pond, and B.J. Bushman. 2010. Sweet revenge: Diabetic status as a predictor of interpersonal forgiveness. *Personality and Individual Differences* 49: 823–826.

Dewey, J. (2000) [1922]. *Human nature and conduct: An introduction to social psychology*. New York: Prometheus.

Elias, M.J., S.J. Parker, V.M. Kash, R.P. Weissberg, and M.U. O'Brien. 2008. Social and emotional learning moral education and character education: A comparative analysis and a view toward convergence. In *Handbook of moral and character education*, ed. L.P. Nucci and D. Narvaez, 248–266. New York: Routledge.

Feshbach, N.D. 1989. Empathy training and prosocial behaviour. In *Aggression and war: Their biological and social bases*, ed. J. Grobel and R.A. Hinde, 101–111. Cambridge: Cambridge University Press.

Fesmire, S. 2003. *John Dewey and the moral imagination: Pragmatism in ethics*. Bloomington: Indiana University Press.

Flanagan, O. 2009. *The really hard problem: Meaning in a material world*. Cambridge: MIT Press.

Flanagan, O. 2012. *What do the psychology and biology of morality have to do with ethics? Ethics as human ecology*. Presentation at The Evolution of Morality: The Biology and Philosophy of Human Conscience. Erice, Sicily, 17–22 June.

Flynn, J.R. 2007. *What is intelligence?* New York: Cambridge University Press.

Fosha, D. 2003. Dyadic regulation and experiential work with emotion and relatedness in trauma and disorganized attachment. In *Healing trauma: Attachment mind body and brain*, ed. M. Solomon and D. Siegel, 221–281. New York: Norton.

Fowler, D. 2010. *Texas' school-to-prison pipeline: Ticketing arrest & use of force in schools; How the myth of the 'blackboard jungle' reshaped school disciplinary policy*. Austin: Texas Appleseed.

Fry, D.P. 2006. *The human potential for peace: An anthropological challenge to assumptions about war and violence*. New York: Oxford University Press.

Galliot, M.T. 2008. Unlocking the energy dynamics of executive functioning: Linking executive functioning to brain glycogen. *Perspectives on Psychological Science* 3(4): 245–263.

George, C., and J. Solomon. 2008. The caregiving system: A behavioral system's approach to parenting. In *Handbook of attachment: Theory research and clinical applications*, ed. J. Cassidy and P. Shaver, 833–856. New York: Guilford.

Ghanem, T., and S. Early. 2006. Vagal nerve stimulator implantation: An otolaryngologist's perspective. *Otolaryngology – Head and Neck Surgery* 135(1): 46–51. doi:101016/jotohns200602037. PMID 16815181.

Gilliam, W.S. 2005. *Prekindergarteners left behind: Expulsion rates in state prekindergarten systems*. New Haven: Yale University Child Study Center.

Gluckman, P., and M. Hanson. 2005. *Fetal matrix: Evolution development and disease*. New York: Cambridge University Press.

Goldberg, E. 2002. *The executive brain: Frontal lobes and the civilized brain*. New York: Oxford University Press.

Gowdy, J. 1999. Gatherer-hunters and the mythology of the market. In *The Cambridge encyclopedia of hunters and gatherers*, ed. R.B. Lee and R. Daly, 391–398. New York: Cambridge University Press.

Grosjean, B., and G.E. Tsai. 2007. NMDA neurotransmission as a critical mediator of borderline personality disorder. *Journal of Psychiatry and Neuroscience* 32(2): 103–115.

Groves, D.A., and V.J. Brown. 2005. Vagal nerve stimulation: A review of its applications and potential mechanisms that mediate its clinical effects. *Neuroscience and Biobehavioral Reviews* 29(3): 493. doi:101016/jneubiorev200501004.

Haidt, J. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review* 8: 814–834.

Haidt, J., and C. Joseph. 2007. The moral mind: How 5 sets of innate intuitions guide the development of many culture-specific virtues and perhaps even modules. In *The innate mind*, vol. 3, ed. P. Carruthers, S. Laurence, and S. Stich, 367–391. New York: Oxford University Press.

Harlow, H.F., R.O. Dodsworth, and M.K. Harlow. 1965. Total social isolation in monkeys. *Proceedings of the National Academy of Sciences of the United States of America* 54(1): 90–97.

Hazan, C., and P. Shaver. 1987. Romantic love conceptualized as an attachment process. *Journal of Personality and Social Psychology* 52(3): 511–524.

Hewlett, B.S., and M.E. Lamb. 2005. *Hunter-gatherer childhoods: Evolutionary developmental and cultural perspectives*. New Brunswick: Aldine.

Hogarth, R.M. 2001. *Educating intuition*. Chicago: University of Chicago Press.

Hume, D. 1739. *A treatise of human nature*. London: John Noon.

Ingold, T. 1999. On the social relations of the hunter-gatherer band. In *The Cambridge encyclopedia of hunters and gatherers*, ed. R.B. Lee and R. Daly, 399–410. New York: Cambridge University Press.

Intergovernmental Panel on Climate Change. 2007. *Climate change 2007: A synthesis report*. Geneva: World Meteorological Organization (WMO) and United Nations Environment Programme (UNEP).

Kahneman, D. 2011. *Thinking fast and slow*. New York: Farrar Strauss & Giroux.

Karr-Morse, R., and M.S. Wiley. 2012. *Scared sick: The role of childhood trauma in adult disease*. New York: Basic Books.

Kochanska, G. 2002. Mutually responsive orientation between mothers and their young children: A context for the early development of conscience. *Current Directions in Psychological Science* 11(6): 191–195. doi:101111/1467-872100198.

Konner, M. 2010. *The evolution of childhood*. Cambridge, MA: Belknap.

Konrath, S., E.H. O'Brien, and C. Hsing. 2011. Changes in dispositional empathy over time in college students: A meta-analysis. *Personality and Social Psychology Review* 15: 180–198.

Lanius, R.A., E. Vermetten, and C. Pain (eds.). 2010. *The impact of early life trauma on health and disease: The hidden epidemic*. New York: Cambridge University Press.

Lewis, T., F. Amini, and R. Lannon. 2000. *A general theory of love*. New York: Vintage.

MacIntyre, A.C. 1959. Hume on 'Is' and 'Ought'. *Philosophical Review* 68(4): 451–468.

MacLean, P.D. 1990. *The triune brain in evolution: Role in paleocerebral functions*. New York: Plenum.

MacMurray, J. 1992. *Reason and emotion*. Amherst: Humanity Books.

MacMurray, J. 1962/1999. *Reason and emotion*. New York: Humanity Books.

McGilchrist, I. 2009. *The master and his emissary: The divided brain and the making of the western world*. New Haven: Yale University Press.

Meaney, M.J. 2001. Maternal care gene expression and the transmission of individual differences in stress reactivity across generations. *Annual Review of Neuroscience* 24: 1161–1192.

Millennium Ecosystem Assessment. 2005. *Ecosystems and human well-being*. Washington, DC: Synthesis Island Press.

Montagu, A. 1978. *Learning nonaggression: The experience of non-literate societies*. New York: Oxford University Press.

Murdoch, I. (1989) [1970]. *The sovereignty of good*. New York: Routledge.

Narvaez, D. 2008. Triune ethics: The neurobiological roots of our multiple moralities. *New Ideas in Psychology* 26: 95–119.

Narvaez, D. 2010. Moral complexity: The fatal attraction of truthiness and the importance of mature moral functioning. *Perspectives on Psychological Science* 5(2): 163–181.

Narvaez, D. 2012. Moral neuroeducation from early life through the lifespan. *Neuroethics* 5(2): 145–157. doi:10.1007/s12152-011-9117-5.

Narvaez, D. 2013. Development and socialization within an evolutionary context: Growing up to become "A good and useful human being". In *War, peace and human nature: The convergence of evolutionary and cultural views*, ed. D. Fry, 643–672. New York: Oxford University Press.

Narvaez, D. 2013. Neurobiology and moral mindset. In *Handbook of moral motivation: theories, models, applications*, eds. K. Heinrichs, F. Ozer, and T. Lovat, 323–340. Rotterdam: Sense Publishers.

Narvaez, D., and T. Gleason. 2007. The influence of moral judgment development and moral experience on comprehension of moral narratives and expository texts. *Journal of Genetic Psychology* 168(3): 251–276.

Narvaez, D., and T. Gleason. 2013. Developmental optimization. In *Evolution, early experience and human development: From research to practice and policy*, ed. D. Narvaez, J. Panksepp, A. Schore, and T. Gleason, 307–325. New York: Oxford University Press.

Narvaez, D., J. Brooks, and S. Hardy. 2013a. A multidimensional approach to moral identity: Early life experience prosocial personality and moral outcomes [Manuscript submitted for publication].

Narvaez, D., J. Panksepp, A. Schore, and T. Gleason (eds.). 2013b. *Evolution, early experience and human development: From research to practice and policy*. New York: Oxford University Press.

Neiman, N. 2008. *Moral clarity: A guide for grown-up idealists*. New York: Harcourt.

Panksepp, J. 1998. *Affective neuroscience: The foundations of human and animal emotions*. New York: Oxford University Press.

Perkins, D. 1995. *Outsmarting IQ: The emerging science of learnable intelligence*. New York: Free Press.

Pinker, S. 2011. *The better angels of our nature*. New York: Viking.

Polanyi, M. 1958. *Personal knowledge: Towards a post-critical philosophy*. Chicago: University of Chicago Press.

Porges, S.W. 2011. *The polyvagal theory: Neurophysiological foundations of emotions attachment communication self-regulation*. New York: W.W. Norton.

Porges, S.W., and C. Carter. 2010. Neurobiological bases of social behavior across the life span. In *The handbook of life-span development*, Social and emotional development, vol. 2, ed. M.E. Lamb, A.M. Freund, and R.M. Lerner, 9–50. Hoboken: Wiley.

Powell, D., D. Fixen, and G. Dunlop. 2003. *Pathways to service utilization: A synthesis of evidence relevant to young children with challenging behavior*. Tampa: Center for Evidence-based Practice: Young Children with Challenging Behavior, University of South Florida.

Rest, J.R., and D. Narvaez (eds.). 1994. *Moral development in the professions: Psychology and applied ethics*. Hillsdale: Lawrence Erlbaum.

Rest, J.R., D. Narvaez, M.J. Bebeau, and S.J. Thoma. 1999. *Postconventional moral thinking: A neo-Kohlbergian approach*. Mahwah: Erlbaum.

Rest, J.R., D. Narvaez, M. Bebeau, and S. Thoma. 2000. A neo-Kohlbergian approach to morality research. *Journal of Moral Education* 29: 381–395.

Schore, A. 2003a. *Affect regulation and the repair of the self*. New York: Norton.

Schore, A. 2003b. *Affect dysregulation and disorders of the self*. New York: Norton.

Selman, R.L. 2003. *The promotion of social awareness: Powerful lessons from the partnership of developmental theory and classroom practice*. New York: Russell Sage.

Sen, A. 2009. *The idea of justice*. Cambridge, MA: Harvard University Press.

Sroufe, L.A., B. Egeland, E.A. Carlson, and W.A. Collins. 2008. *The development of the person: The Minnesota study of risk and adaptation from birth to adulthood*. New York: Guilford.

Taylor, J.B. 2008. *My stroke of insight*. New York: Viking.

Thoma, S.J., and M. Bebeau. 2008. *Moral judgment competency is declining over time: Evidence from 20 years of defining issues test data*. Paper presented to the American Educational Research Association, New York.

Trevarthen, C. 2005. Action and emotion in development of the human self its sociability and cultural intelligence: Why infants have feelings like ours. In *Emotional development*, ed. J. Nadel and D. Muir, 61–91. Oxford: Oxford University Press.

Trevathan, W.R. 2011. *Human birth: An evolutionary perspective*, 2nd ed. New York: Aldine de Gruyter.

Turnbull, C.M. 1983. *The human cycle*. New York: Simon and Schuster.

Twenge, J., and R. Campbell. 2009. *The narcissism epidemic: Living in the age of entitlement*. New York: Free Press.

Walker, L.J., and J.A. Frimer. 2009. Moral personality exemplified. In *Personality identity and character: Explorations in moral psychology*, ed. D. Narvaez and D.K. Lapsley, 232–255. New York: Cambridge University Press.

Weaver, I.C., M. Szyf, and M.J. Meaney. 2002. From maternal care to gene expression: DNA methylation and the maternal programming of stress responses. *Endocrine Research* 28: 699–700.

Part IV
Religion and (Im)Morality

Chapter 15
Atheism and the Basis of Morality

Stephen Maitzen

> 'I'm not "out" at my workplace,' said . . . Claire, a 27-year-old arts administrator who asked that her last name not be used. 'Because most people think atheists have no morals, I could damage the organization if I'm honest about where I stand on the issue.' [Oppenheimer 2010]

15.1 Introduction

Many people think that God, and only God, makes us moral. Academic researchers and journalists alike report that people in the Jewish-Christian-Islamic world typically link morality with belief in the God of traditional monotheism and see good ethical values as a crucial benefit of religion. According to survey results published in the *American Sociological Review*, for example, Americans distrust atheists more than any other group listed in the survey, and this distrust stems mainly from the notion that only believers in God can be counted on to respect morality (Edgell et al. 2006). As the American ethicist James Rachels puts it,

> it is not unusual for priests and ministers to be treated as moral experts.... When the prestigious National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research was organized in the mid-1970s, two seats on the commission were reserved for 'ethicists'; those seats went to a Jesuit priest and a professor at the Pacific School of Religion . . . . Why are clergymen regarded in this way? . . . In popular thinking, morality and religion are inseparable: People commonly believe that morality can be understood only in the context of [theism]. (Rachels 1999, pp. 53–54)

I expect that statistics would show a similar pattern elsewhere in the Americas and in much of Asia, although perhaps less so in Europe.

S. Maitzen (✉)
Department of Philosophy, Acadia University, Wolfville, NS, Canada
e-mail: stephen.maitzen@acadiau.ca

B. Musschenga and A. van Harskamp (eds.), *What Makes Us Moral? On the capacities and conditions for being moral*, Library of Ethics and Applied Philosophy 31,
DOI 10.1007/978-94-007-6343-2_15, © Springer Science+Business Media Dordrecht 2013

It's not only believers who treat morality as inseparable from theism. So do some educated nonbelievers, such as the atheist philosopher J. L. Mackie. Concerning his famous book-length argument against the existence of objective moral obligations, Mackie concedes the following: 'if the requisite theological doctrine could be defended, a kind of objective ethical prescriptivity could be thus introduced. Since I think that theism cannot be defended, I do not regard this as any threat to my argument' (Mackie 1977, p. 48). For Mackie, evidently, traditional theism is both necessary and sufficient for the objectivity of morals.

In this paper, I argue against the widespread tendency to see theism as the actual or potential friend of morality. On the contrary, I argue that morality depends on *rejecting* theism. Our most serious moral obligations – obligations at the heart of what we might call 'ordinary morality' – remain in place only if God doesn't exist.

As befits a philosopher's contribution, my argument is conceptual rather than empirical. I don't, for instance, argue by emphasizing the many endorsements of cruelty found in the theistic scriptural traditions. Nor do I argue that morality psychologically requires atheism except in this sense: my argument suggests that you can accept both theism and certain basic moral duties without cognitive dissonance only if you ignore theism's implications. Nor do I offer sociological data, as others have (e.g. Paul 2005, 2009), showing that belief in God is correlated with morally undesirable behaviour such as violent crime.

Instead, I argue that the existence of theism's perfect God *logically* precludes the existence of certain basic moral obligations on our part, such as the obligation to prevent or relieve terrible suffering by a child when we easily can.[1] In brief, we have that obligation only if no perfect being is *allowing* the child's suffering to occur, and hence only if no perfect being *exists*. I argue, further, that no recognizable morality remains if we lack even the basic obligation to relieve a child's suffering.

In recent years, some atheists have reacted to society's distrust of them by claiming that atheism accommodates ordinary morality just as well as theism does (see, e.g., Kurtz 2009). The truth, however, is even stronger: *only* atheism accommodates ordinary morality. Contrary to what these overly concessive atheists have said, morality *isn't* common ground between theists and atheists. Morality depends on atheism.[2]

[1] For economy in what follows, I'll refer explicitly to *preventing* suffering, rather than *relieving* it, but nothing of substance turns on this choice, because to relieve suffering is simply to prevent more, or worse, suffering.

[2] I don't *define* 'ordinary morality' in this paper, because I don't think it has a nontrivial definition. Even so, we can identify some obligations that belong uncontroversially to it. There are hard cases, but some cases are easy, such as the obligation we at least sometimes have to prevent easily preventable, horrific suffering by a child.

15.2 The Argument

According to theism, there exists a supreme being, God, possessing perfect knowledge, power, and goodness. Theism thus expresses a core doctrine of monotheistic religions – including in particular Judaism, Christianity, and Islam – ostensibly subscribed to by billions of people. It's obviously important, then, to recognize the implications of this doctrine. Having perfect knowledge, God knows whenever a child is suffering, knows how to prevent that suffering, and knows if the suffering is (or is likely to be) necessary for, or the best way of achieving, the child's overall well-being. Having perfect power, God can prevent the child's suffering. Having perfect goodness, God can't do anything morally imperfect.

With these implications in mind, consider what is unfortunately a gross understatement: somewhere in the world, right now, a child is experiencing terrible suffering that the child doesn't want and doesn't deserve. Now suppose that God, although having the knowledge and power to prevent it, lets the aforementioned child experience terrible suffering, not because the child will ultimately benefit from it but for some other reason, or perhaps for no reason at all. The suffering is intense, the child doesn't deserve to undergo it, and the child doesn't volunteer for it (as someone might volunteer for the pain of donating bone marrow). In allowing the suffering, God exploits the child and thereby acts imperfectly.

The treatment is what's important here, not the exact term we use for it, so you can reject my assumption that we actually use the word 'exploit' in this way. My claim is that no supreme being could treat the child in the way I've described, whichever label we choose for that treatment. To put it mildly, there's something less than perfect about letting a child suffer terribly *for* the primary benefit of someone else – whether for the benefit of a bystander who gets a hero's chance to intervene, or for the benefit of a child-abuser who gets to exercise unchecked free will. If you doubt the previous sentence, consider whether you would dream of letting a child you love suffer abuse in order to secure either of those benefits.

But is it always wrong, one might ask, to exploit people? Don't we sometimes justifiably use innocent people for the benefit of others? If the child in my example contracts an untreatable, fatal and highly contagious disease, might we not justifiably quarantine him or her when it's the only (or best) way to prevent the spread of the disease? Might we not justifiably isolate the child in a way that benefits others at the expense of the child? Don't we justifiably perform triage, letting some patients suffer so we can attend to more urgent cases? Yes. But these practices reflect our imperfection: it's only limitations in our knowledge and power (in this case, medical) that make us resort to triage or quarantine. We regret having to do it; we wish we had the resources to make these practices unnecessary.

A perfect God, however, isn't subject to our limitations in knowledge or power, or indeed any real limitations in knowledge or power. So no perfect God has an excuse for exploitation, even if we sometimes do. Furthermore, if God were to face an actual moral *dilemma*, a case in which he does something immoral no matter what he does, then he wouldn't count as morally perfect, on the obvious assumption that a morally perfect being never does anything immoral. If I'm right, then God can't possibly allow a child's intense, undeserved and involuntary suffering *unless* the suffering is necessary (or if not necessary then optimal) for the child's overall benefit.[3]

Yet many children endure intense, undeserved and involuntary suffering every day, all over the world. What does ordinary morality tell us to do when we encounter them? Obviously, we ought to act compassionately toward them. We have a duty to prevent their suffering, at least when we easily can. Their suffering is very bad for them, which is the most important reason we ought to act with compassion in the first place. But wait. If God exists, then that suffering must be needed – somehow, even if we can't see how – for the overall benefit of those very sufferers. In that case, what happens to the moral duty we thought we had to prevent their suffering? It disappears.

If a perfect God exists, then any suffering that occurs is suffering that God allows to occur, since any perfect being has the power to prevent any occurrence, including any case of suffering. If perfection also rules out exploitation, as I've argued it does, then God allows the suffering of children only if those children ultimately benefit *from* the suffering. The word 'from' matters here. It's not enough if God merely compensates children for suffering he lets them endure; the suffering must be necessary (or at least optimal) for their greater good.

Why? Because compensation doesn't count as justification. I can compensate you after harming you, and indeed a court of law may make me compensate you, but no amount of compensation will *justify* my harming you. The only thing that could justify my harming you would be my need to harm you in order to stop you from harming me or some innocent third party. Likewise, then, for God's permission of a child's suffering: the suffering must be needed, or at least optimal, for the child's overall benefit. Otherwise God can't justifiably allow it.

But if the suffering is needed, or optimal, for the child's overall benefit, then it's like the pain from a needle when the needle is the only (or the best) way to deliver a vaccine. If so, then we never have a duty to prevent the suffering of children; after all, we don't think we have a moral obligation to prevent painful vaccinations when they're beneficial. Even those who oppose childhood vaccinations wholesale do so because they think vaccines do more harm than good. We don't think it's truly compassionate to prevent all vaccinations just because needles hurt.

If we never have a moral obligation to prevent suffering by children – a consequence implied by the core doctrine of theism – then which moral obligations *do* we have? None, as far as I can see. I can't see how we can be objectively obligated

[3] A number of prominent theistic philosophers have defended precisely this reasoning, among them the Christian philosopher Eleonore Stump, who writes that 'if a good God allows evil, it can only be because the evil in question produces a benefit for the sufferer and one that God could not produce without the suffering' (Stump 1985, pp. 411–412) and 'other things being equal, it seems morally permissible to allow someone to suffer involuntarily only in case doing so is a necessary means or the best possible means in the circumstances to keep the sufferer from incurring even greater harm' (Stump 1990, p. 66).

to refrain from theft, fraud, bigotry, or slander if we never have the even more basic obligation to prevent suffering by children. If we lack a moral obligation to prevent even the worst suffering by children, then morality falls apart, or at best it becomes frivolous because it no longer concerns the most serious kinds of harm.

Theism causes a related problem as well. Suppose I'm wrong and we *can* reconcile God's existence with a duty on our part to prevent at least some suffering. Theism still encourages a bizarre 'reverse triage': the *worse* an innocent person's suffering, the more reason theism gives us for thinking that the suffering must be needed for the sufferer's own good, and hence the *less* reason it gives us to prevent the suffering. Theism implies that we ought to prevent mild suffering first, extreme suffering later. Far from shoring up our moral outlook, adding God to it turns it upside down.

15.3 Can God Exploit?

Some might try to answer my argument by rejecting its key premise, by countering that a perfect God *can* allow a child to suffer for the primary benefit of others. Oxford philosopher Richard Swinburne, for instance, says that God has moral permission to exploit any human being because all human beings owe their existence to God, God is on balance their benefactor, and furthermore 'being of use is a good for the victim' who gets used (Swinburne 1995, p. 81). Does Swinburne's reply work? Not at all, as we'll see.

Imagine that I clone a child into existence from a single one of my skin cells, and I treat the child splendidly for all but the final minute of its life. But during that final minute, I let someone abuse the child to death in order to show onlookers just how revolting child abuse is and thereby deter them from ever abusing a child. Think of it as aversion therapy. The child owes its existence to me (via my use of technology), and I'm on balance its benefactor, treating it well for all but the final minute of its life. Moreover, its horrific death isn't purely gratuitous; it serves as an object lesson for the benefit of others, not only deterring some potential child abusers but also protecting children they might otherwise have abused. Nevertheless, in this story I behave imperfectly, to say the least. Yet I behave just as Swinburne imagines God does. Even granting Swinburne's premises, therefore, his conclusion doesn't follow. His defence of exploitation on the part of a perfect God therefore fails.

15.4 Net Benefit?

I've argued that no perfect being can exploit a child by allowing the child to experience undeserved, involuntary suffering unless it's necessary or optimal for the child's overall good. In other words, the suffering must be necessary, or at least optimal under the circumstances, for securing a *net benefit* for the child, whether in this life or the next. Maybe the suffering is an essential *means* of securing the child's net benefit, or maybe it's an unavoidable *by-product* of doing so. One might wonder, however, whether a perfect being must secure for the sufferer a net benefit or instead merely *some* benefit or other.

To see why perfection requires securing a net benefit for the sufferer, imagine that God lets Jack endure undeserved, unwanted, and unbearable pain because it's the only way to get Jill (the object of Jack's unrequited affection) to send Jack, of her own volition, a get-well card that he'll read just before he dies from his painful condition. Jack secures *some* benefit from the suffering – a freely sent get-well card from Jill – but suppose that his suffering is involuntary in that he wouldn't regard the benefit as remotely worth the suffering even if he knew that not even God could produce the benefit any other way. Surely God's conduct in that case falls short of moral perfection. It falls short even if we also suppose that Jack's suffering produces significant benefits for *others* obtainable no other way; maybe news of his suffering triggers generous donations that his hospital wouldn't otherwise have received. It falls short of moral perfection because it's unjust to Jack. It violates a moral standard demanding fairness in the treatment of individual persons, a standard that no perfect being could have an excuse for violating even if we imperfect beings perhaps could.[4] Jack gets some reward but not enough, because his reward fails, by any reasonable measure, to *offset* his suffering.

15.5 Not Truth but Belief

I've argued that the truth of theism would undermine an obligation at the heart of ordinary morality: the obligation we have on at least some occasions to prevent undeserved, involuntary suffering, such as that experienced by children. One might object, however, that it's not the truth of theism that undermines this obligation so much as the *belief* that theism is true. According to this objection, we lack an obligation to prevent suffering only if we *believe* that suffering always benefits the sufferer, regardless of whether our belief is true.

I won't try to settle here the complex issue of how our obligations depend on our beliefs. Even if the objection succeeds, however, it's noteworthy enough if ordinary moral obligations dissolve in the presence of theistic *belief*. But I doubt the objection succeeds in any case. Our ordinary moral obligation to prevent at least some kinds of suffering surely depends on the presumption that suffering is in fact often bad overall for the sufferer. Granted, we recognize that suffering is sometimes on balance worth it for the sufferer; otherwise we'd feel obligated to prevent every vaccination that hurts or every surgery that leaves the patient with a painful recovery. But we feel obligated to prevent suffering in other cases because

[4] Note that this moral standard can constrain the conduct of a *perfect* being even if the standard isn't part of *ordinary morality* – which, as the label suggests, concerns the conduct of imperfect beings like us.

we confidently presume that the suffering *isn't* in the sufferer's best interest, or is at least vanishingly unlikely to be. Consider, for instance, the case of the 4-year-old boy in Michigan who was tortured to death for wetting his pants.[5] Had you been in a position to prevent that torture, easily and at no risk to yourself, ordinary morality would have regarded you as obligated to try. Why? At least partly because, we assume, the torture wasn't – or was vanishingly unlikely to be – in the boy's overall interest. Ordinary morality itself thus presupposes that not all undeserved, involuntary suffering is for the overall good of the sufferer, whereas theism implies that it *must* be. Hence theism and ordinary morality conflict.

15.6 Autonomy

One might try to reconcile theism and ordinary morality by showing that we can have a moral obligation to prevent some action *even if* we know the action won't cause anyone to suffer except as needed for securing his or her overall good. Suppose that an otherwise normal adult is unwilling to receive a vaccination that he knows will benefit him overall (even factoring in the pain of the needle) by making him immune to a virus ravaging his community. According to this objection, we can have a duty to prevent the vaccination, at least if we easily can, because it violates the recipient's *autonomy*.[6]

The objection fails for three reasons. First, it's not at all clear that an otherwise beneficial violation of an agent's autonomy is something that a *third party* has a duty to prevent. Even if I have a duty not to violate your autonomy by benefitting you against your will, it's another matter whether (say) your friend has a duty to try to stop me from doing so. The former duty wouldn't imply the latter.

Second, the objection is irrelevant to my argument. I've argued that theism threatens our ordinary moral obligation to prevent *suffering*. Yet in the objector's example it's not suffering that we're obliged to prevent but instead a violation of autonomy, and violations of autonomy can be painless and no less violations of autonomy for being painless. (Imagine a serious violation of your privacy that you never learn about.) So the example doesn't in fact rescue an obligation to prevent suffering.

Third, the objection has very limited scope even if it does establish some obligation to intervene: it works only for autonomous agents. Yet our ordinary moral obligation to prevent suffering is often weightiest when the sufferer *lacks* autonomy, as in the case of a young child. We have an especially strong moral obligation to prevent the torture of a child when we easily can but no obligation at all to prevent an unwilling child's beneficial vaccination. Autonomy, equally lacking in both cases,

[5] 'Police: Man tortured 4-year-old to death for wetting his pants,' http://www.cnn.com/2010/CRIME/04/15/michigan.child.torture (accessed 26 May 2011).

[6] I owe this objection to Robert Lovering.

does nothing to account for this difference. Our duty to prevent the torture, but not the vaccination, stems from something else: the *extreme net harm* that the torture causes the child. But I've argued that extreme net harm can't *happen* to the child if God exists.

15.7 Free Will

In my experience, theists feel strongly inclined to reply to my argument by claiming that God must never interfere with the freedom of a human agent, not even to stop the agent from torturing a child, or at least that God's desire to respect the torturer's freedom can justify God in allowing the torture.[7] On this view, contrary to what I've argued, God can allow suffering that it then becomes *our* duty to prevent, and so we retain the ordinary moral obligation that I've argued theism threatens. The obvious rejoinder, however, is that God's allowing child-torture *so as* not to interfere with the torturer's freedom is a clear case of exploiting the child for some other end, something no perfect being could do. Indeed, it's worse than imperfect; it's morally monstrous.

If anything, this frequent refrain about the sacrosanct value of human freedom shows just how alien theism is to our ordinary moral outlook. If you decide to play the role of spectator while a child is tortured, even though you could easily stop the torture at no risk to yourself, ordinary morality won't excuse your inaction because you didn't want to curtail the torturer's freedom.[8] It's hard to see why theists imagine that the very same excuse could exonerate a morally perfect God.[9]

15.8 Heaven and Retrospective Consent

According to a theodicy I call 'Heaven Swamps Everything', compensation paid to an exploited individual *can* justify or excuse the original exploitation, provided the compensation is big enough. On this view, God can justifiably let a child be tortured provided that God eventually sends the child to heaven, even if the child's suffering is in no way *necessary* for attaining heaven. Again, however, such reasoning conflicts with ordinary morality because it *conflates* compensation and justification,

[7] Theistic appeals to free will arise whenever I present this argument, including at a session of the American Philosophical Association Eastern Division Meeting at which my commentator based his criticism of my argument almost entirely on the idea that God would never curtail human freedom.

[8] As Derk Pereboom notes, from the ordinary moral perspective 'the evildoer's freedom is a weightless consideration, not merely an outweighed consideration' (Pereboom 2005, p. 84, citing and expanding on Lewis 1993, p. 155).

[9] For a more detailed refutation of the free-will reply, see Maitzen (2009, pp. 120–122).

and it may stem from imagining an ecstatic or forgiving state of mind on the part of the heaven-dweller: in heaven no one bears grudges, even the most horrific earthly suffering is as nothing compared to heavenly bliss, and all past wrongs are forgiven. But ‘are forgiven’ doesn’t mean ‘were justified’: the blissful person’s disinclination to dwell on his or her earthly suffering doesn’t imply that a perfect being was justified all along in allowing it. By the same token, our ordinary moral practice recognizes a legitimate complaint about child abuse even if, as adults, its victims should happen to be on drugs that make them uninterested in complaining. Even if heaven swamps everything, it doesn’t thereby justify everything.

Alternatively, one might suppose that, assuming everyone goes to heaven, everyone on due reflection eventually *consents* (after the fact) to any undeserved and otherwise involuntary suffering he or she experienced while on earth.[10] But this response does nothing to diminish the theistic threat to ordinary morality: our ordinary moral obligation to prevent at least some undeserved, involuntary human suffering disappears on the assumption that its victims will always on due reflection eventually consent to that suffering.[11]

15.9 God’s Commands

Theists might reply that we have a moral obligation to prevent suffering, at least on some occasions, not because such suffering is bad overall for the sufferer but simply because God has *commanded* us to prevent suffering. In that case, allegedly, theism would be consistent with our ordinary moral obligations after all.

This reply fails for three reasons. First, it requires what Jeff Jordan calls a ‘recalibration’ of ordinary morality: ‘the replacement of concern and sympathy and compassion with the obedience to commands. One alleviates suffering not out of compassion for the sufferer, but rather because one is told to do so’ (Jordan 2004, p. 176). Therefore, even if this reply manages to rescue a duty to prevent involuntary, undeserved suffering, it doesn’t rescue the ordinary duty to prevent such suffering out of compassion for its victim, because again the suffering in question is necessary or optimal for the sufferer’s greater good if God exists.

Second, the reply has the puzzling upshot that God has commanded us to prevent undeserved, involuntary human suffering, at least when we easily can, even though such suffering *always benefits* the sufferer overall. Why, then, have us prevent it? Notice that distinguishing *acts* from *rules* won’t help here, at least not if the

[10] Compare Alston (1996, p. 112), which defends a view quite close to this one.

[11] Yet another view is that intense suffering is always a gift from God, a blessing, in part because it is an analogue of Christ’s suffering. Christopher Hitchens attributes this view to Mother Teresa of Calcutta (Hitchens 1995, p. 41). Even if we ignore the highly questionable features of this view, it fails to blunt the theistic threat to morality for which I argue here, since if intense suffering is always a blessing in disguise, we never have an ordinary moral obligation to prevent it.

sufferer's well-being is our paramount concern, as it should be. Rule utilitarians say that they can consistently endorse a rule adherence to which *sometimes* fails to maximize utility provided that following the rule *generally* maximizes utility. I've argued, however, that our following the rule 'Prevent involuntary, undeserved suffering' *never* maximizes the sufferer's utility if God exists, because we never thereby increase the sufferer's total utility relative to what God would have secured for him or her in any case. Not even the staunchest rule utilitarian will endorse a rule requiring us to prevent suffering if following the rule never increases the sufferer's utility.

Third, just where *has* God commanded us to prevent suffering? That particular command isn't easy to find in the monotheistic scriptures, to put it mildly. Moreover, we've seen that a generic command to act *compassionately* doesn't translate, if God exists, into a command to prevent suffering. For a command to act compassionately isn't a command to show *misguided* compassion – for instance, compassion that causes someone to prevent even those vaccinations that greatly benefit their recipients. By the same token, then, God doesn't command equally misguided compassion compelling us to prevent suffering that, I've argued, must always produce a net benefit for the sufferer.

15.10 An Imperfect God?

In this paper, I've assumed the classical theistic conception of God as possessing perfect knowledge, power, and goodness. Some may try, therefore, to evade my argument by positing a God who lacks one or more of those divine perfections. Because I've argued that the existence of a *perfect* God would undermine morality, the response 'Well, that's not *my* God!' doesn't of course refute my argument so much as it changes the subject. But let me give what I think are compelling reasons *not* to change the subject.

Serious theological problems arise for those who imagine that their God has limitations and imperfections. To begin with, this view of God rules out any *a priori* arguments for God's existence, such as the ontological argument, that proceed from the mere concept of a perfect being; indeed, it abandons the entire project of perfect-being theology. But worse, this view invites all manner of awkward questions about the God it imagines. If God is imperfect, why think that God always has existed and always will exist? An imperfect God might be only finitely old and might go out of existence just when we need him most! If God is imperfect, why think that God has the power to make the universe out of nothing, or even the power to fashion the universe out of pre-existing stuff, or the power to achieve justice in the end? The affirmation 'With God, all things are possible'[12] is supposed to comfort believers,

[12] Matthew 19:26 (KJV); for similar affirmations, see also Job 42:2, Jeremiah 32:17, and Luke 1:37, all cited in Leftow (2011), p. 106.

but if God is imperfect, what assurance do they have that all things *are* possible with God? Classical theism avoids those awkward questions by insisting on God's perfection. Furthermore, the more limited and imperfect one imagines God to be, the more God resembles the deities that polytheistic religions invoke to explain various aspects of the natural world: a god for the Sun, another for the Moon, another for fertility, and so on. But surely deities of that sort have been *outmoded* by science's ability to explain those aspects of the universe in purely naturalistic terms.

With respect to perfect knowledge in particular, the non-classical version of theism known as 'open theism' claims that God lacks infallible knowledge of the future, or at least the part of the future that depends on the libertarian free choices of creatures. But this departure from classical theism doesn't evade my argument. For even on open theism, God can surmise (and surely better than we can) that a child is *about* to be tortured by a libertarian free agent or, at a minimum, that the agent is likely to *continue* the torture once it's underway.[13] It doesn't take anything close to infallible knowledge to make those judgments. Possessing the requisite power and benevolence, therefore, even the God of open theism would prevent child torture (or its continuation) unless again, in God's supernaturally best judgment, it was necessary or optimal for the child's greater good. Hence open theism fails to avoid the dire implications for morality that I've argued stem from classical theism.

In my experience, those who respond to my argument by invoking God's imperfection more often portray God as lacking the *power* to prevent suffering that he sees coming and would prevent if only he could.[14] Again, problems arise for any view that retreats on God's power. First, even if God may lack the power to prevent *every* case of horrific suffering by children, it strains credulity to think that the suffering that God makes it his priority to prevent would have been even worse than the horrific suffering by children that God *does* allow to occur.

Second, why think that a God lacking perfect power would still possess perfect knowledge, perfect goodness, or indeed perfection in *any* attribute? The classical Anselmian conception of God famously treats the divine perfections as a package deal: reasoning from the concept of a *greatest possible being*, we're supposed to conclude that God must have *every* intrinsically great-making quality in its highest possible degree. On this view, the divine perfections stand or fall together. Hence abandoning one of the perfections requires abandoning all of them, which in turn risks portraying God as merely a powerful extra-terrestrial, existing only contingently, finitely old, mortal, and so on. Without the classical Anselmian conception of God, we lack *a priori* grounds for ruling out that scenario, and it's hard to see how we could have *a posteriori* grounds for ruling it out either.[15]

[13] As William Hasker, himself an open theist, emphasizes (Hasker 2010, p. 308).

[14] See, e.g., Gellman (2010), responding to Maitzen (2009) and replied to in Maitzen (2010). Oddly, Gellman explicitly declares that he's describing a perfect God (2010, p. 191), but I see no way to characterize as *perfect* a God who suffers from the apparently severe limitations in power that Gellman sketches in his article.

[15] See Oppy (2011) for discussion of a similar worry.

Third, and relatedly, once you allow for the possibility of an imperfect God it seems *arbitrary* to abandon omnipotence rather than omniscience or omnibenevolence. Maybe God knows perfectly well how horrifically children sometimes suffer and could prevent it if he wanted to, but he finds their suffering entertaining or beautifully poignant. Or maybe God lets children suffer horribly *precisely so that* we can have serious moral obligations to intervene on their behalf,[16] which I submit would be a morally abominable case of God's exploiting those children. Furthermore, once you admit God's imperfection, where do you stop? If human and animal suffering is evidence at all of God's imperfection, it's by no means clear that it isn't evidence of a *radically* imperfect God: a God too impotent, or ignorant, or morally indifferent to intervene in *any* of the cases of undeserved, involuntary suffering we observe – and a far cry from the God that theists say they worship.

To sum up, I've argued that the existence of our most basic moral obligations logically implies the nonexistence of any perfect being, such as theism's perfect God. Logically speaking, morality *isn't* common ground between theists and atheists. Theism logically precludes, for example, your moral obligation to prevent terrible suffering by a child even when you easily enough can prevent it. In terms, therefore, of the question that motivates this volume – What makes us moral? – the answer is, in part, a logical (if often unacknowledged) commitment to atheism.

References

Alston, W.P. 1996. The inductive argument from evil and the human cognitive condition. In *The evidential argument from evil*, ed. D. Howard-Snyder, 97–125. Bloomington: Indiana University Press.

Edgell, P., et al. 2006. Atheists as 'other': Moral boundaries and cultural membership in American society. *American Sociological Review* 71: 211–234.

Gellman, J. 2010. On God, suffering, and theodical individualism. *European Journal for Philosophy of Religion* 2: 187–191.

Hasker, W. 1992. The necessity of gratuitous evil. *Faith and Philosophy* 9: 23–44.

Hasker, W. 2010. Defining 'gratuitous evil': A response to Alan R. Rhoda. *Religious Studies* 46: 303–309.

Hitchens, C. 1995. *The missionary position: Mother Teresa in theory and practice*. London: Verso.

Jordan, J. 2004. Divine love and human suffering. *International Journal for Philosophy of Religion* 56: 169–178.

Kurtz, P. 2009. Opening statement by Paul Kurtz. In *Is goodness without God good enough? A debate on faith, secularism, and ethics*, ed. R.K. Garcia and N.L. King, 25–39. Lanham: Rowman & Littlefield.

[16] As proposed by Hasker (1992) and defended in Hasker (2010). Hasker's explanation assumes that God has no moral obligation to prevent such suffering, and hence God can't be faulted for letting it occur, even though God's letting it occur creates for *us* a moral obligation to prevent it if we easily enough can, an obligation *we* can be faulted for failing to honor. I can't see how God could permissibly delegate such an obligation – that is, delegate it without thereby *exploiting* the sufferers in a morally objectionable way.

Leftow, B. 2011. Why perfect being theology? *International Journal for Philosophy of Religion* 69: 103–118.

Lewis, D. 1993. Evil for freedom's sake? *Philosophical Papers* 22: 149–172.

Mackie, J.L. 1977. *Ethics: Inventing right and wrong*. Harmondsworth: Penguin.

Maitzen, S. 2009. Ordinary morality implies atheism. *European Journal for Philosophy of Religion* 1: 107–126.

Maitzen, S. 2010. On Gellman's attempted rescue. *European Journal for Philosophy of Religion* 2: 193–198.

Oppenheimer, M. 2010. Atheists debate how pushy to be. *New York Times*, October 16, p. A12.

Oppy, G. 2011. Perfection, near-perfection, maximality, and the Anselmian God. *International Journal for Philosophy of Religion* 69: 119–138.

Paul, G. 2005. Cross-national correlations of quantifiable societal health with popular religiosity and secularism in the prosperous democracies: A first look. *Journal of Religion and Society* 7: 1–17.

Paul, G. 2009. The chronic dependence of popular religiosity upon dysfunctional psychosociological conditions. *Evolutionary Psychology* 7: 398–441.

Pereboom, D. 2005. Free will, evil, and divine providence. In *God and the ethics of belief*, ed. A. Dole and A. Chignell, 77–98. New York: Cambridge University Press.

Rachels, J. 1999. *The elements of moral philosophy*, 3rd ed. Boston: McGraw-Hill.

Stump, E. 1985. The problem of evil. *Faith and Philosophy* 2: 392–423.

Stump, E. 1990. Providence and the problem of evil. In *Christian philosophy*, ed. T.P. Flint, 51–91. Notre Dame: University of Notre Dame Press.

Swinburne, R. 1995. Theodicy, our well-being, and God's rights. *International Journal for Philosophy of Religion* 38: 75–91.

Chapter 16
What Makes the Martyr (Im)Moral?

Anton van Harskamp

The Martyr ...! Who is he or she? Today martyrs are no longer identified with radical believers who witness, by suffering, the glory of God (the word 'martyr' derives from the Greek for a first- hand witness). No, these days 'martyrdom' is often associated with suicide terrorists. Suicide terrorism seems to be a new form of martyrdom: 'predatory martyrdom' (Barlow 2007, pp. 3, 129ff.).

Many Westerners will deem it evident that suicide terrorism is extremely immoral. In the world of suicide terrorism however, suicidal 'predatory' actions are considered to be actions for a good cause – a terrorist, writes Bruce Hoffman, also the suicide terrorist, 'is fundamentally an altruist' (Hoffman 2006, p. 37; cf. Post 2007, p. 16; Bloom 2007, pp. 76f.; Merari 2010, pp. 145ff.). These contradictory moral evaluations of suicide terrorism bring forward the issue of justification of 'predatory martyrdom': what precisely 'makes' the new martyrs utterly immoral or utterly moral? What justifies the act of the martyr as moral, or what denies the morality of suicide terrorism? In this essay we'll deal with the latter question. The views of three eminent moral philosophers will be presented. All three argue that suicide terrorism is (almost) always terribly immoral: Michael Walzer (2004), Igor Primoratz (2004, 2007/08) and Marcel Hénaff (2007/08). We will propose that the arguments of these philosophers are not *per se* false, but may not deal with the question satisfactorily for two reasons. First, because they tend to overlook the interplay of terrorism with cultural, economic, political and individual violations by non-state actors, by state actors and by structural violence (the *trias violentiae,* Schinkel 2010) *and,* second, because they don't reckon with the religious dimension of the actions of the suicide bombers. In the last sections of this paper we'll argue that this religious dimension may induce a Kierkegaardian teleological suspension of ethics, of thinking about the (im)morality of suicide terrorism. The question has

A. van Harskamp (✉)
Faculty of Philosophy and Department of Social and Cultural Anthropology, VU University, Amsterdam, The Netherlands
e-mail: anton@bezinnen.nl

B. Musschenga and A. van Harskamp (eds.), *What Makes Us Moral? On the capacities and conditions for being moral*, Library of Ethics and Applied Philosophy 31,
DOI 10.1007/978-94-007-6343-2_16, © Springer Science+Business Media Dordrecht 2013

to become: ‘What makes this religious act of the suicide bomber a good or a false religion?’! Only if we ask this question first – not a very popular one these days – we may be entitled to seek for an answer to our main question. Let’s now first start with trying to define suicide terrorism.

16.1 Defining Suicide Terrorism

Defining ‘suicide terrorism’ is not that easy. In the literature on terrorism one may find three problems with which ‘definers’ of (suicide) terrorism are struggling.

First, the problem that the concept ‘terrorism’ – of which suicide terrorism is a subcategory – refers to a phenomenon that seems to resist ‘objective’ definition: many definitions in our days are loaded with a pejorative, political subjectivity (Jenkins 1980; cf. Neumann 2009, pp. 6ff.). Defining terrorism pejoratively however, suggests that we live in a Manichean world of a good ‘self’ and a bad or even evil ‘other’. Just like the concept of ‘fundamentalism’ once was used in Western liberal discourses as the rejection of the enemy other, so today the concept of terrorism is often used as a form of Othering, a basically political construction of the world in ‘Us’ and evil ‘Them’, so Asad in his *On Suicide Bombing* (Asad 2007, passim; cf. Nanninga 2007, pp. 47–48). So an inherently pejorative way of defining terrorism does not really touch on the object itself, but is an expression of a simplified worldview, often a surrender to the false myth of pure evil (Eller 2010, pp. 16–17).

Secondly there is the problem of the inclination to define terrorism as the intentional, singular acts of evil people. But as sociologist Charles Tilly has made clear, we cannot define terrorism as singular acts of some, politically deluded people with an evil disposition (Tilly 2004, cf. Schinkel 2010, pp. 136–153). Terrorism is an always conditional process (Atran 2006, p. 13; Bloom 2007, pp. 79, 90), of which historical and actual events, as well as reactions (and denunciations) of the states and parties which are fighting ‘wars on terrorism’, are part. So we have to deal with the observation that terrorism has to be presented as part of the history of the conflicts in which terrorist actions are embedded.

Thirdly there is the problem of the determination of the overall goal of terrorism. In many existing definitions it is claimed that terrorism is purely political. Bruce Hoffman defines terrorism as the deliberate creation and exploitation of fear through violence *in the pursuit of political change* (Hoffman 2006, p. 40). Domenico Tosini states: terrorism is the use of violence against civilians by non-state entities *for specific political purposes* (Tosini 2007, p. 667). In the recently designed Academic Consensus Definition of Alex Schmid – a definition based on responses of 50 experts – terrorism is considered to be at the one hand a doctrine about a tactic of coercive, political violence, and at the other hand a calculated, demonstrative practice performed for its propagandistic and psychological effects (Schmid 2011, pp. 86–87). One may ask however whether terrorism, in particular suicide terrorism, cannot also be considered to be a practice performed for its ‘theatrical’ effects, as Mark Juergensmeyer has made clear (Juergensmeyer 2003, pp. 121–147). And one may wonder whether this horrible theater of suicide and murder does not imply that it has to be defined as being more than 'political'. Sociologist Jeffrey Alexander, writing about 9/11 as a message to the world in the form of a horrific public, 'theatrical' performance, considers on the one hand terrorism as a form of political action, but indicates on the other hand that terrorism is also a post-political phenomenon. The seemingly non-rational use of violence, the isolation of terrorism from 'normal' political structures, and above all the stunning vagueness of the religious ideology of most global terrorists, e.g. an Islamic millennium, all these characteristics point to a departure from what is 'normal' politics. Rather than defeating its opponents through political struggle, the seemingly demonic willingness to sacrifice the lives of others and their own, seeks mainly to draw blood and to deliver chaos and reciprocal despair. As such, according to Alexander, terrorism may start with politics but also overwhelms the political world (and our secular worldview) with acts which transcend the political (Alexander 2004, pp. 88ff.; cf. Wieviorka 2009, pp. 95, 112–113). Just because of the possibility of this seemingly 'transcendent', maybe even religious or quasi-religious nature of suicide terrorism, it is so difficult to establish a definition of terrorism.

So far we may conclude that in a definition: (a) we may not express our normative political subjectivity; (b) we may not suggest that terrorism is an isolated phenomenon or that terrorism is an action of evil or deluded perpetrators; and (c) we may assume that the nature of terrorism is characterized by a transcendence of 'the political'. Now, when we realize that these acts of suicidal violence are motivated by spreading fear amongst a non-combatant collectivity which is presumed to be the enemy and the oppressor of the terrorists (cf. the survey of definitions of terrorism by Vallis et al. 2006, p. 7), we may define suicide terrorism as:

> The practice of direct suicidal violence against non-combatants by non-state actors who claim to react against structural or political violence by a presumed adversarial collectivity. This practice is performed for its political, propagandistic and psychological effects and demonstrates a post-political 'theatrical' symbolism.

16.2 Philosophers on Suicide Terrorism

Let's now discuss three moral condemnations of suicide terrorism. And let's say before we start, that when we are going to criticize the philosophical-moral critique on suicide terrorism, we are not going to defend or justify suicide terrorism. We will 'only' bring forward that the grounds for condemning suicide terrorism have to be found on a deeper level than the level of 'normal' ethics.

16.2.1 Michael Walzer

In *Arguing about war* (2004) Michael Walzer postulates that terrorism is utterly wrong. Why? Because of 'the killing of innocent people', and because of 'the intrusion of fear into everyday life, the violation of private purposes, the insecurity of public spaces, the endless coerciveness of precaution' (Walzer 2004, p. 51). Terrorism is according to Walzer even worse than the crime of murder, because a murder victim has been chosen for a purpose, while in terrorism the victims are chosen indiscriminately. Walzer actually considers terrorism as a manifestation of evil; as the deliberate and wilful harming, misusing, humiliating or destroying of innocent others (Zimbardo 2007).

Walzer's specification that terrorism is utterly wrong, even evil, because of the intrusion of fear and the violation of private purposes, the insecurity of public spaces, the endless coerciveness of precaution – he writes on the effects of 9/11 on American society – indicates that his approach to terrorism is purely political (as he himself suggests, 2004, p. xiv). He seems to assume that politics is the field of a struggle of interests, a field where enemies have to be expelled. So it is no surprise that, precisely because he deems terrorism to be evil, he endorses the need for an expulsion of evil, which actually is a 'war on terrorism' (2004, pp. 136ff.) We may, however, surmise that his purely political approach falls short of understanding the horrifying mystery that suicide terrorism is. Why could that be the case? There is, to begin with, the framing of terrorism into the theory of just war. For Walzer it seems to be clear that starting a war and waging a war can sometimes (theoretically) be legally and morally justified, while, compared to war, terrorism should always be wrongful, because of the violation of the principle of innocence (2004, pp. 36, 136, 152). However, against the framing of terrorism into the theory of just war one may bring forward that just war theory was developed as a moral doctrine to govern conflicts among sovereign nation-states (Schwartz 2004, p. 292). And terrorism starts and is perpetrated by non-state actors, by those who claim not to have the political instruments of nation-states. Now, it seems to be that, thinking from the legal and moral perspective of *ius ad bellum* (say: *ius ad terrorem*) Walzer excludes beforehand the possibility that non-state actors can have a morally legitimate conflict with states. He seems to imagine a political universe in which moral legitimacy can only be an issue in conflicts between nation-states. And thinking about the level of *ius in bello* (say: *ius in terrore*) we may ask whether Walzer's framing of terrorism into the just war theory does not imply some idealization of 'normal' wars? We suspect that also states, modern states, have pursued and will pursue terroristic policies, and will violate the principle of innocence, for instance by bombing and shelling civilians. In short, Walzer does not accept that terrorism, being not an expression of a conflict between nation-states, can neither be justified nor condemned like just, respectively unjust war' can (Morton 2005, p. 85). It is further remarkable that when Walzer refutes some of the excuses for terrorism (Walzer 2004, pp. 53–60), and counters the argument that terrorism is the last possible option for the weak and the powerless, he uses only one argument: 'every actor is a moral agent and makes an independent decision' (Walzer 2004, p. 60). This argument reveals a focus strictly on the evil individual perpetrator. This argument is an expression of a way of thinking that in the end there are no external factors which can influence, determine or 'make' the moral responsibility of an individual actor. So Walzer apparently focuses on the evil actor alone, while he seems to deny that when we wish to evaluate terrorism, we should not focus on the evil perpetrator alone, but we should take into account the reaction of terrorists on the (presumed) violence of their enemies. One may ask whether Walzer uses an inherently contradictory argumentation: if it is the case that a moral act is only the responsibility of an individual actor, how can one then pass a moral judgment – in Walzer's case a basically political judgment – on a political, hence public act? And there is a flaw in Walzer's approach. This flaw comes with his very own version of the just war theory. In *Just and Unjust Wars* (2000) he worked out an argument on emergency ethics. He reiterated this argument in *Arguing about War* (2004) by stating that in the event of a supreme emergency, i.e. when the survival of a political community is at stake, a leader of a nation, representing his political community, may, no perhaps even *has to*, override fundamental humanitarian principles, like the principle of not killing innocents (Walzer refers to the years 1940 and 1941: Churchill sending bomber planes to German cities). An individual, argues Walzer, may never kill innocents, even if he can save himself by doing so. 'A moral president' however, or 'a prime minister or military commander' (Walzer 2004, p. 41) is allowed in situations of supreme emergency to kill innocents, 'if and only if' he knows he is doing wrong! Walzer's arguments take a peculiar turn here, for he gives an existential twist to the inner feelings of guilt of the wrongdoing and simultaneously rightful acting leader:

> A morally strong leader is someone who understands why it is wrong to kill the innocent and refuses to do so, refuses again and again, until the heavens are about to fall. And then he becomes a moral criminal (like Albert Camus's 'just assassin') who knows that he can't do what he has to do – and finally does. (Walzer 2004, p. 45)

This argument is not convincing. Why? Because Walzer seems to consider a purely individual feeling of guilt as a precondition for public political-moral acting. But, as Talal Asad has observed: guilt of an individual strong leader is neither a legal nor a moral-political judgment; it is only a private feeling, an individual sensibility (Asad 2007, p. 18). The existential possibility of the presence of feelings of guilt in an individual can never be an argument for condoning the acts of any political leader. So, one may suppose that Walzer has a double moral standard. As Andrew Valls rightfully asks about Walzer's views: 'But why is it that the territorial integrity of, say, Britain [or, let's add, the U.S., AvH] justifies the resort to ... violence that targets civilians [the U.S. in Iraq and Afghanistan. AvH] but the right of self-determination of a stateless nation never does?' (Valls 2000, p. 73; also cited by Steinhoff 2004, p. 105). Why does Walzer actually condone existentially driven violations of basic moral standards like not killing innocents of (Western) states , and can he be quite certain that similar violations of terrorists are only evil?[1]

[1] Walzer's answer could be: 'Because terrorists are never in a situation of supreme emergency'. He writes: 'Terrorism has not been a means of avoiding disaster but of reaching for political success' (Walzer 2004, p. 54). One may wonder whether this 'political' answer holds: how can Walzer be sure that terrorists do not find themselves in a situation of supreme emergency?

16.2.2 Igor Primoratz

Let's take a look at another moral consideration of terrorism: the Israeli philosopher Igor Primoratz. He argues in a less political way than Michael Walzer. He does not presuppose that only states can be the legitimate perpetrators of violence. And he acknowledges that states can act violently in immoral ways and that for instance the bombing of German and Japanese cities in World War II or numerous actions of the IDF in Lebanon, Gaza and the occupied territories, did have a terrorist character; these actions were violations of the principle of not harming and destroying innocents (Primoratz 2007/08, p. 39). Besides that, Primoratz does *not* judge terrorism by focusing on the actor's evil identity. He argues that when we wish to answer the question of what makes the terrorist moral or immoral, we have to pay attention only to *what* is done and to what is the proximate *aim* of doing. This insight brings him to this definition of terrorism: 'The deliberate use of violence, or threat of its use, against innocent people ... with the aim of intimidating some other people into a course of action they otherwise would not take' (Primoratz 2004, p. 24; Primoratz 2007/08, p. 40). It goes without saying that the crux of the question of what makes the terrorist (im)moral will be the meaning of 'innocent people'. According to Primoratz the targeting of innocents is an essential trait of terrorism (Primoratz 2004, p. 20). Let's follow Primoratz's arguments on innocence, and let's realize that the concept of innocence refers to a characteristic that is ascribed by one group of people to another group of people.

Primoratz firstly claims that the immediate targets of terrorism have to be considered to be innocent not only from the view of those who are targeted, but potentially also from the view of the terrorists themselves (Primoratz 2007, p. 42)! In order to avoid a relativistic view Primoratz tries to give 'innocent' an objective meaning which *could be* subscribed to by the terrorists themselves.

Primoratz's second claim is that innocents are those who *on any credible understanding of responsibility and liability* for the injustice the terrorists fight against, cannot be responsible. They are innocent to the degree that makes them not liable to be killed or maimed (Primoratz 2007, p. 41).

The third and decisive step in his argument is complex. On one side Primoratz realizes that even in just war theory, non-combatants can lose their innocence, that is: can lose their immunity for violent attacks. He even suggests that if the injustice is not real, but merely alleged by the terrorists, this allegation can be enough for losing one's immunity against violence (Primoratz 2007, p. 41). But on the other side Primoratz argues that the understanding of responsibility which is expressed in the terrorist's very own justifications, like Osama Bin Laden's justification of 9/11, is simply 'preposterous' and false. Osama's basic idea – a 'Manichean' idea in which the world is divided into strictly good and strictly evil – was that all people who pay taxes to the American government and vote for their president are enemies and may be attacked. So this view is according to Primoratz preposterous and false. And he states that because of the killing of people who were really innocents, the act of 9/11 was *prima facie* simply wrongful. So Primoratz suggests on the one hand that terrorists could know – *on any credible understanding!* – that their targets were innocent people, while on the other hand he holds the opinion that the terrorists deliberately choose to act in an wrongful way by killing them. Now we may understand that Primoratz's answer to the question of what makes the suicide terrorist moral or immoral is that terrorism is almost (always) absolutely wrong (Primoratz 2007/08, p. 51). Why almost? Because Primoratz argues that the violation of the principle of innocence does not need to be *in any situation* the criterion that makes a terrorist act wrongfully. He can even suggest that terrorism as the killing of innocents can be justified whenever a state or a group of people are facing a moral disaster and terrorism is the only possibly effective method available (Primoratz 2004, p. 49). In our view, this implies that, although Primoratz is definitely right in observing that the killing of innocents is *prima facie* morally wrong, by admitting that terrorism can in principle be theoretically justified, the killing of innocents cannot be the decisive factor which makes the martyr always immoral.

We may conclude by now that a political verdict on terrorism like Walzer has brought in, does not really touch on the phenomenon of suicide terrorism, while the more philosophical approach of an author like Igor Primoratz could not indicate crystal-clear what exactly did and does make suicide terrorism immoral (or moral). A normal, political or secular-ethical approach of suicide terrorism seems not to be appropriate to uncover by interpretation the horrible nature of terrorism.

16.2.3 Marcel Hénaff

Before we're going to consider in a minute that horrible (= religious) nature of terrorism, let's turn to another way of condemning suicide terrorism, a way in which the focus is not on the criterion of not killing innocents, but on the motivations of the terrorists. Where Walzer and Primoratz both elaborated on the issue of innocence, philosopher and anthropologist Marcel Hénaff focuses on the impulses and motivations of the terrorists (of 9/11). According to him the hijackers of 9/11 were not seeking practical military and political advantages, but global vengeance by 'the production of a spectacular image', an event which should be shocking, intimidating and destabilizing the Western world (Hénaff 2007/08, pp. 73ff.). Although Hénaff realizes that Al Qaeda claimed to have political goals and that actions like 9/11 were proclaimed to be religious – the terroristic reasoning had to be a religious one, if only because terrorists like the late Osama bin Laden claimed to fight a basically metaphysical war against the forces of evil – he argues that in the last resort these kind of suicidal attacks can be considered neither political nor religious (Hénaff 2007/08, p. 93). These kind of attacks were according to Hénaff outside of the political as we understand it, because they should be 'only' randomly nihilistic, symbolic images of destruction by which the terrorists wished to punish a supposedly guilty civilization. And they should not have a religious character because the community of Islam and the 'authorized representatives' of the Quranic tradition did condemn the attacks as contrary to the teachings of the Koran. Hénaff concludes that the acts of terrorists are acts by which they remove themselves from any understanding. He summarizes his arguments by stating that we cannot give meaning to these kind of terroristic acts, and by concluding that the terroristic acts must therefore be simply criminal:

> We find ourselves bewildered because we are deprived of [any] ... frame of reference. We are confronting acts that radically remove themselves from any deliberate space and whose motivations are of a purely dogmatic and absolute character: their aim is to punish and therefore to kill, nothing more. The death of those who carry out the attack, which is part of the act, gives their act a sense of ultimacy, which amounts to saying: 'Your judgment does not concern us; we remove ourselves from humankind.' This is of course proclaimed in the name of transcendence that claims to be religious. But if this assertion is rejected by the community of believers to which the attackers claim to belong, then the only thing that remains of the act is pure criminal violence.... (Hénaff 2007/08, p. 94)

We could ask Hénaff whether it is correct to connect our bewilderment with the purely criminal character of 9/11 and other terroristic acts. And we could wonder whether Hénaff does not step into an a-historical, and meta-physical trap by considering 9/11 as an ultimately criminal, not really understandable, inhuman action. A colleague of Hénaff, anthropologist Niccolo Caldararo, has refuted views like those of Hénaff. He points out that Westerners are inclined to hold myopic views on suicide terrorism and Islam, because one actually suggests that public suicidal actions are not in the first place outside human history but specifically outside our Western history, while one 'forgets' that also in Western history suicidal actions could have a public function; we may according to Caldararo think of the (public) glorification of Christian martyrdom to waves of self-sacrificial terror attacks of anarchists in the nineteenth century, even to contemporary actions of 'suicide by cops' (Caldararo 2006). Against views like those of Hénaff, Caldararo suggests that although suicidal political actions are a relatively new phenomenon, they have to be located within an Eastern (and Western) global tradition.

16.2.4 Faisal Devji

In particular Hénaff's assumption that 9/11 is not religious because the action is not justified by a community of believers, seems not to be correct. American historian Faisal Devji pointed some years ago to the new-religious character of the Jihad. In a brilliant analysis of the 'medial' landscapes of the Jihad – that is of the patterns of beliefs in the West which are produced by the public images of suicidal terrorist acts – he argues that the novelty of 9/11 was not in the first place the barbaric taking of many innocent lives, but the manifestation of a radically new, mediated image of individualized and at the same time publicly manifested religion! (Devji 2005, p. 32). According to Devji it is the predicament of all modern, radical religion that it is individuated, and that its content and behaviour are divorced from collective solidarity with a community of believers. Devji demonstrates extensively how modern jihadis considered a violent, external jihad (war against non-Islamic enemies) to be a purely individual duty, and how the new jihadis in acting according to this duty wrested themselves away from traditional Islamic authority; while at the same time they justified their actions by drawing upon the flotsam and jetsam of received Islamic ideas and upon remembered histories spanning in the Muslim tradition (Devji 2005, p. 41).[2] (The Islamist understanding of jihad and martyrdom, observes Roxanne L. Euben, is a religious continuous and simultaneously discontinuous, while individualizing and innovative reading of an religious, Islamic path: Euben 2002, p. 21). In particular an analysis of texts of 'great' jihadis like Osama the Laden, Mohammed Atta, and Ayman al Zawahiri revealed for Devji that on the one hand the modern jihadis destroyed traditional Islamic authority, while on the other hand their actions did actually have a fully religious (public, mediated) character (Devji 2005, pp. 113ff.).

Following Devji's interpretations we may say that the actions of modern jihadis were not only religious because they were pursued by believers or because they were justified by (individually based) religious arguments. No, they had a religious character because in the end suicide terrorism has a self-justifying 'mystical' nature which was/is publicly displayed. Devji agrees with Olivier Roy that the individualization of the jihad means that the individual concerns of the terroristic advocates of personal faith, repentance and individual salvation, have transformed the suicidal terrorist practices into 'mysticism', that is to say into 'post-political-mystical' acts. In line with Roy's arguments Devji postulates that the terrorist suicidal act has to be interpreted as a self-contained and self-justifying manifestation of individualized religion. Why 'self-justifying'? Because the mediation of the jihad to the world in the media brings us the message: there is no secular ground for this act, only a 'decisionistic' religious one! (Devji 2005, pp. XVI, 42–43, 48, 52ff. 114ff.). No wonder that Al Qaeda can be seen as a religious movement which can be favourably compared with a 'mystical brotherhood'. No wonder that there is a plausible ground to consider Osama bin Laden in the first place not as a political, but as a (terrorist) spiritual leader! (Post 2007, p. 197). Let's now pay more attention to the religious nature of terrorism, and to the question whether this new, basically individualized religion, may be the factor that makes the predatory martyr (im)moral.

[2] In an interview René Girard states that with 9/11 we've witnessed a return of an archaic sacrificial (religious) mechanism (Girard and Doran 2007, p. 25). He even seems to suggest that 9/11 is the most radical manifestation of the return of (archaic) religion to a secularizing world! (Girard and Doran 2007, pp. 28, 31). Following Devji's interpretation it seems more correct to assume that 9/11 signals the coming of a politically individualized 'new religion'. Jihadism seems to be a post-political expression of a 'new religiosity' which has gone astray!

16.3 On Suicide Terrorism and Religion

Suicide terrorism is ever and again infected by religion. Even in political-secular forms of terrorism one can detect religious elements. The Tamil resistance for instance was a secular nationalist movement, but the cultural atmosphere from which the actions of the Tamil suicide bombers took place, showed religious characteristics: friends and relatives of Black Tigers were encouraged to erect shrines in remembrance of the bombers who were glorified as martyrs for the highest cause; religious themes were used by the Tamil leadership not only to generate support for the Tamil cause, but also to motivate and to justify the actions of suicide bombers; and the candidate bombers were infused with the idea that dying on the battlefield was ‘only’ the shedding of their body from their immortal souls, and that their death was a sacrificial service to the basically transcendent, fully utopian ideal of the Tamil political community (Barlow 2007, pp. 135–136; cf. also pp. 122–123) for the religious elements in the supposed purely secular-military kamikaze actions of Japanese pilots.

In particular for religiously inspired suicide terrorists it is clear, wrote Bruce Hoffman, that their violence has to do with martyrdom, for their violence ‘is first and foremost a sacramental act or divine duty executed in direct response to some theological demand or imperative’ (Hoffman 1993, p. 2; Neria et al. 2005). Although research of Mohammed M. Hafez (2006) and Ariel Merari (2010) into the personal convictions of imprisoned Palestinian ‘would- be’ suicide bombers (those who ‘failed’ and who were arrested before their martyr operation) revealed that the ideals of religious martyrdom are not the only factors in the making of a suicide terrorist (Merari 2010, p. 243) and that nationalistic motives, revenge for Palestinian suffering etc. played a role, it seems to be evident that the desire for more or less clearly formulated individual *redemption* also played a major role: suicide bombing is not only an opportunity to punish an oppressive enemy, and to fulfil God’s command to fight injustice, but above all a privilege and a reward for the individual bombers who are committed to their faith and their values (Hafez 2006, p. 44; cf. also p. 2). One may assume that the act of martyrdom is not only a religiously inspired act but is also in itself a religious act. It is an act in which those who cannot assert themselves individually, politically and culturally, can assert themselves in death (Khosrokhavar 2005, pp. 49, 229)! How can we understand that composition of religion (sacramental act), martyred self-sacrifice, death, and reward?

In the first place by realizing that there can be a connection between the ways the great mystics strove for ultimate religious rewards, i.e. for the joyful (re-)union with God, *and* absolute, ultimately self-annihilating submission. Philosopher of religion Ariel Glucklich, who in his work focuses on self-directed aggression in religion, pointed recently to the mystical impulses in the actions of suicide bombers. His general idea in his *Dying for Heaven* (2009) is that religious rituals – and a suicide bomber attack is the public performance of a ritual – can be directed at three types of religious enjoyment: celebrative, ecstatic and, the most paradoxical type: self-immersive, that is ultimately self-annihilative ‘enjoyment’ (Glucklich

2009, pp. 77ff.). According to Glucklich we may interpret the rituals of predatory martyrs as being specific to the last type. In the mystical tradition, this type of enjoyment consists in gaining the capacity to control and master one's own need for mere pleasure (of being alive). The greatest religious enjoyment is for the martyr is absolute self-control, a control that expresses itself in the end in fiery self-annihilation in the name of God. Glucklich points to the affinity between great religious figures like St. Teresa, Gandhi, and great Jewish mystics on the one hand and Iranian Basiji, Black Tigers and Palestinian human bombers on the other hand:

> Clearly, Gandhi never killed a soul, and just as clearly the (suicide) killer was not trying to strengthen anyone's moral fiber. But in their distinct ways, all of these are primarily religious actors. All choose their own personal disintegration … in order to attain something far grander than pragmatic solutions to practical problems. They seek nothing short of a beatific vision – a religious ideal that offers up permanent happiness. (Glucklich 2009, pp. 30–31)

In the second place we should reckon with what is already indicated by Glucklich and which we may call the 'secret economy' of self-sacrificial rituals (Eller 2010, pp. 108ff.). Every sacrificial ritual contains elements of violence, but the religious believer might not perceive the bleeding, or beating or killing of the sacrificial victim as repulsive violence at all. The mystery of all sacrificial rituals is that in the symbolic or real shedding of blood – of victims and of perpetrators – not the taking of life but the giving of life is really fundamental (as anthropologist E.O. James brought forward already in the 30s). This 'mystery' also goes for suicide terroristic actions like 9/11: we may interpret 9/11 as being an attempt at the giving of life: in fulfilling God's plan, 9/11 was, for the hijackers, simultaneously the pursuit of the utopian life of the *umma* which transcends the single lifetime of each one of them *and* the individual act of getting redemption (as well as the believed immortalization of their very own deeds) (Euben 2002, 10ff.).

One may, however, wonder whether it is correct to consider modern suicide bombers as martyrs, because one could suggest that 'classical' martyrs neither deliberately chose to be killed nor were oriented toward killing others. Yet we surmise that it can be correct to consider suicide terrorism as a form of 'predatory martyrdom' (and that the designation of the terrorists as martyrs does not need to be mere propaganda). For in the first place we have to realize that in their anticipation of the glory of individual or collective resurrection 'classical' martyrs could also have *willingly and actively* embraced death, even, as Hugh Barlow recently made clear, to the point of provoking their death (which had – Glucklich! – to be considered as the consummation of the love for God; Barlow 2007, p. 42, and passim; cf. Middleton 2011, pp. 33, 44, 49ff). In the second place we'll have to realize that in particular in Islamic traditions, but also in Christianity, the notion of martyrdom can be amalgamated with the figure of the combatant, not only with the victim: the combination of glorified martyrdom and being a warrior is part and parcel of the monotheistic religious traditions. In the third place although one may suggest that the suicide terrorists transformed martyrdom from commemoration of violence suffered to justification of violence, we'll have to endorse the idea that this transformed martyrdom in its individualized and at the same time (post)political Gestalt shows up a continuity with the traditional components of martyrdom. These components are: the pressing presence of a highest, spiritual and (post-political) cause or truth – the life of the utopian *umma*; further a threat or challenge to that cause or truth – occupation and repression by the great enemy; there is moreover a person willing to give his life for the cause – the literarily ab-solute believer, claiming to be radically religiously submissive; and there is of course the audience! (Eller 2010, pp. 150f.). Regarding this last component of martyrdom, let us not forget that audiences were always essential for sacrificial rituals, and that the system of sacrifice and the system of war, as Grace Jantzen has made clear, were never only pure religious acts but had a function for the community of those who 'witnessed' the 'witnessing operation' of the martyr/warrior (Jantzen 2009, pp. 132–133).

So we may surmise that suicide terrorists, like most of the Palestinian bombers and the hijackers and organizers of 9/11, actually and not *per se* consciously, built, in an individualized, innovative way, on an old religious tradition: a tradition in which 'the system of sacrifice' and 'the system of war' were united in a way that the martyr could commit acts of extreme aggression against his own body and against the bodies of his enemies (Glucklich 2009, p. 134).

The ultimate question is of course: but how could he (or she) do that and transgress 'normal' moral boundaries by killing non-combatants? The (part of an) answer has to be: because he (she) acted out of a religious atmosphere in which a cruel, capricious God is considered to be the one and only foundation of morality. The terrorists actually selected from their religious tradition the idea that the great models of faith, 'the knights of faith', had to be figures like Abraham and Isaac/the silent Ishmael – Mohammed Atta, the leader of the 9/11 hijackers considered himself being an Ishmael! They were models of men who in complete obedient submission to God could make killing of a human being and being killed into a God-pleasing act! We could state that suicide terrorists – 'altruists'! – pursued with their 'operations' the mechanism which once was designated by Kierkegaard in his interpretation of Genesis 22 in *Fear and Trembling* as the 'teleological suspension of ethics'. The foundational idea in this suspension of ethics is that the only source of morality has to be God, and that an act is not good because human beings consider it good, but that an act is good because it is God's will. It is the purely religious suspension of ethics, this radical attempt to fulfil God's morality which is beyond human morality, that could make Abraham and Isaac/Ishmael, *and* the suicide terrorist, perpetrate and accept something which would otherwise evoke moral revulsion (Jantzen 2009, p. 125).

So it is no surprise that one cannot connect directly a coherent relation between suicide terrorism and morality. And it is no surprise that it is in the last resort too simple to state that suicide terrorism is evil. As being an ultimately religiously inspired act, even a religious act in itself, we have to consider 'the autonomous nature of the religious impulse' (Lawrence 1990), that is to say the possibility that the religious act is carried out in quite a different grammar – according to the terrorists a much higher grammar in which all human morality has to be framed – than the non-religious, moral act (Philips 1964, p. 412). The terrorists of 9/11, suggested Malise Ruthven, were virtually claiming autonomy for their acts: they were so close to God that they identified their action with the will of God, and in doing so left to God the moral consequences of their act (Ruthven 2002, p. 35).

16.4 Conclusion

So, do we have to conclude now that, just because of the religious autonomy of the terrorist act, we may not actually assess what the martyr should make moral or immoral, because one can claim that in the end the terroristic-religious act distracts and suspends itself from an ethical judgment?

I surmise we still can. But we should search deeper into the religious nature of terrorism and delve into the religious life world of terrorists and terrorists organizations. And in our approach to terrorism we should proceed from political and moral philosophy to a way of thinking normatively and theologically! So maybe what was true of some ‘classical’ Christian martyrs and religious self-renouncers, is also true of ‘predatory martyrs’: that they possess a sense of uniqueness and destiny because they feel obliged to take up the burdens of mankind; that their determination to sacrifice themselves is accompanied by an equal willingness to sacrifice others, and that, just because they fully identify with God’s will, their act is ultimately not God’s end, but ‘only’ the highest manifestation of their very own will: the martyr’s absolute submission and sacrificial renouncement of the self could be an act to dispose of one’s life and to control other men’s life in the manner of an absolute monarch (Smith 1997, p. 15; cf. Eller 2010, pp. 158–159). And it could be that cultural theorist Terry Eagleton is right when he acknowledges on the one hand the fully religious character of the act of the martyr:

> In destroying their own flesh and blood . . . suicide bombers bear witness to a power which is in their view even more formidable than the state. The flesh may be corruptible, but in the very process of its dissolution the Idea which spurs them on stands forth radiant, sublime, unkillable. (Eagleton 2005, pp. 95–96)

While also acknowledging on the other hand that:

> What cannot be annihilated is the very will which drives them to annihilate themselves. To be able to will yourself out of existence is to be godlike indeed. It is both to imitate God and to oust him, since as a creator in your own right you have usurped his divine prerogative. *Suicide is the death of God. He who is ready to kill himself becomes a God*, declares Kirillow in Dostoevsky’s *The Devils*. What more breath-taking form of omnipotence than to do away with yourself [and a number of others, AvH] for all eternity? (Eagleton 2005, pp. 96–97)

When we realize, with ‘dialectical’ theologians like Karl Barth, that the fiercest criticism of religion is often delivered in religion itself by religious thinking and doing, and that, as Eagleton points out, religion can cause the death of God (!), we could now surmise that the question of what makes the martyr immoral is in the first place answered, not by moral-political reasoning and by pointing at the killing of human beings, but by ‘theologically’ pointing to the killing of God by religious martyrs. The most plausible answer to our central question ‘What makes the martyr (im)moral’ could become: It is their own individuated, decisionistic version of ‘bad religion’ (Devji) that makes the martyr immoral.

Acknowledgments Thanks to Dominic Cronin and Floor van Harskamp

References

Alexander, J.C. 2004. From the depths of despair: Performance, counterperformance, and “September 11”. *Sociological Theory* 22(1): 88–105.

Asad, T. 2007. *On suicide bombing*. New York: Columbia University Press.

Atran, S. 2006. The moral logic and growth of suicide terrorism. *The Washington Quarterly* 29(2): 127–147.

Barlow, H. 2007. *Dead for good: Martyrdom and the rise of the suicide bomber*. Boulder: Paradigm Publishers.

Bloom, M. 2007. *Dying to kill: The allure of suicide terror*. New York: Columbia University Press.

Caldararo, N. 2006. Suicide bombers, terror, history, and religion. *Anthropological Quarterly* 79(1): 123–131.

Devji, F. 2005. *Landscapes of the Jihad: Militancy, morality, modernity*. Ithaca: Cornell University Press.

Eagleton, T. 2005. *Holy terror*. Oxford/New York: Oxford University Press.

Eller, J.D. 2010. *Cruel creeds, virtuous violence*. Amherst: Prometheus Books.

Euben, R.L. 2002. Killing (for) politics: Jihad, martyrdom, and political action. *Political Theory* 30(1): 4–35.

Girard, R., and R. Doran. 2007. Apocalyptic thinking after 9/11: An interview with René Girard. *SubStance* 37(1): 20–32.

Glucklich, A. 2009. *Dying for heaven: Holy pleasure and suicide bombers – Why the best qualities of religion are also its most dangerous*. New York: Harper Collins.

Hafez, M.M. 2006. *Manufacturing human bombs: The making of Palestinian suicide bombers*. Washington, DC: Unites States Institute of Peace.

Hénaff, M. 2007/08. Global terror, global vengeance? *SubStance* 37(1): 72–97.

Hoffman, B. 1993. *Holy terror: The implications of terrorism motivated by a religious imperative*. Santa Monica: RAND.

Hoffman, B. 2006. *Inside terrorism*. New York: Columbia University Press.

Jantzen G.M. 2009. *Violence to eternity: Death and the displacement of beauty*, ed. Jeremy Carrette and Morny Joy. London/New York: Routledge.

Jenkins, B.M. 1980. *The study of terrorism: Definitional problems*. Santa Monica: The Rand Paper Series.

Juergensmeyer, M. 2003. *Terror in the mind of God: The global rise of religious violence*. Berkeley: University of California Press.

Khosrokhavar, F. 2005. *Suicide bombers: Allah’s new martyrs*. London/Ann Arbor: Pluto Press.

Lawrence, B. 1990. *Defenders of God: The fundamentalist revolt against the modern age*. San Francisco: Harper & Row.

Merari, A. 2010. *Driven to death: Psychological and social aspects of suicide terrorism*. Oxford/New York: Oxford University Press.

Middleton, P. 2011. *Martyrdom: A guide for the perplexed*. London/New York: T&T Clark International.

Morton, A. 2005. *On evil*. New York: Routledge.

Nanninga, P. 2007. Representing fundamentalism: Academic discourses on Islam and violence: The impact of 9/11. Thesis Research Master Religious Symbols and Traditions, University of Groningen, Groningen.

Neria, Y., et al. 2005. The Al Qaeda 9/11 instructions: A study in the construction of religious martyrdom. *Religion* 35: 1–11.

Neumann, P.R. 2009. *Old and new terrorism: Late modernity, globalization and the transformation of political violence*. Cambridge: Polity.

Philips, D.Z. 1964. Moral and religious conceptions of duty: An analysis. *Mind*, New Series 73(291): 406–412.

Post, J.M. 2007. *The mind of the terrorist: The psychology of terrorism from the IRA to Al-Qaeda*. New York: Palgrave-Macmillan.

Primoratz, I. 2004. What is terrorism? In *Terrorism: The philosophical issues*, ed. I. Primoratz. Houndmills: Palgrave Macmillan.

Primoratz, I. 2007/08. A philosopher looks at contemporary terrorism. *Cardozo Law Review* 29: 33–51.

Ruthven, M. 2002. *A fury for God: The Islamist attack on America*. London: Granta Books.

Schinkel, W. 2010. *Aspects of violence: A critical theory*. New York: Palgrave-Macmillan.

Schmid, A. (ed.). 2011. *The Routledge handbook of terrorism research*. London: Routledge.

Schwartz, J.M. 2004. Misreading Islamist terrorism: The "War Against Terrorism" and just-war theory. *Metaphilosophy* 35(3): 273–302.

Smith, L.B. 1997. *Fools, martyrs and traitors: The story of martyrdom in the Western world*. New York: Knopf.

Steinhoff, U. 2004. How can terrorism be justified? In *Terrorism: The philosophical issues*, ed. I. Primoratz, 97–109. Houndmills: Palgrave Macmillan.

Tilly, C. 2004. Terror, terrorism, terrorists. *Sociological Theory* 22(1): 5–13.

Tosini, D. 2007. Sociology of terrorism and counterterrorism: A social science understanding of terrorist threat. *Sociology Compass* 1(2): 664–681.

Vallis, R., et al. 2006. *Disciplinary approaches to terrorism: A survey*. Newport: University of South Wales.

Valls, A. 2000. Can terrorism be justified? In *Ethics in international affairs*, ed. A. Valls, 65–80. Lanham: Rowman & Littlefield.

Walzer, M. 2000. *Just and unjust wars: A moral argument with historical illustrations*. New York: Basic Books.

Walzer, M. 2004. *Arguing about war*. New Haven: Yale University Press.

Wieviorka, M. 2009. *Violence: A new approach*. London: Sage.

Zimbardo, P. 2007. *The Lucifer effect: Understanding how good people turn evil*. New York: Random House.

Chapter 17
Moral Lessons from Monstrosity: *The Kindly Ones* and the Reader

Bettine Siertsema

Seeking an answer to the question 'What makes us moral?' I follow philosophers like Martha Nussbaum and Rüdiger Safranski in their taking literature as the territory where questions like this are posed (Nussbaum 1990; Safranski 1998). I concentrate on a controversial novel in which the main character, who also is the narrator, seems to represent a total lack of morality: Dr Max Aue in Jonathan Littell's novel *The kindly ones*, published in French in 2006, and translated into English in 2010. Aue insists that he is not fundamentally different from the reader or any other ordinary person. I will examine his reasoning and compare it with the arguments found in interviews with real-life SS-men, with the thoughts of Hannah Arendt on the theme, and with the analysis by sociologist Stanley Cohen of the ways in which responsibility for certain evil events is denied. This helps to see where and in what way perpetrator reasoning like Aue's goes astray. The flawed reasoning is, indeed, all too recognizable. Aue presents himself as 'Everyman'. He seems to be the personification of Hannah Arendt's 'banality of evil'. And indeed, 'kindly' readers may be willing to accept much of his reasoning and arguments, but only up to a certain point, I assume – a point where his actions and cool attitude towards them stretch the reader's ability and willingness to empathize too much. However, I think that there is still another reason why it is doubtful that the claim of 'ordinariness' that author Jonathan Littell has Aue insist on will hold. This reason does not regard Aue's thoughts and justifications, nor his actions, but the peculiarity of his not war-related life, with which the author has adorned him.

B. Siertsema (✉)
Faculty of Philosophy, VU University, Amsterdam, The Netherlands
e-mail: g.siertsema@vu.nl

B. Musschenga and A. van Harskamp (eds.), *What Makes Us Moral? On the capacities and conditions for being moral*, Library of Ethics and Applied Philosophy 31,
DOI 10.1007/978-94-007-6343-2_17, © Springer Science+Business Media Dordrecht 2013

17.1 Theoretical Background

17.1.1 Historiography

Prior to writing his novel Jonathan Littell spent 7 years on historical research. He doesn't mention his sources anywhere (since it is a novel, indeed he wasn't obliged to), but one can make an educated guess as to some of the books he used for background. In this section I briefly go into some works that he may have consulted, as far as they are of interest for the theme of the morality of the main character.

In 1992 American historian Christopher Browning published his study of the members of a reserve police battalion involved in the mass executions of Jews in the Lublin district in Poland, under the revealing title *Ordinary Men*. One of his conclusions was that the men who were assigned to shoot defenceless civilians were given a free choice in advance and did not suffer any punishment or disadvantageous repercussions if they refused. Another, no less shocking conclusion was that men who in the beginning felt disgust and shame about these shootings, to the point of fainting or vomiting, after some time were numbed, while some even began to enjoy their tasks (Browning 1992).[1]

As a kind of answer to Browning's book, American political scientist Daniel Goldhagen published *Hitler's willing executioners* (Goldhagen 1996). It focused on the same police battalion 101, and on two other instances of direct confrontations of 'ordinary' Germans with the suffering of Jewish prisoners: the labour squads from the concentration camps and the death marches. Goldhagen's thesis is that the Holocaust was caused by a virulent, 'eliminationist' anti-Semitism, that is typical for the German people since the Middle Ages. The book raised an intense debate.[2] It met with great acclaim by the general public, especially in the US, but historians heavily criticized it.

17.1.2 Social Sciences

Goldhagen's conclusions were not only contested by historians but also by social scientists. Leonard S. Newman, for instance, showed how Goldhagen's argument is

[1] In 2012 Browning returned to the theme of *Ordinary Men*. In his 'Auschwitz: Never again' lecture he took social-psychological research results into consideration, including works on more recent mass murder. He comes to the 'dire conclusion that modern governments that wish to commit mass murder will seldom fail in their efforts ... to induce "ordinary men" to become their "willing executioners" (Browning 2012, p. 23).

[2] In 1996 and 1997 at least five issues of journals (*Commentaire, Le Débat, Les Temps Modernes*, *Documents: Revue des questions allemandes* and *Psyche: Zeitschrift für Psychoanalyse und ihre Anwendungen*) and six collections of essays and reviews (edited by respectively: Edouard Husson, Franklin Littell, Sabine Marquardt, Julius Schoeps, Wolfgang Wippermann and Rolf Binner et al.) were dedicated to Goldhagen's book (Binner et al. 1998, pp. 260–263).

refuted by several well-known earlier social-psychological studies and experiments, of which he most famous is the Stanley Milgram's 'Behavioral Study of Obedience' in 1963 (Newman 2002). Newman concludes from the various experiments that 'A monocausal framework for understanding genocide [such as Goldhagen's concept of German eliminationist anti-Semitism, BS] cannot do justice to the dynamic interplay between persons and situations that is characteristic of human behaviour.' All of the experiments he describes, by the way, have to do with the question 'what makes us moral?' and show how easily an individual refrains from practicing the morale he/she theoretically agrees with and how easily his/her judgment on him/herself and others is manipulated by group dynamics or higher authorities. That doesn't mean however, that persons can appeal to these social-psychological findings as an excuse for their individual acts: evidently the fact that most people in a given situation react in a certain way doesn't relieve someone from his/her individual responsibility. In another context this point is stressed by sociologist Stanley Cohen. In his study *States of Denial* he analyses the strategies of denial and justification that both governments and individuals employ to absolve themselves for their wrongdoings or lack of intervention. In view of the distance to the subject I don't suppose that Littell used Cohen's book, but I found it helpful for the section 'Justifications' in this paper (Cohen 2004).

17.1.3 Philosophy: Hannah Arendt

When dealing with the Holocaust perpetrator philosophically one cannot ignore Hannah Arendt's book on Adolf Eichmann (Arendt 1963). Eichmann was one of the first Holocaust perpetrators to receive wide attention as a person. Since his trial in 1961 the Holocaust became the centre of attention where the Second World War was concerned.[3] Not only was the trial an international television event, it also was elaborately covered by numerous newspapers that sent writers of high repute as reporters to Jerusalem. In her report for *The New Yorker* Hannah Arendt introduced the term 'the banality of evil', referring with 'evil' not to the crime of the Holocaust but to the person who committed (or helped to commit) that crime by unthinkingly obeying orders. Apart from the term 'banality of evil', it is Arendt's belief in Eichmann's lack of anti-Semitism and his obedience as the heart of his motives in facilitating (or taking part in) the Holocaust that has since been contested. Several authors have cast doubt on or even disproven the truthfulness of this characterization (Sharpe 1999; Browning 2003; Cesarani 2004; Lipstadt 2011).

[3] The 1945–1946 Nuremberg Trials of high-ranking Nazi officials concentrated more on the procedures and the documents that would establish their guilt than on their characters and moralities. Moreover, the persecution and murder of the Jews was viewed merely as one of the many war crimes to be judged (Wieviorka 2006).

Much of Arendt's views, contested as they may be, also apply to the fictional Max Aue. This goes even more for her considerations in the 1964 essay 'Personal responsibility under dictatorship', an answer to the criticism on her book (Arendt 2004). Here Arendt argues that in a totalitarian state only those who give up their roles in public life are able not to partake in unjust acts. An appeal to the required obedience to orders or the law doesn't hold, because obedience is a form of support. She states that most of the mass murderers and their accomplices did not believe in the ideological justification of those murders. For them, the fact that it happened according to the will of the *Führer* was sufficient reason. She refuses to apply the concept of collective guilt on the German people as a whole, because it would mean ignoring those persons who did resist or protest.

17.1.4 Biographical and Visual Input

I found parallels with the fictional Aue in other portraits of major Holocaust perpetrators: the autobiography of Rudolf Höss, the commandant of Auschwitz (Höss 1959), and the interview-based biography of Franz Stangl, the commandant of Treblinka, by Gitta Sereny (1974). An insiders' view of Holocaust perpetrators of lower ranks is given in a book by Ernst Klee et al. that presents accounts and photographs of the Holocaust, which soldiers sent home in letters to their families (Klee et al. 1988). The photo album of the Ukrainian *Aktionen* Aue makes and sends to his superiors in Berlin (which is to his credit and advances his career), may well be inspired by Klee's presentation, and also by the German exposition in the late 1990s, 'Vernichtungskrieg. Verbrechen der Wehrmacht 1941–1944'.[4]

17.2 Just Like You?

17.2.1 From Bystander to Perpetrator

Returning to the novel, let us first make Aue's acquaintance: he is an SS-officer, holding a doctorate in law, who tells in great detail (the novel numbers a little under 1,000 pages) his experiences in the years 1941–1945, in the Ukraine, Stalingrad, Hungary, Auschwitz and Berlin. These include the massacre at Babi Yar and other *Sonderaktionen*, the starvation and defeat of the German army at Stalingrad, where he gets seriously wounded, the deportation of the Hungarian Jews, the death marches from concentration camps as the Red Army draws near, and the allied bomber

[4] See for the background of this exposition and the controversy about it: http://en.wikipedia.org/wiki/War_crimes_of_the_Wehrmacht#Wehrmachtsausstellung (last consulted 5 March 2012).

attacks on Berlin. The tasks he is successively assigned to is to write reports on the progress of the *Endlösung*, on the morale of the Wehrmacht, and on the living conditions in the camps. Therefore his function makes him not so much a perpetrator as a sort of official 'bystander' (to use the phrase coined by Raul Hilberg), but he is a bystander from very close by. And, not wanting to distance himself from the dirty work, Aue doesn't shy away from personally shooting civilians, including women and children, on certain occasions. He survives the war and settles under another name in France, where he writes his memoirs some 40 years later. He tells about the atrocities he witnessed and committed in a businesslike, detached manner, which leaves the reader disturbed, appalled even.

The first part, titled 'Toccata' (all the parts have a title referring to a piece of music, the parts of a suite or a partita) gives the most explicit moral lesson of the book. Indeed, in the first sentences it characterizes itself as 'a morality play', but also as 'a search for truth'. The text begins with an address to the reader, perhaps alluding to Beethoven's Ninth Symphony: 'Oh my human brothers, let me tell you how it happened.' Aue then denies being an evil man. He didn't ask for the tasks he executed, but he doesn't plead the infamous excuse *'Befehl ist Befehl'* either: 'What I did, I did with my eyes open, believing that it was my duty and that it had to be done, disagreeable or unpleasant as it may have been.' He admits being guilty, and then explicitly addresses the readers once more: 'I am guilty, you are not, fine. But you should be able to admit to yourselves that you might also have done what I did.' After all, 'every one, or nearly every one, in a given set of circumstances, does what he is told to do; and, pardon me, but there's not much chance that you're the exception, any more than I was' (p. 20).[5] It is a proposition that is indeed sustained by socio-psychological experiments like Stanley Milgram's and Zimbardo's (Milgram 1974; Zimbardo 2007). Aue then declares: 'The real danger for mankind is me, is you' (21). And as a conclusion of the first part, he once again states emphatically: 'I am a man like other men, I am a man like you. I tell you I am just like you!' (p. 24).

I don't know if there are many unsuspecting readers who haven't read any reviews or the blurbs before reading the novel, but if any, they will soon be cured from their neutral attitude. In the next part, 'Allemande', Aue is sent by the *Sicherheitsdienst* on a mission in the Ukraine, to report on the behaviour and the morale of the *Wehrmacht* during the liquidation of partisans, Jews and other enemies of the Third Reich.[6] He witnesses many horrendous scenes, and sometimes partakes in them as well. Yet there are several aspects that earn him some goodwill on the part of the reader. For one, he always tries to avoid unnecessary cruelty. There is

[5] This proposition is indeed sustained by socio-psychological experiments like Stanley Milgram's (1974).

[6] Recently the mass killings in East Europe, such as the infamous slaughter at Babi Yar that appears in Littell's novel, became known as the 'Holocaust by bullets', after the title of a book by the French priest Patrick Desbois. But, as this book was first published in 2008, it can hardly have been a source for Littell.

an incident when he reprimands a fellow officer who, in an unprovoked rage, had beaten an old man to death with a spade. In retribution, this ill-disposed officer plots Aue's reassignment to the starving army at Stalingrad. Furthermore, the few officers Aue takes a liking to are not die-hard National Socialists, but men who seem to look at the Nazi ideology with an ironic, rather superior, and sometimes even explicitly critical eye. When he is in the Caucasus there is a debate whether the so-called *Bergjuden* are Jewish or over the centuries have become so much mixed with other Caucasian peoples that genetically they are no longer Jewish at all. Aue clearly holds this last view, and takes it as a real setback when his superiors decide otherwise. Finally, in 1944, when he is asked to raise the output of the forced labour in the war industry, he makes efforts to better the working and living conditions of concentration camp inmates, fighting against indifferent, unwilling, incompetent camp commanders, and the general shortage of everything (food, clothing, building materials).

17.2.2 Standing by His Conviction

The reader could easily be taken in by these actions and attitudes, but not by the reasoning he uses in discussions to justify the murder of unarmed citizens including women and children. His attempt to protect the *Bergjuden* by stressing the arguments for them not being Jewish, is motivated by military interests: the *Bergjuden* are very anti-communist and could therefore be valuable informers, and they are very well integrated with the local population, which means that their persecution may become a source of animosity and social unrest in the region. His efforts (apart from his official orders) to improve the living conditions of concentration camp prisoners are prompted by his conviction that the war can only be won if the war industry would be much more productive. He may not be a raving anti-Semite, but he does not ever object to the reasoning underlying the *Endlösung*, and seems to have quite easily picked up and internalized the National Socialist racial theory, that looks upon Jews as enemies of the state, regardless of their political views, age, gender and social-economic status. That he stands by this conviction is clearly illustrated by some incidents with individual Jews, who he doesn't hesitate to kill (or let them be killed by subordinates), such as a very old, very calm and dignified *Bergjude*, and a little girl during some *Aktion* in the Ukraine:

> At the edge of a grave, a little girl about four years old came up and quietly took my hand. I tried to free myself, but she kept gripping it. In front of us, they were shooting the Jews. '*Gdye mama?'* I asked the girl in Ukrainian. She pointed toward the trench. I caressed her hair. We stayed that way for several minutes. I was dizzy, I wanted to cry. 'Come with me', I said to her in German, 'don't be afraid, come.' I headed for the entrance of the pit; she stayed in place, holding me by the hand, then followed me. I picked her up and held her out to a Waffen-SS: 'Be gentle with her', I said to him stupidly. (p. 109)

Aue's emotional disturbance (dizziness, wanting to cry) points to a feeling of personal involvement, of empathy with the little girl, and, I think, to the moral intuition that killing her might be wrong. Nevertheless there seems to be no hesitation on his part to let her be killed.

Just as repulsive (at least for me) is the episode with the young Jewish orphan Yakov, around the same time. This boy plays the piano beautifully and is adopted as a kind of mascot by Aue's military unit. Aue even orders sheet music from Berlin for him. But, while repairing the car of one of the officers, Yakov severely hurts his hand, which has to be amputated. And consequently Yakov, no longer able to play the piano, is killed. The incident itself takes place off the picture, but is no less disturbing for it. Here, not the killing of an amorphous crowd of anonymous, dehumanized people is at issue, but the killing of an individual, who is spoken with, looked into the eye, appreciated and even liked, in short someone with whom a personal relationship has been established. This relationship turns out to carry almost no weight at all for Aue and his colleagues in view of the ultimate goal of the *Endlösung*. *Almost* no weight, because afterwards one of the officers, Bohr, attacks a fellow officer who makes light of the murder (with someone else commenting on his consequent arrest: 'But it's too bad for Bohr: a good officer, and he's ruined his career for a little Jew. It's not as if there were a lack of Jews, over here'), and Aue himself deeply deplores the murder as well and in the following years thinks back of the talented boy on several occasions. In this, Yakov is different from the little girl and from a teenager whom Aue gives a cigarette before he is killed: they never seem to haunt him afterwards.

17.3 Unease

17.3.1 Guilt

Aue's self-portrait is double-sided. In the episodes where he is directly confronted with the Holocaust, he suffers from severe digestive disorders (retching and throwing up, sometimes even before the meal is finished) and bad dreams about his own or other people's excrements. In some episodes he also seems to suffer from a Lady Macbeth syndrome, with an obsessive need to clean himself and his uniform (pp. 109, 178). On the text level there are words and statements that point both to an insight into his guilt or at least into the irrationality and evil of the persecution of Jews and other groups, but also to his 'innocence', his not being responsible (apart from the many pronouncements in which he shows himself to be aware of his guilt in the eyes of the world). The following quotations can be seen in this light:

> I emerged from the war an empty shell, left with nothing but bitterness and a great shame, like sand crunching in your teeth. (p. 12)

This statement seems to implicate that the war has destroyed his inner self. The bitterness indicates his view on the outward world as the cause of this destruction; the shame indicates his understanding of his own part in it.

> But the ordinary men that make up the State [as opposed to psychopaths, BS] – especially in unstable times – now there's the real danger. The real danger for mankind is me, is you. (p. 21)

Clearly these words pronounce his guilt: he is a danger for mankind. However, at the same time he diminishes his guilt by saying that he is not special, he is just an ordinary man, and that other ordinary people have acted and will act in the same way as he has done. The suggestion is that when everybody is (potentially) guilty, no one is.[7]

The next fragment is about a plan to request the Jews to report for forced labour, and then to kill everyone who shows up:

> That night, I stayed awake for a long time: I was thinking of the Jews who would be coming the next day. I thought the method adopted very unfair; the Jews of good will would be punished, the ones who might have come to trust the word of the German Reich. (p. 42)

Aue thinks the method 'unfair', morally bad, but he does not question the goal of killing Jews itself. In a sense this echoes the scene with the little Ukrainian girl cited above: he asks the soldier to be gentle with her, as if he deplores the method, but he doesn't call the justness of killing her into question.

> I knew that these decisions were made at a much higher level than our own; still, we weren't automatons, it was important not just to obey orders, but to adhere to them; yet I was having doubts, and that troubled me. (p. 43)

Although Aue first expresses the well-known excuse that the orders came from higher up, he then states that more than sheer obedience was asked of him. He had to approve. He is troubled not by this supposed approval, nor by the content of the orders, but by his very doubts (while those doubts would probably make him more acceptable, or at least more understandable, in the eyes of his readers).

17.3.2 Doubt

> I can in all honesty say that I had doubts about our methods: I had trouble grasping their logic. I had talked with Jewish prisoners; they told me that for them, bad things had always come from the East, and good ones from the West; in 1918 they had welcomed our troops as liberators, saviours; those troops had behaved very humanely .... Now, we were killing them. And undeniably, we were killing a lot of people. That seemed atrocious to me, even if it was inevitable and necessary. (pp. 80–81)

[7] Considering the support of the German people for the Nazis, and the insistence of the Nazis that their ideology was held by the whole of Germany, Hannah Arendt concluded in 1964 that this is a misrepresentation and a false excuse: 'If everyone bears guilt, no one is guilty' (Arendt 2004:54). This is a re-phrasing of an earlier statement: 'Where all are guilty, nobody in the last analysis can be judged', put forward in an essay in 1945, 'Organized guilt and universal responsibility' (Arendt 1994, p. 126).

Again Aue expresses his doubts, but now they are more ambivalent. At first he doubts only the methods, and then the doubts seem to shift towards the goal itself: the Jews had looked upon the Germans as friends and liberators, why then should they be killed? His equivocal position is most clearly visible in the last sentence: he considers this mass killing at the same time atrocious and necessary.

Less ambiguous is his thought after witnessing an *Aktion.* Aue visits the now empty *shtetl* that testifies to the intense poverty of its inhabitants:

> Certainly nothing of this had much to do with the *internationales Finanzjudentum.* (p. 108)

Here Aue explicitly denies the *Führer*'s justification of the persecution of the Jews by using his very words in speeches like the one to the *Reichstag*, January 30th 1939. He sees no relation whatsoever between the very poor Ukrainian Jews and the 'international finance Jewry', and thus – implicitly – no reason to have them killed. He sees through this pretext, as well as through the one about Jewry as the hidden force behind communism, as the next fragment shows. Hearing a discussion among his colleagues about the strange fact that the communists do so little to save the Jews, whereas the Jews are said to dominate the Communist Party, he reacts as follows:

> I smiled to myself, but bitterly: as in the Middle Ages, we were reasoning with syllogisms that proved each other. And these proofs led us down the path of no return. (p. 122)

Eventually he comes to the conclusion that the killing of Jews is completely – in his retort in the next fragment he himself ironically understates it as 'somewhat' – arbitrary. When Orpos, police under the command of the SS, fires at random children in the street, an officer protests and declares they should be court-martialed, Aue answers:

> 'That's going to be difficult ... we've been making them shoot children for months; it would be hard to punish them for the same thing.'– 'It's not the same thing! The children we execute are condemned! These were innocent children.'– 'If you allow me, Standartenführer, the basis on which the condemnations are decided makes such a distinction somewhat arbitrary. (p. 184)

17.3.3 Unwanted Habituation

The next fragment concerns the psychology of the main character as he sees it himself. He is appalled by the horror of the mass executions, but recognizes in himself the tendency of getting used to it, and tries to recover his initial feelings of dismay by making himself watch even more killings, not surprisingly to no avail. Christopher Browning (1992) has shown how inevitably this mechanism of habituation works. Yet terms like 'insurmountable', 'monstrous violation', 'scandal', 'rupture' and 'infinite disturbance' signal the depth of his (initial) emotion.

> By inflicting this piteous spectacle [of mass executions, BS] on myself, I felt, I wasn't trying to exhaust the scandal of it, the insurmountable feeling of a transgression, of a monstrous

> violation of the Good and the Beautiful, but rather this feeling of scandal came to wear out all by itself, one got used to it, and in the long run stopped feeling much; thus what I was trying, desperately but in vain, to regain was actually that initial shock, that sensation of a rupture, an infinite disturbance of my whole being. (pp. 178–179)

17.3.4 Revulsion

The last two fragments that I cite are from further on in the novel. They show that in spite of the observed habituation, the experience of the mass killings in the Ukraine – again not surprisingly – keeps popping up from time to time, bothering him as a nagging conscience, to the point of strong physical reaction. When reprimanded by a superior upon asking how many Jews in the whole of Europe are already killed by 1943, he muses:

> Why had I asked him that idiotic, useless question? How did that concern me? It had been nothing but morbid curiosity, and I regretted it. I wanted to take an interest in nothing but positive things now: National Socialism still had a lot to build; that's where I wanted to direct my energies. But the Jews, *unser Unglück*, kept pursuing me like a bad dream in early morning, stuck in the back of my head. (p. 464)

Observing the well-educated, middle-class Jews from Hungary being deported, he is reminded of the fate of the Ukrainian Jews:

> All that made the scene even more oppressive, despite their yellow stars, they could have been German or at least Czech villagers, and it gave me sinister thoughts. I imagined those neat, tidy boys or those young women with their discreet charm being gassed – thoughts that turned my stomach, but there was nothing to be done, I looked at the pregnant women and imagined them in the gas chambers ... memories of the Ukraine flowed in, and for the first time in a long time I wanted to vomit, vomit my powerlessness, my sadness, my useless life. (p. 789)

It is clear that those feelings of guilt, or at least unease about the Holocaust (I left out episodes concerning partisans and the killing of mentally or physically ill people), appear throughout the book, though definitely more so in the earlier than in the later parts. This is partly due to the process of habituation, but more to Aue's outward adventures, such as the defeat of the German army at Stalingrad. It is not until his assignment concerning the Jews in Hungary and his vain effort to increase the labour output of prisoners, that he is confronted again with the *Endlösung*.

17.4 Justifications

Instances of the justification of the Holocaust and his own role in it, which started in the first chapter, 'Toccata', are found more in the second half of the book. I'll cite them extensively, and then try to lay bare the strategies of denial and justification as analysed in general by sociologist Stanley Cohen (2004).

17.4.1 Denying the Victim

A quotation from the earlier part of the book:

> Political philosophers have often pointed out that in wartime, the citizen, the male citizen at least, loses one of his most basic rights, his right to life .... But these same philosophers have rarely noted that the citizen in question simultaneously loses another right, one just as basic and perhaps even more vital for his conception of himself as a civilized human being: the right not to kill. No one asks for your opinion. In most cases the man standing above the mass grave no more asked to be there than the one lying, dead or dying, at the bottom of the pit. You might object that killing another soldier in combat is not the same thing as killing an unarmed civilian; the laws of war allow the one but not the other; as does common morality. A good argument in theory, but one that takes no account of the conditions of the conflict in question. The entirely arbitrary distinction ... (pp. 17–18).

Aue considers the distinction between soldiers and civilians entirely artificial. He doesn't substantiate why the laws of war and common morality would be wrong, why there would not be a difference between soldier and civilian. Much later in the novel his indignation at the civilian casualties of the allied air raids on Cologne shows his inconsistency on this point. The bombing of cities in the Second World War – on both sides: Coventry, Rotterdam, and besides Cologne also Dresden, Hiroshima and Nagasaki – makes plain that as far as victims are concerned the distinction between soldier and civilian is indeed a feeble one (although one could with good reason argue that these bombings were war crimes). Aue however denies the difference for the wrongdoers just as much: to his mind in wartime both soldiers and civilians lose their 'right not to kill'. Unfortunately he refrains from putting forward any arguments for this daring viewpoint. What is worse, he completely ignores the fact that 'the man standing above the mass grave' eventually had a choice, to shoot or not to shoot, while 'the man on the bottom of the pit' had no choice at all.

For Stanley Cohen this may be a mixture of denial of responsibility (because of the required obedience, and the alleged necessity) and a special form of denial of the victim, in this case a reversal or, more accurately, a sort of exchangeability of victim and perpetrator (2004, pp. 89–96).

After accusing the French and British of harsh measures in their colonies that impinged on human rights, thus 'proving' that the moral sense of the Western powers did not fundamentally differ from the German one, but also predicting that after the Jews, the gypsies and the disabled, it would have been the turn of 30–51 million Russians and Poles, Aue reflects:

> You must think I'm explaining all this to you rather coldly: that's simply in order to demonstrate to you that the destruction by our deeds of the people of Moses did not stem solely from an irrational hatred of Jews – I think I've already shown how poorly the emotional type of anti-Semite was regarded by the SD and the SS in general – but above all from a firm, well-reasoned acceptance of the recourse to violence to resolve the most varied social problems, in which, moreover, we differed from the Bolsheviks only by our respective evaluations of the categories of problems to be resolved: their approach being based on a

> horizontal reading of social identity (class), ours on a vertical one (race), but both equally deterministic (as I think I've already stressed) and reaching similar conclusions in terms of the remedy to be employed.... (pp. 669–670)

Apart from the somewhat veiled language ('recourse to violence' and 'remedies' for mass murder or genocide), the difficulty with this argument is the reduction of groups of people and whole peoples to a social problem, thereby negating their humanity (in the sense of human dignity). By the way, this type of thinking was for Hannah Arendt the ultimate reason why she thought the death penalty the appropriate sentence for Adolf Eichmann.[8] In the comparison with the Bolshevik wrongdoings this justification shows the simple reasoning that 'I cannot be blamed because my enemy does the same thing', which Cohen counts as the condemnation of the condemners in an almost a childlike 'everyone is doing this – why pick on me', and questioning the critics' right to criticize (Cohen 2004, pp. 97–98).[9]

17.4.2 Blaming the Victim

There are two other large relevant passages, but here the justifications are brought forward by other characters: a not very military-like colleague, Ohlendorf, and the evil genius in the background of Aue's life, the powerful industrial Mandelbrod. Both passages are near monologues, in reaction to Aue expressing his unease over the murder of the Jews and other civilians. He doesn't counter their arguments, and therefore it is suggested (but not decisively so) that he is convinced by them. Ohlendorf argues that the initial idea was to make the German Jews emigrate to Poland and further on to the East or to Madagascar, and that it was due to a lack of cooperation by others that these plans came to nothing and a harsher solution was necessary. Ohlendorf deplores the *Vernichtungsbefehl*, and considers it a mistake and a failure, but at the same time agrees to its necessity:

> It's a mistake because it's the result of our inability to manage the problem in a more rational way. But it's a necessary mistake because, in the present situation, the Jews present a phenomenal, urgent danger for us. (p. 220)
>
> The war goes on, and every day that this enemy force remains behind our lines reinforces our adversary and weakens us.... The Jews are praying and striving for our defeat, and so long as we haven't won we can't nourish such an enemy in our midst. And for us, who have received the heavy burden of carrying out this task to the end, our duty toward our people, our duty as true National Socialists, is to obey. (pp. 223–224)

[8] Arendt closes the Epilogue of her report of the Eichmann-trial with a virtual address to Eichmann, of which the last sentences are: 'And just as you supported and carried out a policy of not wanting to share the earth with the Jewish people and the people of a number of other nations – as though you and your superiors had any right to determine who should and who should not inhabit the world – we find that no one, that is, no member of the human race, can be expected to want to share the earth with you. This is the reason, and the only reason, you must hang' (Arendt 1963, pp. 255–256).

[9] In a sense it was this argument that fired the *Historikerstreit* in the 1980s, and more recently the (milder) controversy over Timothy Snyder's *Bloodlands* (Richard J. Evans 2010).

This seems to me a typical case of blaming the victim: without their persecution the Jews would not, of course, have prayed so fervently for the defeat of Germany. It is again a strategy of denial of the victim. Cohen sees this as drawing on 'just world thinking': 'In a just world, suffering is not random; innocent people do not get punished arbitrarily. They must have done something. They deserve to suffer because of what they did, must have done, support doing, or will do one day (if we don't act now)' (Cohen 2004, p. 96). The last sentence of the Ohlendorf quotation shows the appeal to higher loyalties, characteristic of the rhetoric of nationalism (98), and also a kind of reversal of perpetrator and victim ('the heavy burden of carrying out this task').[10]

Mandelbrod shows another, rather daring (if perverse) example of the denial of the victim by reversing, or rather equating perpetrator and victim, National Socialism (or Aryan Germany) and Judaism. He considers the Jews the only worthy enemy of Germany, because they were the first to adhere to purity of race, the first to acknowledge the bond between people and land, *Blut und Boden*, and the first to see themselves as the chosen people. ('There's room on this earth for only one chosen people, called on to dominate the others: either it will be them, as the Jew Disraeli and the Jew Herzl wanted, or it will be us' p. 456). I leave aside the misconception that in Jewish thought being chosen would mean: called on to dominate the others. The reasoning is not very consistent with, even diametrically opposed to, other considerations, which look upon the Jews as inferior people, who, like the disabled and mentally handicapped, pose a threat to the genetic quality of the German people. This latter perception is dominant in Aue's views, as is shown by his surprise about the rising of the Warsaw Ghetto and the fight the starving, poorly armed Ghetto inhabitants put up. He cannot but have respect for their achievement. For Mandelbrod it would have been the confirmation of his theory.

17.4.3 Referring to a Higher Authority

Aue is shocked when he hears about the operation T-four, in which the German war cripples who wouldn't recover enough to resume their military duties, were killed in gas trucks. Then he starts musing about inhumanity and guilt and comes to the conclusion that officer Döll of T-four is considered a criminal only because he had the misfortune to be born in Germany instead of France or America. Referring to Thomas Hobbes, he states that man needs an outside authority to make him accept the constraint of the Law. That authority shifted from God to the king and then to the people or the nation. National Socialism sought it in the *Volk*. The Führer embodies this sovereignty of the *Volk*.

[10] Cohen gives the example of Golda Meir who reproached the Arabs 'for "making" nice Israeli boys do all those terrible things to them' (Cohen 2004, p. 96).

> From this sovereignty the Law is derived, and for most men, in all countries, morality is nothing but Law: in this sense Kantian moral law ... stemming from reason and identical for all men, is a fiction like all laws (bur perhaps a useful fiction)....[11] So for a German, to be a good German means to obey the laws and thus the Führer: there can be no other morality, since there would be nothing to support it. (And it's not by chance that the rare opponents of our power were for the most part believers: they preserved another moral reference point, they could judge Good and Evil on another basis than the will of the Führer, and God served them as fulcrum to betray their leader and their country ...). (pp. 591–592)

Consciously or not, the narrator uncovers here the basic flaw in his morality (and that of most other national socialists): the fact that he as an individual has parted with his own moral judgment altogether and placed it in the hands of the Führer. It is exactly this mentality that Hannah Arendt saw in Eichmann and in her opinion it did in no way count as an excuse, though the majority of Germans acted the same. Persons who 'went really only by their own judgments' were the exceptions, and the only ones who behaved 'normally'.[12]

17.5 Comparison with Historical Characters

This kind of deliberate reasoning is rare in contemporary perpetrator texts. German SS-men do not often appear to be concerned with the justification of their actions. When they show themselves doubtful over certain incidents, their main concerns turn out to be things like military decorum, the stress for their subordinates, and pride in a well-performed job (cf. Klee et al. 1988). These features are also to be found in *The kindly ones*, but more so in the narrative parts than in the pages where Aue is reflecting as a narrator post hoc.

In interviews with Dutch SS-men, conducted some 20 years after the war, one of the striking themes is the pride they take in their physical and mental abilities, the fact that they had proved to be equal to the severe demands of the military elite corps (Armando and Sleutelaar 1967). As for the Holocaust they show the same range of excuses that Stanley Cohen analyses in *States of denial*: some bad things happened, but the number of six million is quite exaggerated; the bad things happened out of

[11] Some 25 pages earlier Aue tells about a social gathering with the Eichmann family, where chamber music is played and Eichmann discusses Kant's *Critique of Practical Reason* with Aue. In the same line of reasoning as the quotation, they agree that there is no contradiction between the *Führerprinzip* and the Kantian imperative: 'Act in such a way that the Führer, if he knew of your action, would approve of it' (p. 566). According to Hannah Arendt, Adolf Eichmann might have known Hans Frank's re-formulation of the categorical imperative: 'Act in such a way that the Führer, if he knew your action, would approve it.' (Arendt 1963, p. 121), referring to Hans Frank, *Die Technik des Staates*, Munich: Rechtsverlag, 1942, pp. 15–16).

[12] Cited after Barry Sharpe (1999), p. 8. The way Sharpe sums up the attitude of Eichmann also pierces through Aue's learned philosophical reasoning: 'Eichmann, and many others like him, sought security and comfort in letting others choose for him and abdicating choice and responsibility' (Sharpe 1999, p. 6).

impotence (e.g. the hunger in the concentration camps); we didn't know about it; the Jews were morally despicable people and so were to blame for their fate (e.g. the only murder one SS-man witnessed, was of a Jewish man who pleaded for his life, offering his wife and daughter in exchange – presented as proof of his moral deterioration), etcetera. In their simplicity these considerations differ pointedly from Max Aue's reasoning. The fact that these Dutchmen were not, for the most part, high-ranking officers, and speaking 20 years later to fellow countrymen of the nation they, according to common opinion, had betrayed by joining foreign armed forces, may to a great extent account for this difference.

Aue's biography shows similarities with the lives of some non-fictional, high-ranking perpetrators, and his justifications have much in common with theirs. With Rudolf Höss, the commander of Auschwitz, he shares the early loss of his father and a negative attitude towards the Roman Catholic Church, in the case of Höss due to a betrayal by his confessor, in the case of Aue due to the unpleasant Catholic boarding school he is sent to by his mother and stepfather (Höss 1959). With Oswald Pohl, who was responsible for the industrial and economical management of the concentration camps, he shares an alleged aversion to unnecessary cruelty and the pride to have counteracted it (Krondorfer 2010).[13] With Adolf Eichmann he shares the emphasis on his lack of personal hatred for the Jews, and his interpretation of Kant's categorical imperative (Arendt 1963, p. 121). With Franz Stangl, the commander of Treblinka, he shares the start of his career in the SD, which is motivated by a certain measure of fear. The Austrian Stangl had, as a policeman before the Anschluss, discovered a secret Nazi arms supply; he was afraid that this previous history would harm him or even put him in danger when the National Socialists took over in Austria, and this made him comply with orders to perform tasks he didn't like, such as assisting in the euthanasia programme in Schloss Hartheim, and the command of Treblinka, at least according to his own view (Sereny 1974). With Aue it is an arrest for homosexual actions in 1937, from which he is saved by his later friend Thomas Hauser, who engages him for the SD. In the first phase of his career Aue makes the mistake to write an honest report on the political situation in France instead of reporting what his superiors want to hear. He fears to be slotted into a dead-end function in Berlin, and so he eagerly grasps the opportunity to go to the front in the Ukraine. Afterwards Stangl may have overdone his fear as the motivation of his compliance, by way of justification; Aue on the other hand downplays the initial negative motivation in the development of his career, although the threat could have been stressed more, since he continues to have homosexual relationships (or more correctly: encounters), that are, in the light of the Nazi severe ban on homosexuality, potentially dangerous for him.

[13] Oswald Pohl wrote in his religiously coloured confession *Credo* he had 'Unmenschlichkeiten nachweisbar energisch entgegengetreten' (Cited after Krondorfer 2010, p. 205).

17.6 Transgressive Sexuality

It is not until this homosexual encounter in 1937 that the disclosure of Aue's uncommon position *in sexualibus* starts. Gradually we learn that he had a childhood sweetheart and was cruelly separated from her. This girl much later turns out to be his twin sister, Una. As children they had sex, anally since she started to menstruate. When this incestuous relationship is discovered, he is sent to a strict boarding school, where he gets into a homosexual relationship with an older boy; during the sex he tries to identify with the feelings of his sister (pp. 200–203). In his early twenties he meets Una again after years of separation, and they have sex once more, but only one single time, because Una sees their childhood relationship as past, and doesn't want to engage in it anew. But Aue's story becomes even more peculiar. After he is wounded and on recuperation leave, he goes to the South of France, where his mother lives with his stepfather. He meets the twins, boys about 7 years of age, and he murders his mother and her husband, without having any recollection of it later on. The twins, who mysteriously disappear after the murder, turn out to be his and Una's children. In one of the last chapters he visits the Pomerania estate of his sister, who is married to a Prussian *Junker*. The couple is abroad, and in the solitude of the large, cold house his feverish dreams and fantasies and his rambles in the woods are sex ridden; he, for instance, penetrates himself with a tree branch (p. 898). He ends up in Berlin, killing both a former lover and his best friend, Thomas. The murder of his former lover could be explained out of fear for exposure as a homosexual; the murder of the best friend is unprovoked. Perhaps it is the first preparation for his new post-war life by covering up his tracks as a perpetrator.

The theme of sexuality, becoming more and more extravagant and perverse, places the novel in the tradition of the so-called literature of transgression, like the Marquis De Sade, Octave Mirbeau and George Bataille.[14] Literary scientist Birgit Dahlke sees the incest motif as a way of reflection on the story-telling itself, as is not unusual with incest in literature. At the same time it provocatively puts Aue in line with mythological heroes like Oedipus and Orestes, and thus with the great cultural tales of Europe (Dahlke 2010).

The transgressive streak can be seen as closely intertwined with the realistic, recognizably historical plot line (which Birgit Dahlke denies however): the cover-up of his homosexuality, which in turn is the consequence of his incestuous love for his sister, is the starting point of Aue's SS-career; the love for his disappeared father, a sadistic brute as he in the end turns out to have been, and the hatred for his French stepfather lead to his youthful abhorrence of the bourgeois life the latter is leading, and thus to his susceptibility to German nationalism and the pursuit of the absolute. Generally the protagonists of transgressive fiction are supposed to be outsiders, who feel confined by the moral boundaries set by culture and society,

[14] In his insightful review in *The New York Review of Books* Daniel Mendelsohn, who is a classicist, goes into both the mythological and the transgressive dimension (Mendelsohn 2009).

whereas Aue is supposed to represent the attitude of the average German, or at least the average National Socialist. Indeed, in the first pages the narrator does everything to convince the reader that he is just like him. This view, unsettling as it may be where his actions and attitude in the Holocaust are concerned, is stretched to its limits by the transgressive plot line, that towards the end gains in strength until it dominates the realistic line altogether. Thus this at first rather convincing view is not only stretched *to* its limits, but *beyond* them, and so in my opinion the element of transgression decidedly weakens the moral claim that is the heart of what the novel wants to convey. It gives the reader an easy way out, providing him/her with an excuse to think: I am not at all like you!

17.7 Gender

Many a book of transgressive fiction can be accused of misogyny, but this reproach wouldn't be true for *The Kindly Ones*. On the contrary, if any it is the female characters that criticize the National Socialist ideology and the practice of warfare.

His sister Una questions the necessity of war, when upon seeing convalescent soldiers on crutches and in wheelchairs, she sighs: 'What a waste' (p. 481). In the same conversation she asks Aue if he himself has ever shot unarmed people, and what he felt upon doing so, and her only reaction to his answer is: 'I'm happy I'm not a man' (p. 482). Later on she counters Aue's anger at the allied air raids on Cologne, by equalling the German victims, innocent women and children, to the civilians he himself helped killing.

His mother, whom he loathes since she had his father declared dead and remarried, questions him accusingly on the fate of the Jews, and when he makes evasions and excuses, referring to the French helpfulness in the persecution of the Jews, she concludes the conversation with an angry 'You are completely mad' (p. 525).[15]

Another important female character is Hélène, a young military widow, with whom Aue has a sort of romantic (but non-sexual) relationship during the second half of the war. When he is very angry at her for nursing him during an illness, he throws some Holocaust facts at her. Afterwards she asks him if those facts were true, he confirms it, and she is deeply bothered with it (biting her lips, averting her face, tears coming to her eyes), and says that they will have to pay for it, even if they will not lose the war (818). From then on her nursing seems to get a colder, more automatic quality.

A very minor character is Eichmann's wife, who is the only one not amused when Aue tells a joke about his landlady commenting on the gassing of the Jews; she

[15] In the original French this denouncement regards the National Socialists in general, not only her son: 'Vous êtes complètement fous.' (484). Since she is talking to her son, the 'vous' must be interpreted as a plural and not as the polite form of address.

seems to want to ask a question, but holds back (p. 565). Her attitude resembles that of the wife of Franz Stangl, as recorded by Gitta Sereny: knowing that something is not right according to her own moral principles but refraining from going into it, in order to avoid having to take a stand against her husband. In the female characters a second voice can be heard, a melody from which morality has not disappeared altogether, but on the whole it is a very weak, subdued voice – as it was in the historical reality.[16]

17.8 The Link to Classical Mythology

The murder of the mother and stepfather links Littell's novel to the classical Orestes tragedies. So does the title: a translation of 'Eumenides', the title of the third part of Aeschylus' *Oresteia*. It is another name of the Erinyes or Furies, the goddesses of vengeance. In Greek mythology the Erinyes pursue criminals, driving them crazy. They became Eumenides when a wrongdoer was cleansed from his guilt. In the Aeschylus tragedies Orestes kills his mother and her lover because they had killed his father. Killing one's mother is one of the worst possible crimes; the Furies who pursue him reject his defence that he had to avenge his father, an obligation just as compelling as the prohibition of matricide. The goddess Athena finally sees to it that a court of law is established that acquits him of the charge. Willy-nilly the Furies agree and become Eumenides. In tragedies by Sophocles and Euripides the role of Orestes' sister Elektra is more prominent. Orestes grows up at a foreign court, while Elektra stays with her parents. After her father is murdered Elektra is married off to a poor farmer. When Orestes and Elektra are reunited she fervently supports his plan to avenge their father.

The similarities and dissimilarities with Littell's novel are clear: Aue kills his mother and his stepfather, because he blames her for the disappearance of his father. In the classical tragedies there is no incestuous love between brother and sister (although Elektra gave her name to the female version of the Oedipus complex, so her name may have a slightly incestuous connotation). The way brother and sister grow up apart corresponds, though the reasons are different. The title seems to suggest that Aue, like Orestes, is acquitted as well. The last sentence of the novel confirms this, yet the formulation is a bit ambivalent. Aue is left in Berlin, near the Zoo, with the Russians coming near, and he has killed his friend Thomas who

[16] Also fanatically National Socialist women figure in the novel, such as the beautiful, very Aryan and very intimidating female body guards of Mandelbrod and Leland. They look so identical that Aue can't tell them apart. And there is the very unattractive female linguistic scientist in the Caucasian episode, whose advice is called upon in the matter of the *Bergjuden*. She is, against common sense and many historical data, determined to declare them really Jewish, thus condemning them to death. Both the body guards and the scientist are more cartoon-like than realistic characters, and can hardly be reckoned to form a counterweight to the (mildly) critical other women in the novel.

a minute earlier shot a police officer, who, like a Fury, had pursued Aue since the murder of his mother:

> I felt all at once the entire weight of the past, of the pain of life and of inalterable memory, I remained alone with the dying hippopotamus, a few ostriches, and the corpses, alone with the time and the grief and the sorrow of remembering, the cruelty of my existence and of my death still to come. The Kindly Ones were on to me. (p. 975)

The phrasing ‘were on to me’ (in French: *avaient retrouvé ma trace*) expresses the feeling of being pursued, haunted by the past, while the name ‘the Kindly Ones’ in its reference to the role in the classical tragedies indicates the opposite: finally coming to rest, in harmony with the world.

17.9 Conclusion

Generally there will be little argument whether Max Aue’s actions are good or bad. Not many readers will vindicate him for the coldblooded killing (or consenting to be killed) of children and other unarmed civilians. The fact that the author singled out a dignified, very old man and children as some of the individual victims seems to indicate that he wanted to stress, maybe even augment, the moral guilt of the main character. Why is it that the reader may well lose his empathy with Max Aue especially in these episodes? He shows respect for the old *Bergjude*, an appreciation for and congeniality with the musical talent of Yakov, tenderness for the little girl, a certain sympathy with the teenager – and nevertheless does not hesitate to let them be killed. His victims are not part of an anonymous, amorphous mass; they have become individuals for him. It is as if he spits on the philosophy of Emmanuel Levinas: he did look his fellow humans in the face, but the ethical appeal of this face seems to be completely lost to him. The claim that he is in moral respect just like the reader, will not sit easily with that same reader, however forcefully this claim is confirmed by social-psychological experiments. Luckily for the reader, he is offered a comfortable way out of seriously considering the claim in view of Aue’s peculiar private life.

The many considerations about whether or not the killings are legitimate or justifiable make it impossible to look upon Aue as an a-moral character. The basic cause of his immorality in the historical plot line is the dominance of the collective over the individual in his way of thinking. This goes as much for his own individuality, which he to a great extent sacrifices to the idea of the nation or the *Volk*, embodied in the *Führer*, as it does for the individuality of the victims, whom he sees first and foremost as parts of a hostile collective, however incredible their potential danger as such may be. With regard to morale the importance of an individual outlook, in respect both of one’s own moral judgment and of the possible victims, thrusts itself upon the reader. Yet if any moral lessons are to be learned from the fictional character of Max Aue (and from the many real Holocaust perpetrators like him), it is foremost the fragility of morale and the sense that philosophical reasoning doesn't protect us from wrong choices.

The Kindly Ones shows that moral self-reflection is not a sufficient answer to the question 'what makes us moral?' (understood as: 'what makes us act morally?'). Keeping up one's individual judgment and not handing it over to the collective's morale may be an indispensable element of being moral or acting morally, but how to do so under pressure, and how one's individual judgment is built up in a way that is not only empowering but also ethically right, is something else. It is a question that *The Kindly Ones* leaves it to the reader him/herself to find an answer for.

Acknowledgments I want to thank Professor Dr A. van Harskamp, Dr M.H. Schenkeveld and Drs Liesbeth Eugelink for the careful attention they gave to early drafts of this paper. In particular I am appreciative to the anonymous reviewer of the next to final version. I made grateful use of their remarks, as far as I was able to.

References

Arendt, H. 1963. *Eichmann in Jerusalem. A report on the banality of evil*. London: Faber and Faber.

Arendt, H. 1994. Organized guilt and universal responsibility. In *Essays in understanding*, ed. Jerome Kohn, 121–132. New York: Harcourt Brace.

Arendt, H. 2004. Persoonlijke verantwoordelijkheid onder een dictatuur. In *Verantwoordelijkheid en oordeel,* ed. Jerome Kohn, 51–75 (translation of *Responsibility and judgment*, 2003). Rotterdam: Lemniscaat.

Armando, and H. Sleutelaar. 1967. *De SS'ers*. Amsterdam: De Bezige Bij.

Binner, R., et al. (eds.). 1998. *Wiens schuld? De impact van Daniel Jonah Goldhagen op het Holocaustdebat*. Houten: Van Reemst.

Browning, C.R. 1992. *Ordinary men: Reserve Police Battalion 101 and the final solution in Poland*. New York: HarperCollins.

Browning, C.R. 2003. *Collected memories. Holocaust history and post war testimony*. Madison: Wisconsin University Press.

Browning, C.R. 2012. *Revisiting the Holocaust perpetrators: Why did they kill?* (Nooit meer Auschwitz lezing). Amsterdam: Nederlands Auschwitz Comité, NIOD & Sociale Verzekeringbank.

Cesarani, D. 2004. *Becoming Eichmann; Rethinking the life, crimes and trial of a 'desk murderer'*. Cambridge: Perseus.

Cohen, S. 2004. *States of denial; Knowing about atrocities and suffering*. Cambridge: Polity Press.

Dahlke, B. 2010. Nuda Veritas? Zum Effekt des Pornographischen in Jonathan Littells Roman "Die Wohlgesinnten". In *Scham und Schuld. Geschlechter(sub)texte der Shoah*, eds. Figge Maja, K. Hanitzsch, and N. Teuber, 301–307. Bielefeld: Transcript Verlag.

Desbois, P. 2008. *The Holocaust by bullets; A priest's journey to uncover the truth behind the murder of 1,5 million Jews* (translated from the French (2007) by Catherine Spencer). New York: Palgrave Macmillan.

Evans, R.J. 2010. Review of *Bloodlands: Europe between Hitler and Stalin* – by Timothy Snyder. *The London Review of Books* 32(21):21–23.

Goldhagen, D.J. 1996. *Hitler's willing executioners; Ordinary Germans and the Holocaust*. New York: Alfred Knopf.

Höss, Rudolf. 1959. *Commandant of Auschwitz; The autobiography of Rudolf Höss* (translated from the German by Constantine FitzGibbon). London: Weidenfels and Nicolson.

Klee, E., et al. (eds.). 1988. *"Schöne Zeiten". Judenmord aus der Sicht der Täter und Gaffer*. Frankfurt am Main: S. Fischer Verlag.

Krondorfer, B. 2010. Männlichkeit und Selbstmitleid; Religiöse Rhetorik in Selbstzeugnissen von NS-Tätern. In *Scham und Schuld. Geschlechter(sub)texte der Shoah*, ed. Maja Figge, K. Hanitzsch, and N. Teuber, 195–221. Bielefeld: transcript Verlag.

Lipstadt, D.E. 2011. *The Eichmann trial*. New York: Schocken.

Littell, J. 2010. *The kindly ones*. London: Harper Perennial (translated from the French by Charlotte Mandell, *Les Bienveillantes*, 2008).

Mendelsohn, D. 2009. Transgression. *The New York Review of Books* 56(5), 26 Mar 2009. www.nybooks.com/articles/22452.

Milgram, S. 1974. *Obedience to authority; An experimental view*. New York: Harper & Row.

Newman, Leonard S. 2002. What is a "social-psychological" account of perpetrator behavior? The person versus the situation in Goldhagen's Hitler's willing executioners. In *Understanding genocide: The social psychology of the Holocaust*, ed. L.S. Newman and R. Erber, 43–67. Oxford: Oxford University Press.

Nussbaum, M.C. 1990. *Love's knowledge. Essays on philosophy and literature*. New York: Oxford University Press.

Safranski, R. 1998. *Das Böse oder das Drama der Freiheit*. München: Carl Hanser Verlag.

Sereny, G. 1974. *Into that darkness; From mercy killing to mass murder*. London: Deutsch.

Sharpe, B. 1999. *Modesty and arrogance in judgment: Hannah Arendt's Eichmann in Jerusalem*. Westport: Praeger.

Wieviorka, A. 2006 *The age of the witness.* Ithaca: Cornell University Press (translated from the French by Jared Stark, *L'ère du témoin*, 1998).

Zimbardo, P. 2007. *The Lucifer effect; How good people turn evil*. London: Rider.

Part V
Morality Beyond Naturalism

Chapter 18
Society and the Origin of Moral Law: Giambattista Vico and Non-reductive Naturalism

David Edward Rose

18.1 The Naturalistic Framework

The following paper is conjectural and a little radical. For these reasons alone, it is probably wrong or, at best, on the wrong track. The central supposition is, though, at least worth entertaining, namely that the tendency in theoretical ethics to concentrate on supposed 'natural' traits, such as aggression or compassion and their historical manifestations in animal behaviour or primitive societies, in order to explain the origin of moral law scientifically entails that the story we tell ourselves about these origins is descriptively erroneous and, worse than this, has negative normative consequences. Rather, one should perhaps look for the origin of ethics and moral law in another human capacity (and not a desire) as Vico does, viz. the imagination.

Prior to developing the central theme of this paper, it would be pertinent to reveal the axiomatic background that supports such scientific stories. If one is to ask what the origin of moral law is, one implicitly assumes that there was a time in which there was no moral law and, then, there was a succeeding time in which there was a morality. An animal, in short, became what we would readily recognize as human. The difference between the pre-human animal and the human animal is to be understood as the emergence of rule-following, so that no matter how pressing an instinct or an immediate desire, morality is present when a being can be observed acting contrary to some deep animalistic desire. The ideal case would then be putting one's life at risk for the sake of a principle (and not just for the love of a family member).

The explanation of the emergence of moral law is most acceptable when consistent with scientific rationalism and scores even more points when it coheres

D.E. Rose (✉)
Department of Philosophy, Newcastle University, Newcastle upon Tyne, UK
e-mail: david.rose@newcastle.ac.uk

B. Musschenga and A. van Harskamp (eds.), *What Makes Us Moral? On the capacities and conditions for being moral*, Library of Ethics and Applied Philosophy 31,
DOI 10.1007/978-94-007-6343-2_18, © Springer Science+Business Media Dordrecht 2013

with reductionism, that is when a complex phenomenon (morality) can be explained using concepts belonging to a simpler and more robust level of explanation (biology and, ultimately, physics). Such a species of scientific rationalism has certain prevalence in contemporary ethical thinking (Jackson 1997; Railton 2003). Simply put, the position is that any explanation about rule-following in human beings has to be capable of being reduced to elements that could feature in a psychological explanation of behaviour. And such a position has modern, historical precedence in the writings of Hobbes (1996) and Locke (1988). At the level of sociology, the atomistic units of explanation are individuals and, at the level of psychology, the atoms of explanations are instincts and desires. And, of course, instincts and desires are reducible to chemical reactions in the brain and explicable under the conditions of natural-scientific law. One cannot just help oneself to an *ad hoc* faculty of moral choice or practical reason. To do so is only to move the focus of the explanation: what separates the pre-human from the human with practical reason?

It seemingly follows that, to understand the origin of law, it would be natural to look for what those first human beings, and their immediate predecessors, share with the civilized human being. And such a common property is to be found in the structures of the mind, understood by psychology as the motivations and instincts that make us who we are. Much of the contemporary work devoted to the origins of ethics and morality concentrates on the naturalistic origins of evaluative practical reason. Since such an endeavour is assumed to be scientific (and scientific is most readily understood in contemporary thought as evolutionary biology) then such an approach is eminently sensible. And the easiest way to understand morality on such a picture is as an adaptive trait responding to the fact that humans, to better further their interests, live in groups. The pivotal element in such a picture is aggression, whether it is the cause of the moral conscience, or whether innate compassion is the reason that human beings, unlike other animals, can repress their aggression. What we share with our ancestors is either an aggressive instinct or the innate capacity to repress such an aggressive instinct. In either case, it is the ability to repress and act against the aggressive instinct which is the first form of moral rule-following.

Since the seventeenth century, intellectual ethical thought has manifested itself in a ‘core’ or standard articulation of the conditions and aims that an explanation of a moral system ought to meet, as well as defining the ‘permissible alternative’ way of thinking about what it means to be a moral being. The naturalistic tradition in moral philosophy assumes that the basic capacities and characteristics of the individual exist prior to his or her socialization and that morality arises out of the necessity for cooperation. The argument of the present paper proceeds by, first, demonstrating that if reductive naturalism is offered as a description of the origin of rule-following, it cannot explain the origin of moral laws in any ordinary sense at all. Second, if reductive naturalism is a normative theory from the outset, then it begs the question in favour of certain normative theories and therefore its claims rest on the truth of the description of the origin of law. Finally, the paper will outline an alternative Vichian moral sociology which is naturalist but non-reductive and also indicate what the empirical consequences of this new methodology may be.

18.2 The Standard Explanations of the Origin of Moral Systems

The explanatory dichotomy of moral systems is characterized by, on the one hand, exemplars such Hobbes and Freud who argue that human beings possess a moral conscience because they renounce aggression and, on the other hand, in writers such as Locke, that humans renounce aggression because they have a moral conscience (Hobbes 1996; Freud 2002; Locke 1988).[1] Both explanations are consistent with natural science in that an individual atom is identified (the person) with a natural property (aggression or compassion) that can be reduced to a more basic science (first psychology and then evolutionary biology) via a simple account of practical reason (egoism). Neither, though, is completely coherent, especially since their supposed main virtue is their coherence with contemporary scientific rationalism.[2]

The first – or ‘core’ – explanation of morality is that human beings follow rules because they renounce aggression or, less superficially, because aggression is internalized (Hobbes 1996; Freud 2002). Our aggressive natures explain moral values consistent with psychology: basic egoism dictates that an agent represses aggression when cooperation would benefit him or her and these repressions allow for the proper functioning of harmonious society. Such a simplistic psychology grounded in drives and desires is reducible to the more robust natural sciences.

The second – or ‘permissible alternative’ – explanation of the existence of moral systems is the opposite in that human beings are able to renounce aggression because they possess an innate predisposition to be cooperative, but this itself would be reducible to terms in psychological discourse such as compassion (Locke 1988). Social existence is made possible by the fact that humans have behavioural instincts that allow them to work co-operatively with others in order to further their own shared and collective interests, Having a moral conscience is a psychological form of reasoning akin to having a linguistic faculty that can occupy a place in the naturalistic universe.[3]

[1] I here use ‘Hobbes’, ‘Freud’ and ‘Locke’ to stand in for standard positions, but pay little attention to the subtlety and nuances of their particular approaches. For this I apologize, but do so for reasons of expediency. For example, I believe that the indifference to at least thinking through the theological aspects of Locke’s theory does serious damage to his theory as a whole.

[2] It should be noted that, for Locke, practical reason is ad hoc and not scientific: the justification of the moral conscience is theological in nature and hence not consistent with modern science. However, there exists a secondary justification in self-ownership in which the human being is aware of his body as his own and reciprocally realizes that others, too, own their bodies (Locke 1988, § 27). Such philosophical explanation has its counterparts in evolutionary discourse: a debate has opened up between those who believe that human cooperation and group regulations can be traced to, on the one hand, the aggressive instincts of the chimpanzee and those, on the other hand, who trace them to the social instincts of the bonobo (Peterson and Wrangham 1997; Kano and Vineberg 1992). So, it is possible to use Locke’s general description in a way consistent with natural science and independent of theological assertions.

[3] By mentioning only Hobbes and Locke, there is a felt lack of the third principal character of modern, social contract theory, viz. Rousseau. Rousseau is more complex and like Vico he sees

So far, so coherent with our overall explanatory system: one defines the ideal, standard theory consistent with science, that is naturalism in which human beings are animals explicable through instincts that are reducible to simple psychology and biology, but at the same time, allowing for a permissible alternative, less consistent with reductive naturalism, but still possibly explicable in terms of evolutionary biology, that we are innately group animals, whereby aggression plays second fiddle to cooperation.

Both accounts of moral rule-following are descriptively problematic in the domain of scientific naturalism. The first problem is obvious in that if moral laws are reified behavioural instincts, then there is a fact–value gap. Scientific laws describe, whereas moral laws prescribe and both Hobbes and Locke aim to show how human beings *should* behave, not how they *do in fact* behave. A human being acts in a cooperative way, i.e. follows rules, because of the need to further his or her own interests (Hobbesian) or because of the existence of an innate, compassionate disposition (Lockean). Yet, this explains nothing. Let us imagine a rather poor evolutionary biologist explain this: by cooperating, an individual is more successful and therefore reproduces more often (one assumes living longer and also being cooperative in the sexual activity itself aids reproduction!). So, why does an agent keep promises? Because an agent is a rule-follower and agents are rule-followers in order to better reproduce and pass on genetic material. (Notice here how a by-product of success becomes the aim of success!) You ought to keep your promises, states the biologist, in order to better pass on genetic material.

Now, if the story being told is descriptive and merely holds that there are more rule-followers than not because they are more successful in reproduction than this is fair enough, but it cannot tell us how individuals will behave since it is too general, so it has very little predicative power as a scientific theory. If, as often happens, the by-product is transformed into a good (as above), then it seems that the biologist is telling me I ought to keep my promises because then I shall reproduce better, but if I ask why I should actually reproduce better anyway, he cannot without absurdity hold that to be a normative, prescriptive law (putting aside its consequences for population, contraceptive practices and sexual orientations). Take rock climbing: if I partake in rock climbing, I am putting my life in danger. To do so, according to the biologist, is to do something morally wrong because I have a

society as an effect of language, which Locke and Hobbes both think predates society. Unlike Vico, though, Rousseau holds that the rule of nature is innate and guarantees liberty and equality beyond social convention, rather than an imagined effect of society. There is a further 'radical' alternative symptomatic of Aristotelians and Hegelians in that a moral conscience is derived from social existence (like the core explanation above) but that (unlike the core explanation) social existence is natural, an organic development of our existence. Human beings naturally and necessarily exist in societies. Such an alternative cannot be considered here and can be dismissed by the terms of the present debate. Neither Hegelians nor Aristotelians are coherent with modern science. For the former, ethics and politics are part of a discourse that is inflated from poorer and simpler scientific ones (rather than reduced to it); and for the latter, explanation is based in a metaphysics wholly contradictory to modern science. Vico, as we shall hopefully see, is different because perhaps he gives us a way to link the science and morality in a different way.

moral obligation (and not just a disposition as he is no longer describing action) to preserve my life. But that is surely ridiculous. Facts of this sort cannot give us moral prescriptions at all.

And it is not just straw men who suffer from this fact-value gap. Hobbes, for example, presents a picture whereby aggression is characteristic of human relationships and conflict arises due not only to scarcity of goods but also human nature. Aggression can only be overcome by agreeing to a contract (a non-moral meta-ethical agreement) to observe standard moral laws because such rule-observance is in the agent's particular interest. Yet, the problem is obvious: if human nature is understood as egoism, then the individual agent will act so as to secure his interests. Such individualism causes conflict between individuals and conflict is not in an individual's interest. So, egoism dictates that an agent must act so as to minimize conflict and that entails making an agreement to follow rules. To enter into a contract one has to be rational (and then reason tells the agent what he *should* do), yet the contract is required because one is not rational because reason tells him to satisfy his interests. So, it seems if taken to be a descriptive hypothesis, Hobbes's account of the origin of moral law is just not plausible. Of course, his hypothesis is not actually a description, but rather a justification of political authority, so it is a normative claim and we shall return to that below.

It seems that the core explanation dissolves into the permissible alternative because humans must have a co-operative instinct if they are to make a meta-agreement to be moral. If human nature is egoist, humans would never be able to make agreements and society would not begin. If, however, egoism is tempered by willingness to compromise, then it seems that the permissible alternative of Locke is descriptively the case.[4] Locke describes human nature as innately adapted to cooperation: human beings naturally respect the life, labour and property of others and enforce violations of these rights. Thus, a pre-legal social formation can exist and prosper within limits. Legalization of punishment is required in response to the need to codify the executive right to justice and social institutions are required to condition the use of capital in production.

However, if such a position is supposed to be coherent with naturalism, it has to explain such metaphysically puzzling entities such as right and desert. The simple deontological approach, whereby there are rights and there is just distribution of benefits and burdens, does not actually do any explanatory work at all since it cannot be reduced to simpler elements as demanded by scientific rationalism. There is also the further empirical problem of the existence of violations of law if these rights and duties are natural laws. In a nutshell, the position needs to explain the fact of human wrongdoing. If a human being possesses a moral conscience, then why does he or she commit actions which we would identify as wrong and why do we need

[4] Such a descriptive problem is equally true of Freud's purely psychological description of the formation of the super-ego through the oedipal development phase. If internalization occurs through turning aggressive instincts on oneself, it can only be dependent on the equally important love of the father and not just aggression towards him. In fact, love seems stronger than aggression.

society to enforce moral law at all? Locke puts forward several reasons for conflict in his descriptive account (ignoring here the effects of the invention of money): human passions overriding rationality, zealousness in retribution, or partiality in the treatment of others. Again, there is a very obvious fact–value distinction at the heart of the explanation: morality tells how we ought to act, whereas science describes how we do in fact act.

Given this, and also Hobbes's claim that human nature both causes and resolves conflict, it is clear that, in both descriptive accounts, plausibility rests on a distinction between reason and inclination. However, to be consistent with naturalism, reason has to be (as Hobbes unsuccessfully attempts to demonstrate) a development of inclination and interest-based desire. Both accounts rest on a dichotomy which is not feasible if they are to be reduced to more basic discourses of science. There must, in short, be only inclination to be consistent with reductionism.

Of course, one might argue, neither Hobbes nor Locke is overtly concerned with this since the main reason for their works was the justification (normative) of authority and not its origin (descriptive). As such, the fact–value distinction is not applicable to the case since they are not writing a history of the origin of society but offering reasons why an agent ought to obey the moral law. Hobbes says it is rational, that is in your interest; and Locke says it is what you would do if you were unswayed by strong passions and partiality and the authority did not harm your interests. It would be easy to state here that such an aim is contradictory to the reductionism required of modern accounts. However, even if that point were ignored, there remain deep-seated problems with the core and the permissible alternative explanations if taken to be purely normative because they still rest on the truth-value of the descriptive groundwork.

If one conceives of morality as a compromise mechanism, as Hobbes does, that means it is only to be conceived of in terms of cost–benefit analysis and hence favours welfarist approaches. Yet, such approaches reduce all individuals to sites of welfare rather than persons. It also encourages atomism in that others are seen as entanglements and obstructions. By telling a story that may or may not be descriptively true about the origin of morality, Hobbes excludes certain moral theories (those that pitch moral law against personal inclination) and also begs the question about how to understand human relationships (others are either obstructions or aids to the satisfaction of one's interests). As such, the rationality of ethics is decided prior to any rationalization of the foundations of ethics. And so, to justify his normative preferences, one cannot just dismiss his descriptive story since it plays an integral part in the argument.

Alternatively, if morality is the product of a natural disposition and not a compromise between competing interests, then wrongdoing is to be conceived of as either the malfunctioning of reason or the capriciousness of passion and emotion. Locke's approach, like that of Kant, rests heavily on the conflict between passions and reason and, as such, moral action will be characterized as rational action. Once more, one finds that a descriptive account that is used to justify a normative commitment excludes specific ethical theories: those that, in short, rely on human emotion and its cultivation. Instead, the ideal moral agent is a robotrational-calculator that negates all particularities of character and human emotional responses. And again, Locke's normative commitments rest on the truth status of his descriptive account.

18.3 Vico's Alternative to the State of Nature

The narrative that Giambattista Vico supplies about the origins of human society and the emergence of law does not hinge on the curbing of aggressive instincts for the furtherance of individual interests, or on the natural capacity to repress aggressive instincts through reason.[5] Rather Vico, like Rousseau, holds that there exists a difference between pre-social and social humans to such an extent that the former are animals whereas the latter are humans properly speaking. For both thinkers, what separates the two entities is the possession of language (Vico 1984, § 369; Rousseau 1984).[6] Such a belief is, of course, in contradiction of the shared liberal belief that language is innate and not artificial, an assertion consistent with both Hobbes and Locke. So, the question is not so much how the human animal overcomes or economizes his or her destructive aggression, but rather what phenomenon transforms the non-linguistic animal into the linguistic human. However, it would intuitively seem that without language to make promises and contracts, there could be no society and, according to both Vico and Rousseau, without society there could be no language. Society, on these terms, is impossible.

Vico's social theory is an odd smorgasbord of traditional views. Alongside the social contract theorists, he does not believe society is an organic development, but rather an act of will. Morality is an artificial and not a natural entity. Yet, like the Aristotelians and the Hegelians, Vico holds that the human being is malleable and that society (although artificial) is necessary for a certain species of animal to become human through the re-channelling and redirection of desires and passions into the creation of a moral, socially adjusted being. Unlike either, for him the origins of civilization are non-rational and although society is not an organic development, human beings (as we properly understand them) cannot exist without society. So, how do we make sense of these contradictions since it seems so counter to the two possible positions we believe a rational thinker ought to take?

The answer is to ignore the nature of pre-social rationality (egoism) and instincts shared with other animals (such as aggression or cooperation) and, instead, to concentrate on a faculty or trait that differentiates human beings from animals and to see whether that trait can be used to explain the existence of morality, practical reason and language. If one can articulate what must have been the case for human

[5] As formerly, with Hobbes and Locke, the reading of Vico offered here is pragmatic and expedient, rather than nuanced and attentive, although I do feel it is consistent and ultimately right. Closer interpretation would be required to prove this conviction of course.

[6] All references to Vico's writings are cited by paragraph number (§).

beings to have formed a language, then one can perhaps show how the first linguistic utterances lead to the social rule of law. Vico's thesis is the rather surprising claim that human beings differ from other animals because they possess the faculty of the imagination or poetic imagination, to be precise, and that the first expressivist utterances of human beings were not simple cries of pain or joy, but the extension of the self into the unfamiliar realm of nature; measuring nature 'as if' it were like human beings. Such a trait historically emerges and, once present, even if consistent with naturalism, changes the nature of explanation from physics to sociology and the latter cannot be reduced to the former. The possession of a faculty of imagination is a requirement for foresight, comparison and moral law.

The first human-animals were motivated (like all animals) by self-preservation and a desire to reproduce (Vico 1984, §§ 369, 522–525). They roamed the primeval forests in search of those things that would satisfy egoistic interests. Animals, for Vico, 'know' only through sensation and one simple way to understand this is to hold that for animals there is no distinction between perception and representation. Human ancestors, though, were different; they had a faculty of imagination which was able to transcend the simple perceptual field in order to conceptualize the contents of perceptual states.

Such imagination allowed wonder into the world. Vico tells his reader that 'At the same time they gave the things they wondered at substantial being after their own ideas, just as children do, whom we see take inanimate things on their hands and play with them and talk to them as though they were living persons' (Vico 1984, § 375). The human being extends what he knows about himself into the world of objects, the ability to think: what if nature were like me (and no doubt a precondition for empathy and sympathy). Thus, the natural world was subject to wonder and the imagination furnished it with properties drawn from self-understanding. When there was available food, the earth was being generous, when there was not it was angry. The simplest form of understanding, the original metaphysical system, could not but be a form of animism for Vico and so any system of knowing or abstraction of knowledge must begin 'with a metaphysics not rational and abstract like that of learned men now, but felt and imagined as that of these first men must have been, who, without power of ratiocination, were all robust sense and vigorous imagination' (Vico 1984, § 375). Just as the human roars when it is in pain or angry, so the lightning of the sky is the roar of a being or a will that inhabits the 'body' of the sky (Vico 1984, § 377). The world, like the pre-human himself, could be generous, wicked, angry, and capricious, and the easiest way to conceive of a world was to take the imaginative leap into an animistic world.

These primitive pre-humans used their imagination to create a world which would respond to their actions and so began to fear it and retreat from it. The primal fear of nature is not a natural cause of action (in fact, animals cannot fear anything because they cannot have beliefs) but fear is the consequence of an interpretation of the world as a collection of wills and the creation of a system of symbolic meanings through which these wills communicate with humans. Through imagination, like children, the first humans associated lightning with the anger of a will: 'Thus it was fear which created gods in the world; not fear awakened in men by other men, but fear awakened in men by themselves' (Vico 1984, § 382). The natural world becomes symbolic, a meaningful stratum or medium in which the pre-humans invested their own passions and interests. And this fear, generated through the imagination alone, motivated the pre-humans to group in caves and transform from individuals to familial, group beings (Vico 1984, § 522).

The investment of a personality into nature and the world itself by imagination created also the proto system of moral law through the symbolic and irreal metaphysics of poetic religion. Jove, or the most powerful of the gods, was the cornerstone of such a system:

> Thus Jove acquired the fearful kingdom of the lightning and became the king of men and gods; ... From the first great benefit he conferred on mankind by not destroying it with his bolts, he received the title *Soter,* or savior... And for having put to an end the feral wandering of these few giants, so they became the princes of the gentes, he received the epithet *Stator*, stayer or establisher. (Vico 1984, § 379)[7]

If the will is angry and has so much power, he could destroy, but he chooses not to. So, humans are grateful to him and indebted to him, subject to him. Jove is a 'credible impossible'. Poetry has as its material a 'credible impossibility. It is impossible that bodies should be minds, yet it was believed that the thundering sky was Jove' (Vico 1984, § 383). Imagination works by explaining things and events through metaphor that associates the thing with the human body, its parts and human passions and sensations.

Law and rule-following is a development of poetic attempts to understand the world: '... we begin our treatment of law... at the moment when the idea of Jove was born in the minds of the founders of the nations' (Vico 1984, § 398). Law emerges from a particular consciousness and its conceptual understanding of the world. It is the codification of civil or conventional wisdom which, in the first instance, is an irreal poetic ordering of the world. Such ordering, and the realm of moral life and law, resides in the use of the imagination (Verene 2008, p. 1117). So, the rituals required to placate Jove ask for a particular sacrifice or the repression of egoistic desires on the part of the individual. For the first time, an animal felt a conflict of interests at its heart: there was, on the one hand, the desire to do what further one's immediate interests, yet simultaneously, on the other hand, a motivating fear not to risk the wrath of the gods and perform an action that contradicted one's immediate interests. The sacrifice of a bull is harmful to the interests of a primitive tribe, yet it is performed for the sake of an imaginary reward. It may be impossible that the slaying of an ox will protect the group, but it is a credible belief, especially when shared. The purpose of such a 'credible impossible' is the ability to repress immediate desire and liberate oneself from instinct. Obeying

[7] Vico uses the genetic-etymological method to show that the myths of Jove are then reified into 'ius' or Justice. So, the primal motivation to be good even if it conflicts with other interests is fear of punishment. From such sensuous beginnings, abstract notions of right and wrong are developed through historical embellishments to the irreal system of 'moral' nature. I cannot comment on the status of the etymological proof, though I would take it with a huge pinch of salt. See § 398.

immediate instincts may anger the gods, so one should temper one's desires. The invention of an irreal, symbolic universe creates the means by which pre-humans become humans; by the repression and creation of a libidinal economy for reasons of appeasing the gods, i.e. doing good. So, the imagination explains the emergence of the distinction between reason (rule-following motivated by fear) and immediate instinct rather than assuming it for the purpose of explanation.

But, is such an account consistent with naturalism? Vico grounds his historical science in the faculty of imagination regulated by basic universal facts about human beings. Human imagination creates the basic axioms of practical reason and substantiates them in institutions, a conception of the good and a web of social values and meanings. One can imagine a similar use of the imagination in the understanding of human relations. Just as human beings depend on Jove's good will and their happiness is dependent on his passions and whims, so the sexual bond between two persons is dependent on the whim of the others. To stabilize it and to avoid conflict and anger, partners, children, slaves and social inferiors are conceived of as property in the same way that humans are property of Jove. It is a similar credible impossible: other persons cannot be property in the same way a sky cannot be a person, but it can be believed because it gives 'sense' and has normative consequences for the individual and the group. The basic needs and desires of humans can be satisfied and manifested in myriad ways depending on one's worldview, and the worldview is, according to Vico, an irreal, symbolic creation. The actual manifestation is a product of natural needs, but cannot be reduced to it without losing explanatory power. For Vico, material desires constrain the faculty of imagination which, together, produce the foundations for substantive judgments of reason: the irreality of moral values and obligations cannot make the satisfaction of material needs impossible; otherwise it would not be 'credible'. Although a system of explanation may be 'impossible', it cannot be 'incredible' because its existence relies on the belief invested in it by the group. Beliefs that harm are unsustainable at the group level (or such a group will not survive). Reductive naturalism is too restrictive since it oversimplifies sociology by not supplying the necessary concepts for a full understanding of human behaviour. Humans often sacrifice their own felicity and that of their families for other social goods whether significant or trivial, and unless one can understand the nature of these sacrifices, sociology as a science remains incomplete, The moral conscience is an oscillation between the real needs of the animal and the irreal needs of the social being, thus giving us the foundations for the dichotomy necessary for practical reason.

In Vico's social theory, the humanist influence is keenly felt and no more so in the characterization of the human being as a moral being, not because he renounces aggression nor because his aggression is internalized, but because he or she possesses the capacity of 'imagination'. Once we begin to offer explanations in terms of 'intentions' and 'reasons' then the social world, with all its meanings and conceptual baggage, cannot be discarded and it is this irreal world of meanings which sets the conditions for explanations of actions. In other words, once we begin to talk about agents and their intentions – which any moral theory must do – then we have moved into a very different explanatory domain and cannot reduce such a domain, even if it can be consistent with it, to a mere naturalism (Deligiorgi 2010).

So, for human society and the human itself to emerge out of animalistic savagery, it is a prerequisite that human beings possess the faculty of imagination. The imagination is a necessary prerequisite for free will (the ability to resist immediate motivations) and knowledge (the separation of the perceiver from their field of perception through its symbolic reconstruction) that make social existence possible. Vico delineates three functions of the imagination: '(1) to invent sublime fables suited to the proper understanding, (2) to perturb to excess, with a view to the end proposed: (3) to teach the vulgar to act virtuously, as the poets have taught themselves' (Vico 1984, § 376). One can extrapolate Vico's three functions in a more contemporaneous way (Grassi 1976, pp. 290–291).

1. Poetic creation of the imagination is aimed towards intelligibility. When lightning strikes, an animal's sensory perceptions causes the expected action, i.e. to flee. The pre-human separates the perception from the representation via the faculty of imagination and wonders what if it had been different. By asking why it had not been different, he identifies the credibly impossible cause of the lightning, i.e. Jove's passions. And his new symbolic interpretation – Jove symbolizes his anger through lightning – creates new and different responses: a placating ritual, a prayer, a change in behaviour, a repression of an instinct or a want and so on. The fables that systemize the symbols are an attempt to appropriate the meaning of the world around one. Perceptions are open to interpretation, and imagination allows for the human being to divorce him or herself from the immediate world of causes and reactions (Vico 1984, § 379). The world is no longer an immediate field of sensation, but an object – separate from the 'I' of the perceiver – to be grasped and made intelligible. Such making intelligible is, in the first instance, poetic creation, but is reified into knowledge proper: 'Men at first feel without perceiving, then they perceive with a troubled and agitated spirit, finally they reflect with a clear mind' (Vico 1984, § 218). Radically, Vico is proposing that most of our knowledge is derived from a metaphysical error.
2. Imagination causes fear due to its interpretation of events (Vico 1984, § 382). Yet, because interpretations conflict, this creates an awareness of the ignorance that leads to fear. Systems of thought arise in order to still this fear. Myth is not to be explained as an attempt to explicate the natural order of things and neither does it fail due to metaphorical confusion. Rather, myth is a symbolic expression of the needs and interests of individuals and the fear that arises from our imagined conceptualization of the world is the motivating factor in the formation of ethical motivations.
3. Vico 'conceived of ethics as an expression of the transfer of our instincts and passions, that is, of the nature in us, into a human system of relations priorly designed by our imagination and forming the roots and foundations of our ethical motivation' (Grassi 1976, p. 292). Basic needs, hunger and reproduction, are satisfied not just directly but also indirectly. Work is carried out not just for

the satisfaction of needs but also the satisfaction of pseudo-needs. Materially, iron may be more valuable than gold given its myriad uses, but socially gold is more valuable than iron because symbolically it expresses wealth and power. Agents desire the social meaning often in violation of material needs. Fear and other desires created by the imagination can be appeased through symbolic exchange. The individual farms for himself in order to satisfy material needs, but he also farms for the priest who will convey his thanks to the gods and thus satisfy his symbolic needs. The farmer acts from reasons derived solely from the symbolic world. The lightning will not destroy him because he has performed his duty. Hence imagination makes rules and laws over and above (and sometimes contrary to) simple animal nature.

18.4 Imagination and Practical Reason

The centrality of imagination in Vico's social theory, as a prerequisite to the formation of belief, knowledge and law, has the very desirable trait of explaining the fact–value distinction. Whereas both Locke and Hobbes wish to deduce obligations from facts of human nature, Vico only needs to show that a faculty of imagination is an empirical property of certain minds. The obligations of the social world and moral law are irreal, in the sense that they have no existence outside the group or community, but are produced by a fact about a natural object (the faculty of imagination of the human being) and material desires (the credible impossibles must be consistent with basic needs). The emergence of a moral society is explained by the fear humans have when they reconceptualize the world as a 'credible impossible' and they placate such fear through the performance of expressivist rituals. Standard attempts to describe the origin of morality identify a specific characteristic that entails the need for moral law (aggression, egoism) or explains its pre-social existence (cooperation, compassion). Vico offers us an alternative, explained in this counterfactual, if human beings did not possess imagination, then they would have no knowledge nor morality.

How, though, is one to make sense of this in a way coherent with modern science and acceptable to contemporary public reason and why might we want to move from a consideration of concrete drives and passions such as aggression and compassion to a faculty of thinking in human beings? Surely the former is easier in psychological and evolutionary biological terms and the task of explaining the emergence of the imaginative faculty in evolutionary terms is not helped by the vagueness of the term. The first step is to identify a psychological conceptualization of the what-if structure of thinking. Such a task is an empirical and not a theoretical one and a psychological account needs to be supplied by a cognitive mind theorist to back up a rather theoretical account of the imagination. There would also have to be a historical investigation. Vico's philological-genetic inquiry into the origins of human civilization and the fables of a divine gift or divine retribution (the flood for example) identify the beginnings of human society for him and the birth of moral law. Such a birth must be simultaneous with the emergence of a language and accessible through the hermeneutic reconstruction of myths. For the theoretical side, it is enough to say that there is now consciousness and there once was not. Consciousness must have emerged.

More importantly, consciousness is simultaneous with imagination. A personal anecdote may illuminate this. I recently helped my 6-year-old son with his mathematics homework. One of the questions I initially found rather odd: it asked him to compare two bags containing six balls in total. One distribution was 4 in one bag and 2 in the other. The other distribution was 3 and 3. What was peculiar was the actual question asked. He was required to identify the 'fair' distribution. Fairness is not identical with justice, but neither is it a fact. It was not explained to the student and I did not have to explain it to him. My son immediately pointed to the 3 and 3 distribution. He *perceived* the fair situation as desirable and wanted it more than the world in which he could choose whichever bag. A 'fair' world was a motivation.

There is no fact of fairness, only a symbolic comprehension of the situation. Let us try to reconstruct the way one might think through this problem. The mind of the child asks in such a world where I had to share marbles with my little brother, I could imagine that the distributions could be anything but that which I would most want is the one in which there were equal amounts in each bag.[8] Equality is not fairness unless conceived as such by an act of imagination which brings together the passions of egoism, sympathy, empathy, care, ambition and so on. (One could imagine a real case where a child already owns a hundred marbles and therefore thought it 'fair' that another had all six marbles and he had none.) Theoretically, imagination is the capacity to apply the structure of 'what if?' to the world in order to conceptualize it as lacking something. I imagine different worlds in which I and my smaller brother possess 2, 3 or 4 marbles. I imagine the passions I would feel in each and also the passions of my brother. All such reasoning is immediate and occurs at a very deep level of human understanding, at the very break between perception and representation; what if I were to represent the world as thus (as Jove in the sky and as possessing 4 marbles). Each world would have a mood to characterize it: despair if I were only to have 2; arrogance if I were to have 4, perhaps tinged with pity; and righteousness if I were to have 3. And such worlds would fill me with dread.

Human being is the imposition of symbolic meanings that constitute values over and above simple desire-based relationships to objects: the fear of divine retribution in primitive humans allows them to exercise free will and repress strong, immediate motivations. The symbolic property relationship between members of a group, such as 'this is my son', 'my wife' and so on, is not true in any metaphysical sense, but it is an extension of the thinking that tells me that this is my body and that the Earth belongs to Jove, Whether it is true or false, it has pragmatic consequences: respect for another's partner in a group if it is conceived of as property represses desires that

[8] One problem here is that my child is already socialized. He has completed 1 year of school and possesses a language. He has been taught about 'sharing' and 'fairness' since his birth.

can destabilize the whole. If I have fear for retribution from the gods, it is a stronger passion than the lust I feel for your partner.

Is this the sort of imagination Vico was describing? Perhaps, but his imagination is the way in which we understand human beings as rule-followers rather than mere responses to mechanical chains. Imagination is the invention of a case whereby we ask the question What if the world were thus and so? (which, it could be argued, we perceive simultaneously with how the world actually is) and it requires the conceptualization of the world into a symbolic, irreal place of reasons: we ought to do X and not satisfy desire Y because the gods wish for us to be Z and so on.[9] Morality depends upon the construction of an 'irreal' world (a credible impossible). So, an ethics derived from this Vichian position is an irrealism, but not the simple expression of animalistic desires (egoism, compassion and so on), but more conventions such that statements can be rational or not as long as they cohere to a system which is credible even if it is impossible.

18.5 Some Consequences of This Way of Thinking

The advantages of the approach firstly overcome the deficiencies of our standard stories about the origin of morality. A single drive to power or egoism cannot explain a contractual birth to morality unless it assumes reason apart from passions. Reciprocally, a rational agent has no need for the institution of morality unless his passions can overrule his reason at times. And so both rely on a traditional dichotomy between reason and passions, but animals do not have reason and human beings do, so the true question is: How did reason emerge? A prerequisite for reason would be the imagination as conceived by Vico:

> Men cannot unite in a human society unless they share a human sense that there is a divinity who sees into the depths of their hearts, since a society of men can neither begin nor remain stable without a means whereby some rely upon the promises of others and are satisfied by their assertions in secret matters. (Vico 1982, § 45)

Hobbes cannot explain why egoists would keep the meta-agreement (the agreement to agree) and Locke cannot explain why moral beings would need to make an agreement at all; Vico can. The birth of society for him is non-rational, it requires a leap of belief from a world as it appears, to an arcane world of reasons and values (firstly understood as animism). Reason is made possible by the symbolic, poetic expression of human understanding. Animalistic humans require laws because in groups conflict will harm their interests and they will agree to obey the shared law because they have self-imposed a fear of retribution on themselves. The fiction of the credible impossible world of retribution should be thus and so can only be sustained by a symbolic system of motivation: eternal life, reward and justice, conscience

[9] It is very similar to Sartre's (2003) discussions of self-consciousness as well.

and so on, that overrides our immediate motivations. The faculty of the imagination could explain the origin of morality and given its irrealist foundations, it would not normatively beg the question in favour of a specific system of morality or exclude one. It should also remain coherent with naturalism as long as the imagination can be identified as an emerging characteristic of an entity and not be reduced to natural dispositions.

Finally, a word or two should be devoted to consequences. Ethics, on this picture, is best understood as irrealism: it is not a fact that ‘lying is wrong’, but neither is it a simple emotivism. ‘Lying is wrong’ does not stand in for a simple con-attitude on the behalf of an agent, it expresses an imagined fear of the consequences and hence a motivation to be good. However, since the fear derives from a system shared with other group members, statements can be rationalized, which internally is seen as logical, but externally is akin to changing one’s attitudes, in much the same way that *ad hominem* arguments are used in philosophy. If I want to convince you to stop smoking, I would argue that it is harmful. That would be rational. But, equally, I could show you that George Bush smokes and assume you do not want to be like him. The reasoning applied in moral reasoning is internally akin to the former, but externally akin to the latter. Moral statements are simultaneously minimally rational and also motivating.

Individual action, then, can only be understood in a social context in which the observer predicts, on the basis of the social identity of the individual, what he is most likely to do. There can be no reduction to a universal human nature, except in its most abstract form. To understand the individual, it is first necessary to understand his social commitments.

References

Deligiorgi, K. 2010. Doing without agency: Hegel’s social theory of action. In *Hegel on action*, ed. A. Laitinen and C. Sandis, 97–118. London: Palgrave MacMillan.

Freud, S. 2002. *Madness and its discontents.* Trans. D. McLintock. London: Penguin.

Grassi, E. 1976. Marxism, humanism, and the problem of imagination in Vico’s work. In *Giambattista Vico’s science of humanity*, ed. G. Tagliacozzo and D. Verene, 275–294. Baltimore: John Hopkins Press.

Hobbes, T. 1996. *Leviathan*. Cambridge: Cambridge University Press.

Jackson, F. 1997. *From metaphysics to ethics: A defence of conceptual analysis*. Oxford: Oxford University Press.

Kano, T., and E. Vineberg. 1992. *The last ape: Pygmy chimpanzee behavior and ecology*. Stanford: Stanford University Press.

Locke, J. 1988. *Two treatises of government*. Cambridge: Cambridge University Press.

Peterson, D., and R. Wrangham. 1997. *Demonic males: Apes and the origins of human violence*. New York: Mariner Books.

Railton, P. 2003. *Facts, values, and norms: Essays toward a morality of consequence*. Cambridge: Cambridge University Press.

Rousseau, J.-J. 1984. *A discourse on inequality.* Trans. M. Cranston. London: Penguin.

Sartre, J.-P. 2003. *Being and nothingness.* Trans. H. Barnes. London: Routledge.

Verene, D. 2008. Vichian moral philosophy: Prudence as jurisprudence. *The Chicago-Kent Law Review* 83(3): 1107–1130.

Vico, G. 1982. The first new science: The principles of a new science of the nature of nations leading to the discovery of the principles of a new system of the natural law of the gentes [1725]. In *Vico's selected writings* (trans and ed.: Pompa, L), 81–156. Cambridge: Cambridge University Press.

Vico, G. 1984. *The new science of Giambattista Vico*, 3rd ed, ed. T. Bergin and M. Fisch. New York: Cornell University Press.

Chapter 19
Enacting the Moral; Concrete Particularity and Subjunctive Space

Adam B. Seligman

Hazarding an opinion on the source of human morality is not something to be taken lightly.

I recall literally running out in 1999 to purchase Tzevan Todorov's book, *The Fragility of Goodness*, which dealt with the saving of Bulgarian Jews during WWII by leaders of the Church and some parliamentary deputies (1999).[1] I ran out to purchase the book, to learn why church leaders such as Patriarch Kiril and Exarch Stefan exposed themselves and their families to considerable risk to save the Bulgarian Jews. Indeed, to this day, in early March the leaders of the Jewish community go to the Bachkovo monastery to recite Kaddish for these men.

I approached the book with great expectations of learning what made for moral courage and was ultimately disappointed. Not that it is a bad book; it is a very good book. But there is no explanation of what makes us moral; there is only a narrative history of this or that individual acting in this way and not in another. Individual decisions were taken, decisions that had huge repercussions on the lives of tens of thousands of individuals, decisions which allowed 50,000 Jews to live – and there is no, I have come to believe, there can be no explanation, no theory, no algorithm for why a particular individual decided at a particular point in time to act in one way and not the other. We can impute whatever theory we wish, but ultimately the decision made is akin to a black box that we cannot fathom. This of course assumes that a decision was made, that rational calculations, cost–benefit analyses, utility margins and so on were brought to bear. They may have – but equally may not – have been aspects of these acts of moral courage.

[1] Though we must recall that 11,343 Greek, Macedonian and Thracian Jews who were then under Bulgarian administrative rule, were not saved but were deported and except for 12, all were murdered in concentration camps (Todorov 1999, p. 9).

A.B. Seligman (✉)
Department of Religion, Boston University, Boston, MA, USA
e-mail: seligman@bu.edu

B. Musschenga and A. van Harskamp (eds.), *What Makes Us Moral? On the capacities and conditions for being moral*, Library of Ethics and Applied Philosophy 31,
DOI 10.1007/978-94-007-6343-2_19, © Springer Science+Business Media Dordrecht 2013

19.1 Moral Action

I think though that from this black box we can at least take a first tentative step in our inquiry and so I suggest that we focus our inquiry on the realm of actions: What impels us to act in a manner that we or others deem as moral? It is the actions that I wish to stress rather than any inquiry into particular human qualities that we may or may not possess. For ultimately, I think it is our actions that may be judged as moral or immoral. What we *are*, on the other hand, is anyone's guess. Of course, this circumscription of the moral to actions rather than essences still leaves us in the dark about what moral, or morality itself may mean. And answering that question, a necessary prolegomena to the theme of this volume, is, itself, no easy task. A pretty fair review of some of these problems can be found in Plato's dialogue *Euthyphro*, which deals with the not dissimilar problem of piety and impiety (Plato, pp. 57–88). In rather unique circumstances, to which I will return shortly, Socrates attempts to get Euthyphro to define what makes a pious (and in contrast) an impious act. Some initial, if illusionary progress is made when the pious is identified with what is pleasing to the gods; though it soon becomes clear that different gods will define different acts as pious or impious and I think from our own experience, we could assert that the same god may well deem the same act pious or impious depending on the circumstances.

Indeed the history of morality has shown how the definition of moral acts has changed radically over time. We need think only of its history in the Western tradition. For Roman writers such as Quintillian, Seneca and Cicero (and not a few of their Renaissance followers) the moral life was identified with service to one's political community. By the middle of the sixteenth century this was best represented in the tradition of *republican humanism* that we have learned to identify through the writings of scholars such as Quentin Skinner (2002) and J.G.A. Pocock (1975).

The religious wars of mid-century as well as later seventeenth-century developments, during the English civil wars and in the settlement of New England, saw morality identified with following Godly commandments in the establishment of a 'city upon the hill' (Miller 1959; Bercovitz 1975). By the turn of the eighteenth century and as was to become almost iconically realized in the writings of Immanuel Kant, morality moved inward and was identified with following the call of conscience, the famous 'moral law' within. Different gods, then – and with them very different understandings of morality.

19.2 Morality and Intentionality

In this move inward – towards identifying morality with the workings of the inner self, of conscience – and ultimately, with Kant, of intentionality: we find as well an increasing distancing of morality from actual acts in the world. Morality begins to leave the messy, ambiguous, complicated and ever-changing world of action to take up residence in the relatively secure confines of our mind. We leave the workings of *fortune*, of circumstance – never totally under our control – and circumscribe the realm of the moral to that of moral reasoning, over which – presumably – we have control. We leave the dirt of history, the compromised nature of all particular circumstances and persons, and enter a realm of greater moral clarity and control. This concern with intent has become the touchstone of much of our moral reasoning, most importantly, as noted, in Immanuel Kant's writings on the workings of the 'good will.' As Kant stresses: 'The good will is not good because of what it effects or accomplishes or because of its adequacy to achieve some proposed end; it is good only because of its willing, i.e. because it is good of itself' (Kant 1785/1989, p. 5). As some of you may know, Thomas Nagel and Bernard Williams cogently delineated the limits of this view by showing the insufficiency of a morality that totally eschews consequences and the external world of action in making moral calculations (Williams 1981, pp. 20–40). As Nagel clarified: 'However jewel-like the good will may be in its own right there is a morally significant difference between rescuing someone from a burning building and dropping him from a twelfth-story window while trying to rescue him' (Nagel 1979, p. 25).

I am neither learned enough, nor particularly interested here in going into the important differences between deontological ethics and consequentialist ones that such comment evokes. In fact, personally, I am pretty firmly rooted in a tradition that has some difficulty in recognizing a realm of ethics or morality outside of law and the duties law imposes. There is a crucial expression in Rabbinic Judaism – *tov hamitzuveh*. That is to say, the commanded is always preferred to any act of supererogation. Indeed, it is unclear, and at very least, debatable if in Judaism any ethical value can be assigned to actions not understood as commanded by God.

But the issue of moral luck – which is what is at stake in Nagel and Williams's critique of Kant – does bring us back to the world of action, of events in the world out there – rather than internal states of consciousness and intent. This world of action is a hazardous one; where things often go wrong, good intentions lead to bad results and the mass of concrete particulars: specific acts, agents, events, etc. continually call in question the coherence of any universal set of laws or injunctions that, in their very abstract universality, are always more comfortable in the world of consciousness than in that of praxis.

19.3 Morality and the Particular

It is precisely this concrete particularity of action that I wish to address. For action is, by definition, always concrete: we marry this spouse, attend the funeral of this colleague, give money to this cause, shy away from helping this person in distress, and so on. As concrete and particular, action is always messy, never totally coherent, often ambiguous, given to second guessing and endless improvement. It can never be abstracted from the flow of history, from the world of contingency and of forces beyond our control. It is the polar opposite of the good will – which depends only on itself to be moral and praiseworthy. Action depends on myriad forces, not only beyond our control, but often even beyond our understanding.

And it is in this context that the setting of Plato's *Euthyphro* becomes significant for our inquiry. For when Socrates meets Euthyphro, the latter is on the way to court, to try his father for killing a slave, who had, in turn killed a field labourer. Euthyphro's father had bound the evil-doer in chains and thrown him in a ditch where he expired as the father awaited word of what to do with him. Euthyphro was appalled by his father's actions and reasoned that his father, having murdered this man, must be brought to court – as must any man who had taken another's life.

It is just this very reasoning, in its abstract universality, that shocks Socrates who questions the propriety of bringing one's father to court in any circumstance. It is this then that occasions the unresolved debate between them on the definition of piety and impiety, on what we owe the gods and if the good is good because the gods deem it so, or if the gods deem it so, because it is good in itself. Something similar to the first position would much later be advocated by Islamic Ashaarite philosophy (against Mutazalite thought) as well, as in the Christian tradition by Dun Scotus and William of Occam – who indeed claimed that if God had so willed, lying, stealing and committing adultery would be virtuous activities (Reinhart 1997).[2] St. Thomas of course advocated a much more compatible unity of nature and grace.

Socrates questions the very generality and universality of Euthyphro's position – precisely what most of us would consider its moral foundation – and argues instead for the claims of the particular, in this case one's special obligations to one's father. Abstract knowledge of piety and impiety, of the nature of right and truth, good and evil, just and unjust, honourable and dishonourable action, is – Socrates argues – quite beyond us and appeals to their truth-value do little but get us into interminable muddles, not very different indeed from the gods themselves.

What we do know are our immediate obligations: and these would seem to prohibit taking one's father to court (unless, Socrates implies, the murdered was another relative). Socrates would in fact seem to be mirroring the dictum of another philosopher, living at roughly the same time, but worlds away: the Confucian scholar Xun Zi, who claimed that: 'To honour the beginning is the basis of virtue' (Hsun 1963, p. 91). Whatever else, prosecuting one's father is not to honour one's beginnings.

Beginnings however are always not only concrete, but particular and limited – this father, this mother, this place of birth and time, and not others. And as opposed to our thoughts, our actions are as particular as they are messy. Hence, if morality is to be understood as a predicate of action, rather than of thought or intention, I think then its first prerequisite is that we have the capacity to abide with the ambiguity, impermanence and fallibility that inheres to all action.

What we also learn from the *Euthyphro* is that our moral goals should be equally limited, bound by the very circumscribed nature of our knowledge. We must reject

[2] On Mutazalite thought see: Martin et al. (1997), Makdisi (1991), pp. 50ff; Tierney (2008), p. 96, Stone (1965).

the temptation of the general and all-consuming moral directives, say the lure of history's *telos*, and rather direct or moral actions to the realm of the concrete and limited, our filial obligations say – our debt to our beginnings.

19.4 Ambiguity, Boundaries and Moral Action

By focusing on the concrete and particular I mean less to invoke the type of criticism levelled by Michael Sandel on the universalism of liberal ethical claims (Sandel 1989) and more to recall John Dewey, the father of American pragmatism, who argues for the need to come to terms with the ambiguity and uncertainty that characterizes all practical deeds. He eschews the philosophical tendency to seek any sort of pure Being, any 'disclosure of the Real in itself, of Being in and of itself' (Dewey 1930, p. 17). 'The quest for certainty', he exhorts, seeks 'a peace which is assured, an object which is unqualified by risk and the shadow of fear which action casts' (Dewey 1930, p. 8). Unfortunately, practical activities 'take effect in an uncertain future, and involve peril, the risk of misadventure, frustration and failure' (Dewey 1930, p. 21). Embedded in a world of particularities and change, of probabilities and contingencies, the world of real life decisions is inherently uncertain. Actionable knowledge, rather than theoretical speculation, requires recognizing the necessarily ambiguous nature of the world, as well as of our very categories for knowing it.

Accommodating this ambiguity is, I would argue, a necessary, if perhaps not sufficient perquisite for morally informed acts. Its necessity can be most easily accessed through the very idea of empathy; that imaginative blurring of boundaries that makes social life possible. Like allegory and metaphor, social empathy works by crossing and blurring the boundaries of ego and alter. Just as we can draw verbal analogies ('My love is like a red, red rose'), so we cross the boundaries of affect – his pain may be like my pain. For Aristotle of course, this was the very role of tragedy: to generate empathy and expand the boundaries of the self. Like metaphor, empathy is characterized by the positing of analogies and similarities between otherwise distinct entities. And while I may not be able to actually feel her pain, or participate in his jealousy or share their loss, I can make the imaginative jump from my own experiences of pain, jealousy and loss to empathize with others in their distress (and in their joy).

The social world – the world of quotidian affairs, of families and friends, of citizenship and community, of petty rivalries and dramatic jealousies, of gossip by the water-cooler and baseball with friends, barbeques with in-laws, Christmas dinners and college jaunts – all belong in no small measure to the world of empathy, as it is realized, extended, and circumscribed as the case may be. This is the level of metaphor, of my ability to reach across the boundary between self and other, to apprehend your feelings, at least in part, through the imaginative leap from my own. Empathy thus rests on some elision of self, the willingness to 'play' with fuzzy boundaries. It requires some renunciation of ego's central place in its own symbolic universe, even if momentarily. It rests on some blurring of boundaries and our capacity for metaphor; and is of a dual character. First, it renounces the self's limiting and defining conditions by requiring a willingness to play with the boundaries of the self. Second, it renounces immediate gratification by accepting social rules of order, in the very act of sitting quietly and listening to the other, of trying to avoid putting one's own ego first.

Empathy and – to use psychoanalytic terms – proper ego-functioning both demand recognition of the world as it stands outside the self, and so would seem to go hand in hand. Both involve recognizing limits, but also the capacity to transcend them (if sometimes only in the imagination). They fuse the edges but do not confuse the boundaries of objective and subjective realities. They define our abilities to empathize with others but also to recognize objective reality outside of ourselves. Order's constraints and boundaries, its rules and regulations, bind the ego and circumscribe its action. Yet, however paradoxically, these constraints are necessary if either social order or empathy is to exist. I would add in passing that the contemporary post-modern embrace of ambiguity often loses sight of this dynamic, of the need for the conventional rules and regulations that constitute the boundaries that separate but also connect our shared world. Boundaries may be blurred (or ambiguous), but societies can neither renounce nor absolutize them. Empathy and metaphor rest on the ability to play the edges, retaining some sense of separation and distinction between entities or selves, and yet not create rigid boundaries that forever keep subject and object, world and feeling (and fantasy) apart and inviolate.

If the preceding argument is correct, empathy – and with it, moral action – must rest on a dialectic between boundaries and their dissolution that we are identifying with the sense of the ambiguous. Empathy lies in achieving this balance, rather than in the denial of boundaries or of some ideal free of limits and constraints, a totalizing vision of self, other and society. But neither does the balance come by absolutizing boundaries, by denying the creative need to continually reframe them. It needs the creation and recreation of boundaries, but not their abrogation.[3]

At this point it may be wise to review the ground we have covered in a number of points:

(a) morality is an attribute of actions rather than essences;
(b) its definitions have changed greatly over time, cultures and circumstances;
(c) a necessary component of moral action would seem to be the capacity to empathize with others;
(d) another component would be a limited range of action, cognizant of our limited knowledge;
(e) both rest on our ability to live with ambiguity, with the ever-shifting frames of our reality and interactions with others.

A minimalist position perhaps, one from which many would not demur, but one which may well seem and too individualistic, if not psychological in its implications:

[3] This argument is developed in greater depth in Seligman and Weller (2012).

What makes us moral? Good childrearing practices, informed by the insights of D.W. Winnicott (1971) and other child-analysts. So let me throw a spanner in the works and make the totally counter- intuitive claim that one crucial resource in building a capacity for ambiguity, for leaving a space for the other and so for empathy, is … ritual.

19.5 Ritual and Ambiguity

By ritual, I understand a series of formal, iterated acts or performances that are, in Roy Rappaport's terms, 'not entirely encoded by the performer' (Rappaport 1999, p. 24). That is, they are imbued by meanings and shaped by conventions external to the performer. Such ritual acts are crucial to the existence of the relational self, i.e. of a self who can accommodate ambiguity and engage morally with the other. This is the result of ritual's ability to both recognize and cross boundaries, and of its rhythmic relationship to time.

Rituals create a subjunctive space, a shared 'could be' that constructs individuals in relation to others.[4] This is as true of religious ritual as it is of the rules of civility and etiquette. Ritual, in its formal, iterated and enacted moments, presents a unique human resource for dealing with ambiguity and the multivocal nature of all relationships – with beings human and divine. Ritual defines and binds entities, times and spaces. By creating such borders it also links entities, times and spaces to what lies beyond their immediate field. As such, it presents a coherent and embracing way to live in a plural and hence also deeply ambiguous universe, one where order's rules can never really be known, but still must be acted upon.

When we say that people share a symbol system, or a set of values, or a common idea of the sacred, we in essence assert that they share the potential space of what 'could be', a subjunctive world. Much ritual action provides this shared sense of empathy – sometimes even in terms of a shared 'what if'. When Jews congregate around the Passover Seder table and are enjoined to fulfil the commandment to feel 'as if you yourselves have been liberated from Egypt' they create that shared space where the communality of the 'could be' becomes the basis of the ongoing collective experience. The Shi'ite enactment of the defeat of Imam Hussein at Karbala and the Catholic participation in the Eucharist have similar import. Confucius, famously uninterested in the world of spirits, still insisted that when 'he offered sacrifice to his ancestors he felt *as if* his ancestral spirits were actually present. When he offered sacrifice to other spiritual beings, he felt *as if* they were actually present.' Maimonides enjoins us to attend to our prayers 'as if' we are standing before the Creator of the universe. The moral community that Emile Durkheim outlined in his *The Elementary Forms of Religious Life* exists precisely because it shares the potential space of culture created through ritual. Shared moral

[4] The following argument is taken from A. Seligman et al. (2008).

community is however never the entirety of social experience in its full complexity of misunderstandings, conflicts of interest, and incompatibilities. It is instead a subjunctive construct, a shared acquiescence to convention; precisely the convention of ritual.

In ritual we subject ourselves to externally given categories of order, whose source can be anything from a transcendent deity (as in Judaism) to the natural ordering of the physical and social world (as in Confucianism). Ritual concentrates on the performative nature of the act rather than on its denotative meaning. In its purely formal aspect, ritual puts questions of belief or truth aside in favour of the shared world that its action creates and requires. The external, performative aspects of ritual – especially its repetition and recollection of places and times not given to purely rational or instrumental computation – give it a unique lability. Thus ritual encompasses the ambiguity of life in a unique manner. It allows one to 'play' with such ambiguity without undue concern with the authenticity of one's actions and beliefs. Ritual unshackles the mind from a need to *believe* in a dogma of our choosing, as long as we act within its conventions.

Ritual thus allows us to live with ambiguity and the lack of full understanding. In slightly different terms it allows us to live with the other, with what we do not fully know or understand – as indeed, we can never fully know or understand any other. The presentation of ritual's 'as if' universe, the subjunctive, requires neither a prior act of understanding nor a clearing away of conceptual ambiguity. Performance simply and elegantly side-tracks the problem of shared meaning by allowing participants to express acceptance of an order without requiring a full understanding of it. In this way it resembles all manner of decisions we must make to take any concrete action, where we accept that we have as much understanding as we are likely to get and that action must be taken even though our knowledge is incomplete (as it always must be). This is true for a medical intervention, a financial investment, a marriage commitment, a declaration of war or the planning of a highway – for virtually all forms of human endeavour. Through its emphasis on action, on the performative and its creation of a subjunctive universe, ritual creates a world – temporary, fragile to be sure, but not false – a world where differences can be accommodated, tolerance enacted (if not fully understood) and openness to the other maintained. By creating subjunctive worlds, ritual reminds us that otherness is possible, that the world as it appears is not the only possible world.

19.6 Ritual and Sincerity

While ritual activity carries its own form of intentionality, it is important to note that ritual is not necessarily concerned with what we often call sincerity. In any ritual, as with saying 'please' and 'thank you', performing the act marks acceptance of the convention. It does not matter how you may feel about the convention, if you identify with it or not. In doing a ritual the whole issue of our internal states is often irrelevant. What you *are* is what you *are in the doing*, which is of course an external act. This differs significantly from modernist concerns with sincerity and authenticity. Getting it *right* is not, as in the latter cases, a matter of making outer acts conform to inner beliefs. Getting it right is doing it again and again and again – it is an act of world construction. As an ideal type, the self *who does ritual* is very different from the self *who is sincere.*

Unlike ritual, the sincere involves a search for motives and for purity of motives. Sincerity privileges intent over action. As noted earlier, this concern with intent has become the touchstone of much of our moral reasoning. In contrast to this, the realization that our boundaries are only artifice and that the world is fundamentally ambiguous – which we can find in a ritual approach – allows us to accept and even play with that ambiguity. Sincerity seems by its very definition to exclude ambiguity. Recall that its dictionary meanings include 'being without admixture,' 'free', 'pure', 'whole', and 'complete'. Samuel Johnson lists among its cognates 'unhurt', 'uninjured', 'pure', 'unmingled' and 'incorrupt'. Sincerity, carried to its extreme, is the search for wholeness, for overcoming boundaries and positing a unitary, undifferentiated, uncorrupted reality. It is a utopian impulse, which – in its search for purity –often sacrifices empathy for narcissism.

Anti-ritualist attitudes deny the value to ritual's subjunctive, of play, convention, and of illusion. They seek to root interaction in some attestation to the sincerity or truth-value of all categories or interlocutors. The drive for oneness, for notational wholeness whether expressed in the wish to be at one with oneself and with the world, or for eternal and unchanging truths, will always come into conflict with the ambiguous reality of our existence.

To the extent that it is devoid of content, the very formalism of ritual (and this is, admittedly, never complete) allows room for multiple meanings. 'Empty ritual' may well be a modernist execration, but this seemingly 'empty' (but really formal) aspect of ritual acts is precisely what allows us to move from one world of significance and meaning to another. Rather than vilifying 'emptiness' we should see the formal qualities of ritual as generating a potential space within which otherness can be introduced and experienced. This is true for the frame around a picture, for the crossing of oneself before receiving communion, as well as for the rituals of courtship as a prelude to love. All enable new experience through the construction of boundaries whose very creation allows us to bridge them. Frame and bridge are one in the formalism of these acts and artefacts. This is a significant and often ignored aspect of ritual – by delineating a separate and sacred time and space, ritual also forces us to continually cross-over, into and out of every ritual performance. In essence then a (even if often latent) recognition of multiple realities is structurally built into ritual action.

Certainly the modern world has seen different attempts to construct totalistic and intensely 'sincere' rituals of political belonging (among Nazis, fascists, communists, but in certain liberal democratic societies as well). However, by defining the sacred in immanent rather than transcendent terms, the ritual recognition of the world's inherent pluralism tends to get lost in their 'expressive' aspect. Often indeed modern religious practice takes on these 'expressivist' elements among more fundamentalist (and hence, by definition, modernist) Jews, Moslems, Hindus and Christians. Their ritual activity tends to be about a self-conscious, almost Herderian expression of identity, in a way that traditional ritual practice is not. When ritual becomes expressive of a worldview, rather than a form of subjunctive action – which by its very nature is fragile, particular and cognizant of its own limits – recognition of difference tends to get lost, as the spaces between sacred and profane, ritual and quotidian, self and other are conflated in what is, at the end of the day, a hymn to self.

19.7 Ritual, Ambiguity and Moral Action

For human creation – the greatest of which, are our moral actions – takes place in the spaces in-between: in the not-yet represented, the as-yet undefined, and un-notated. Only here lie the possibilities for the emergence of the new or other. The formal or abstract character of ritual creates that space by positing subjunctive worlds that open up a potential space. Such a space opens up the possibility for something new to come to be, something that was not previously represented, symbolized or imbued with meanings – something that was not previously notated.

Each moment of ordering – daily prayer, purification rite, birthday party – stops the continuum of daily experience, if only for a moment. It reveals and creates a new subjunctive world, even though we know we will leave that world again as soon as we leave off our prayers, re-enter the world of inevitable pollution, or get up on the morning after a birthday. In music, sound orders time. It is abstracted as the beat, an instantaneous stoppage, a geometrical point in time instead of space. Ritual is far more complex than this, but it also fixes the world for a moment and then allows the flow to resume. By being predictable, rhythm defines time itself. We tend to think of rituals just as a series of events or of rhythm as a simple set of repeated claps or the ticking of a clock:

tick tick tick

It might be better, however, to emphasize as strongly the bits in between the claps:

tick TOCK tick TOCK tic TOCK

That is, the ritualized, ordering moments are not the only things that matter. They are just one of at least two subjunctive worlds that we always cross between. Rhythm is not just tick, tick, tick, but rather tick-tock, tick-tock, tick-tock. In between the beats lies the continuum of life, the potential chaos, the un-interpretable silence.

This was Freud's fundamental insight in his classic story of the little boy tossing a toy far away to shouts of '*fort*' (gone) and getting it back with a happy '*da*' (there). Freud recognized that this game of repeated disappearance and return allowed the child to conceptualize and thus accommodate himself to his mother's departures and returns. As Freud put it, 'her departure had to be enacted as a necessary preliminary to her joyful return' (Freud 1955, p. 15). The absence was as necessary as the presence. The child's repetitive play allowed him the ability to view the whole pattern from above, to realize that the absence would lead to the presence. Repetition provides the armature upon which both tick and tock, absence and presence, can be re-presented. Repetition always creates difference, both ticks and tocks, and requires us to accept that difference. Like ritual, it opens up the potential space beyond the self where the other may exist.

Ritual, then, can help teach us the tremendous dangers of trying to build a totally coherent world of authentic, individual truth-claims. It encourages us to recognize the fragmented and discontinuous nature of the world, the endless work entailed in building and refining our multiple and often conflicting relationships within that world, and the ultimate impossibility of resolving its ambiguities. And it helps teach us the powers of ethical action based upon such a vision. Accepting the world's discontinuities and ambiguities means that the work of building and refining relationships will never end. Ritual, at least in its relationship to the rest of experience, is never totally coherent and never complete. Yet doing the work of ritual is one of the most important ways we live in such an inherently plural world.

Ritual lets us be together despite the openness of all our encounters. Its formalism and its repetitions provide a critical vehicle for human sociability while not denying the ambiguity that is part and parcel of sociability itself. Ritual factors in ambiguity, as pointed out above in the example of the tock that is inherent in every tick. If it were truly biologically or neurologically impossible to step outside of roles and social expectations there would be no need for civility, courtesy or the rules of etiquette. The rules apply precisely because such possibilities are always with us – sometimes threatening, sometimes beckoning, sometimes both at once.

Ritual's rhythms provide predictability and so create an arena where people can interact and share a moral universe. If for St. Paul, the ritual law was understood as a necessary propaedeutic to the new dispensation of grace; it may well be that ritual's obligations, with their ability to both order existence and open it, is itself a central prolegomena to moral action. I wouldn't wish to claim that ritual is the only road to moral action – but it is a much neglected one in today's world, and one whose virtues we would do well to be reminded of.

References

Bercovitz, S. 1975. *The Puritan origins of the American self.* New Haven: Yale University Press.

Dewey, J. 1930. *The quest for certainty: A study of the relation of knowledge and action.* New York: Milton Balach.

Freud, S. 1955. Beyond the pleasure principle. In *Standard edition of the complete psychological writings of Sigmund Freud*, vol. 18. New York: W.W. Norton.

Hsun, T. 1963. *Basic writings.* Trans. B. Watson. New York: Columbia University Press.

Kant, I. 1989. *Foundations of the metaphysics of morals.* Trans. L. White Beck. Upper Saddle River: Prentice Hall.

Makdisi, G. 1991. *Religion, law and learning in classical Islam.* Hampshire: Variorum.

Martin, R., et al. 1997. *Defenders of reason in Islam.* Oxford: Oneworld Publishers.

Miller, P. 1959. *Orthodoxy in Massachusetts.* Boston: Beacon.

Nagel, T. 1979. *Moral questions.* Cambridge: Cambridge University Press.

Plato, Euthyphro. n.d. In *The works of Plato*. Trans. B. Jowett, 57–88. New York: Tudor Publishing Co.

Pocock, J.G.A. 1975. *The Machiavellian moment: Florentine political thought and the Atlantic republican tradition*. Princeton: Princeton University Press.

Rappaport, R. 1999. *Ritual and religion in the making of humanity*. Cambridge: Cambridge University Press.

Reinhart, K. 1997. *Before revelation: The boundaries of Muslim moral thought*. Albany: SUNY Press.

Sandel, M. 1989. *Liberalism and the limits of justice*. Cambridge: Cambridge University Press.

Seligman, A., and R. Weller. 2012. *Rethinking pluralism: Ritual, experience and ambiguity*. Oxford: Oxford University Press.

Seligman, A., et al. 2008. *Ritual and its consequences: An essay on the limits of sincerity*. Oxford: Oxford University Press.

Skinner, Q. 2002. *Visions of politics; Renaissance virtues*. Cambridge: Cambridge University Press.

Stone, J. 1965. *Human law and human justice*. Stanford: Stanford University Press.

Tierney, B. 2008. Natural law and natural rights. In *Christianity and law*, ed. J. Witte Jr. and F.S. Alexander, 89–105. Cambridge: Cambridge University Press.

Todorov, T. 1999. *The fragility of goodness: Why Bulgaria's Jews survived the Holocaust*. Princeton: Princeton University Press.

Williams, B. 1981. "Moral luck" pp. 20–40 in his *Moral luck*. Cambridge: Cambridge University Press.

Winnicott, D.W. 1971. *Playing and reality*. New York: Routledge.

About the Authors

Katharine Browne is a postdoctoral research fellow at the Centre for the Study of Mind in Nature (CSMN) at the University of Oslo. She received her PhD in philosophy from the University of Toronto. Her dissertation, 'Reason, Evolution, and the Possibility of Cooperation', examined the relationship between an empirical investigation into the foundations of human cooperation and the traditional normative questions in moral and political philosophy. Her current research focuses on the theoretical and practical implications of a cultural evolutionary analysis of human cooperative behaviours. She also has research interests in health-care ethics.

Markus Christen studied philosophy, physics, mathematics and biology and did a PhD at the Institute of Neuroinformatics at ETH Zurich, Switzerland, involving a research stay at the Max Planck Institute for the History of Science, Berlin, Germany. Currently, he is researcher in neuro-ethics and empirical ethics at the Institute of Biomedical Ethics of the University of Zurich and Visiting Scholar at the Psychology Department of the University of Notre Dame, USA.

Florian Cova is a postdoctoral researcher at the Swiss Centre in Affective Sciences, currently working on the relationship between emotions and moral evaluations. His main interests are experimental philosophy and moral psychology and he has published on these topics in journals such as *Consciousness and Cognition* (on intuitions about free will)*, Mind and Language* (on intentional action) and *The Monist* (on aesthetic judgments).

Jessy Giroux is a philosophy PhD student at the University of Toronto. His research interests are mainly in meta-ethics. He currently works on the role of intuitions in contemporary moral philosophy, which has led him to explore many related topics in moral epistemology and normative ethics, such as the use of the method of reflective equilibrium and the possibility of a priori moral knowledge. His philosophical interests often lead him to explore more empirical questions, such as the sources of moral motivation or the origin of moral norms.

Catherine Herfeld studied economics and philosophy of the social sciences at Heidelberg University, the Technical University Berlin and the London School of

B. Musschenga and A. van Harskamp (eds.), *What Makes Us Moral? On the capacities and conditions for being moral*, Library of Ethics and Applied Philosophy 31,
DOI 10.1007/978-94-007-6343-2, © Springer Science+Business Media Dordrecht 2013

Economics. She currently is a research fellow at the Centre for the History of Political Economy at Duke University. She is about to complete her dissertation (Witten/Herdecke University), tracing the origins and philosophical foundations of rational choice theory. Her research interests include the philosophy of economics, the emergence of mathematical economics, the phenomenon of economics imperialism, and the work of Amartya Sen.

Julia Hermann is Lecturer in social philosophy at Maastricht University. She graduated in philosophy and political science from the University of Heidelberg and received her PhD from the Department of Political and Social Sciences of the European University Institute. Julia has been a visiting scholar at the Philosophy Department of the University of California, Berkeley, and a teaching fellow in the *European Master's Programme in Human Rights and Democratization* in Venice. Her research interests include theories of moral justification and moral knowledge, philosophy of education, global justice, human rights, the ethics of personal relations, and the later Wittgenstein.

Andrés Luco is an Assistant Professor in the Philosophy Group at Nanyang Technological University, Singapore. He holds a PhD in philosophy from Duke University, and has previously taught philosophy at North Carolina State University and the University of Cape Town.

Stephen Maitzen is the W.G. Clark Professor of Philosophy at Acadia University, Nova Scotia, Canada. He has authored articles in epistemology, philosophy of religion, applied ethics, decision theory, and metaphysics. His current research interests include scepticism, God and the problem of evil, the impact of theism on human morality, and the perennial pseudo-question 'Why is there something rather than nothing at all?' In 2006 he received Acadia University's highest award for excellence in teaching.

Gerben Meynen studied philosophy, theology, and medicine. He received a PhD in philosophy (2006) as well as in medicine (2007). He specialized in psychiatry at VU medical centre Amsterdam. In 2007 he received a Veni-grant (NWO), project entitled: 'Free will and mental disorder. Exploring and evaluating the role of psychiatry in the philosophy of free will' (VU University, Amsterdam, Department of Philosophy). Currently, he is an assistant professor at the Faculty of Philosophy VU University and he works as a psychiatrist at GGZ InGeest. He is endowed Professor of Forensic Psychiatry at Tilburg Law School, Tilburg University.

Bert Musschenga is Professor of Philosophical Ethics at VU University, Amsterdam. From 1999 to 2003 he was the Director of the Netherlands School for Research in Practical Philosophy. He is Editor-in-Chief of *Ethical Theory and Moral Practice.* He published monographs on quality of life and on integrity. Besides that, he published many articles and edited many volumes in Dutch and in English on both practical ethical topics and on more fundamental theoretical questions such as the relation between religion and morality, moral reasoning, personal and moral identity. In recent years he published a number of articles on the relation between psychological and moral perspectives on moral intuitions.

Darcia Narvaez is Professor of Psychology at the University of Notre Dame in South Bend, Indiana, USA. She researches the early life experiences of young children and their influence on 'moral brain' development. Professor Narvaez developed several theoretical models: the Integrative Ethical Education model for moral character education, Triune Ethics Theory, which addresses the neurobiology of moral development in early life and adult moral mind-sets, and the Adaptive Ethical Expertise model which addresses how individuals build virtue. She has co-authored or co-edited nine books and published numerous articles and chapters. She is editor of the *Journal of Moral Education*.

Carsten Fogh Nielsen did postdoctoral work at the Centre for Subjectivity Research, University of Copenhagen and is currently part-time lecturer at the Department of Education, Aarhus University and the Department of Law, University of Southern Denmark. He has published several papers on popular culture and moral philosophy in The Blackwell Philosophy and Popular Culture Series and is currently working on a series of papers developing a Hegelian approach to the problem of moral motivation.

Thomas Ott studied physics and did a PhD at the Institute of Neuroinformatics at ETH Zurich, Switzerland, and at the University of Johannesburg, South Africa. Currently he is professor for simulation with specialization in nature-based computing at the Zurich University of Applied Sciences.

Alejandro Rosas received his PhD in Germany for his dissertation on *Kant's refutation of idealism*. In 1992 he obtained a full-time position at the Universidad Nacional de Colombia, in Bogotá. Since 1996 he has devoted his teaching and research to a naturalistic philosophical outlook, particularly to the relevance of evolutionary theory for philosophy. He organized the first Latin American Meeting on the Philosophy of Biology in 2006 in Bogotá, where participants from Argentina, Brazil, Mexico, and Colombia shared their research. He has spent research periods in Germany and Austria and works mainly on the evolutionary origins of morality.

David Edward Rose is a Lecturer in Philosophy at Newcastle University with interests in Hegelian ethics and Counter-Enlightenment thought and their application to contemporary moral and political issues. He has previously published books on the concept of free will and Hegel's Philosophy of Right and has published various articles on Giambattista Vico, ethics and political philosophy.

Katrien Schaubroeck studied philosophy at the University of Leuven, where she obtained her PhD in 2008, and currently works at Utrecht University as a postdoctoral researcher. Her research focuses on issues at the intersection of meta-ethics and moral psychology with a special interest in the nature of practical reason. She has published on a variety of topics within the domain of practical philosophy such as the normativity of law, cultural norms and practical reasons, love's reasons, the authority of morality, free will and responsibility. Her book *The normativity of what we care about; A love-based theory of practical reasons* will come out in 2012 (Leuven University Press).

Adam B. Seligman is Professor of Religion at Boston University and Research Associate at the Institute on Culture, Religion and World Affairs there. He has lived and taught at universities in the USA, in Israel and in Hungary where he was Fulbright Fellow. He lived close to 20 years in Israel where he was a member of kibbutz Kerem Shalom in the early 1970s. His many books include *The idea of civil society* (1992), *Inner-worldly individualism* (1994), *The problem of trust* (1997), *Modernity's wager: Authority, the self and transcendence* (2000), with Mark Lichbach *Market and community* (2000) *Modest claims; Dialogues and essays on tolerance and tradition* (2004), with Weller, Puet and Simon, *Ritual and its consequences: An essay on the limits of sincerity* (2008) and most recently, with Weller *Rethinking pluralism: Ritual, experience and ambiguity* (2012). He is director of the International Summer School on Religion and Public Life.

Bettine Siertsema studied Dutch language and literature at the VU University in Amsterdam. She took her PhD with a dissertation on Dutch diaries and memoirs on the concentration camps. Currently she holds a position in the Department of Philosophy of VU University and works on a book on the Holocaust perpetrator in literature. Her fields of interest are Holocaust literature and the interface of religion and literature.

Anton van Harskamp is adjunct Professor (Emeritus) social and cultural anthropology in: 'Religion, identity and civil society in multidisciplinary perspective', at VU University Amsterdam. His research focuses on modernity and new religions in the Western world, including fundamentalist religions. He now works on a research project on terrorism and religion. Together with André Droogers he works on a volume on the methodology of religious studies. With Bert Musschenga he is co-editor of this volume.

Harry Wels is Associate Professor at the Department of Organizational Sciences of VU University Amsterdam and at the African Studies Centre (ASC) in Leiden. Trained as an anthropologist, his research focuses mainly on structures of organizational cooperation in nature conservation in South and southern Africa. His publications include 'Private wildlife conservation in Zimbabwe. Joint ventures and reciprocity' (Leiden: Brill 2003), and together with Marja Spierenberg 'Conservative philanthropists, royalty and business elites in nature conservation in southern Africa' (*Antipode*). He is series editor of the SAVUSA-ASC Book Series with Brill.

Index

B. Musschenga and A. van Harskamp (eds.), *What Makes Us Moral? On the capacities and conditions for being moral*, Library of Ethics and Applied Philosophy 31,
DOI 10.1007/978-94-007-6343-2, © Springer Science+Business Media Dordrecht 2013