Morality and Emotion

(Un)conscious journey into being

Edited by
Sara Graça Da Silva

ROUTLEDGE

Morality and Emotion

Despite the many attempts to disentangle the relationship between morality and emotion, as is clear from the myriad of approaches that try to understand the nature and importance of their connection, the extent of this synergy remains rather controversial.

The multidisciplinary framework of the present volume was specifically designed to challenge self-containing disciplinary views, encouraging a more integrative analysis that covers various methodological angles and theoretical perspectives. Contributions include discussions on the interrelation between moral philosophy, emotion and identity, namely the clash between grand ethical theories and the practicality of human life; philosophical considerations on akrasia or the so called weakness of will, and the factors behind it; anthropological reflections on empathy and prosocial behavior; accounts from artificial intelligence and evolutionary game theory; and literary and artistic dissections of emotional responses to the representational power of fiction and the image.

The inclusion of chapters from varied scientific backgrounds substantially enriches this debate and shows that several core questions, such as the ones related to identity and to the way we perceive the other and ourselves, are transversal. It is therefore valuable and pressing to further explore these common threads, and to encourage disciplinary dialogues across both traditional and emerging fields to help shed new light on the puzzling and fascinating ways in which morality and emotion are mutually imbricated.

Sara Graça da Silva received her PhD from Keele University in 2008 with the thesis *Sexual Plots in Charles Darwin and George Eliot: Evolution and Manliness in Adam Bede and The Mill on the Floss*. Her research interests include the intersections between literature and science, theories of sexuality and gender, Darwinism, morality and emotion, and the evolutionary study of folktales. Graça da Silva has contributed to the *Victorian Literature Handbook*, the *Dictionary of Nineteenth Century Journalism*, *Utopian Studies*, and *Royal Society Open Science*, amongst others.

This page intentionally left blank

Morality and Emotion

(Un)conscious journey into being

Edited by Sara Graça da Silva

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2016
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
711 Third Avenue, New York, NY 10017

Routledge is an imprint of the Taylor & Francis Group, an informa business

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
Names: Silva, Sara Graça da, editor.
Title: Morality and emotion : (un)conscious journey to being / edited by Sara Graça da Silva.
Description: New York, NY : Routledge, 2016. | Includes bibliographical references and index.
Identifiers: LCCN 2015039097| ISBN 9781138121300 (hbk : alk. paper) | ISBN 9781315651040 (ebk : alk. paper)
Subjects: LCSH: Ethics – Psychological aspects. | Emotions.
Classification: LCC BJ45 .M675 2016 | DDC 171/.2–dc23
LC record available at http://lccn.loc.gov/2015039097

ISBN: 978-1-138-12130-0 (hbk)
ISBN: 978-1-315-65104-0 (ebk)

IELT is supported by National Funds through FCT - Fundação para a Ciência e Tecnologia - under the project PEst-OE/ELT/UI0657/2015

Typeset in Bembo
by HWA Text and Data Management, London

Contents

Contributors

Sara Graça da Silva received her PhD from Keele University in 2008 with a thesis on the rich interplay between nineteenth-century science and literature: 'Sexual Plots in Charles Darwin and George Eliot: Evolution and Manliness in *Adam Bede* and *The Mill on the Floss*'. Her main research interests include the relationship between science and literature; Darwinism and evolutionary theories; gender studies, the application of phylogenetic methodologies to the study of human cultural diversity, and morality. Currently a postdoctoral researcher at the Institute for the Study of Literature and Tradition, New University of Lisbon, Portugal, working on evolutionary readings of folktales. She has contributed to the *Victorian Literature Handbook, Dictionary of Nineteenth Century Journalism, Utopian Studies, Cambridge Scholars Publishing,* and *Royal Society Open Science*, amongst others.

Jesse Prinz is a Distinguished Professor of Philosophy and Director of Interdisciplinary Science Studies at the City University of New York, Graduate Center. His research focuses on the perceptual, emotional, and cultural foundations of human psychology. He is author of *Furnishing the Mind* (2002), *Gut Reactions* (Oxford, 2004), *The Emotional Construction of Morals* (2007), *Beyond Human Nature* (2012), and *The Conscious Brain* (2012). Two other books are forthcoming: *The Moral Self* and *Works of Wonder*.

Vasco Correia is a Postdoctoral Fellow in Philosophy at the Instituto de Filosofia da Linguagem, Universidade Nova de Lisboa (New University of Lisbon), where he works on the topic of biased reasoning in argumentation and decision-making, under the supervision of João Sàágua. He obtained his PhD in 2008 from the Université de Paris IV-Sorbonne with a dissertation on the problem of self-deception. Since then, he has published a book and numerous papers on the topics of rationality, argumentation and philosophy of mind.

Dina Mendonça has a Master's in Philosophy for Children under the supervision of Matthew Lipman and Ann Margaret Sharp, and holds a doctorate the University of South Carolina with a dissertation on

'The Anatomy of Experience – An Analysis of John Dewey's Concept of Experience'. She is currently a member of Instituto de Filosofia, at Universidade Nova de Lisboa, working on a situated approach to emotions, a novel and groundbreaking account that takes emotions as dynamic and active situational occurrences (Mendonça 2012), and explores and identifies further complexities of our emotional world (Mendonça 2013). Recently, she has been expanding the application of this perspective to argumentation theory. In addition to her research work in philosophy of emotion, she promotes and creates original material for application of philosophy as an aid in the creative processes at all stages of schooling.

Augusta Gaspar is an Assistant Professor at the Catholic University of Portugal and full researcher at CIS (Centro de Investigação e Intervenção Social), at ISCTE-IUL (Lisbon University Institute). Her research interests bridge psychology, biology and anthropology, and include the ontogeny and evolution of behavioural expression and physiological correlates of emotion. She is currently focusing on mediators of empathy and emotion recognition. Her work has been published in *Acta Ethologica*, *International Journal of Behavioural Development*, *Psychophysiology*, among others. She has published a book on evolution theory (*Evolução e criacionismo – uma relação impossível*, 2007, Quasi) and contributed to chapters to several books, including *Personality and Temperament in Nonhuman Primates*, edited by Alexander Weiss, James King and Alison Murray (2011, Springer-Verlag) and *The Evolution of Social Communication in Primates: A Multidisciplinary Approach,* edited by Marco Pina and Nathalie Gonthier (2014, Springer: Verlag). She is a founding member of the Portuguese Association of Primatology (APP).

Luís Moniz Pereira was Professor of Computer Science at Universidade Nova de Lisboa (retired), and founder and director of CENTRIA (1993–2008), its Artificial Intelligence research centre. He received an honorary doctorate from T.U. Dresden in 2006, was elected Fellow of the European Coordinating Committee for Artificial Intelligence (ECCAI) in 2001, and since 2006, has been a member of the Board of Trustees and the Scientific Advisory Board of IMDEA, the Madrid Advanced Studies Software Institute. He was the founding president of the Portuguese Artificial Intelligence Association, and is on the editorial boards of several scientific journals. His research interests centre on knowledge representation and reasoning, logic programming, and the cognitive sciences, having hundreds of publications under his name. At present, he is affiliated to the NOVA Laboratory for Computer Science and Informatics (NOVA-LINCS).

Eileen John is an Associate Professor of Philosophy at the University of Warwick, UK. She studied in the US, receiving her PhD from the University of Michigan and her BA from Yale. She taught previously at the University of Louisville. Her publications are primarily in aesthetics, with a special focus

on the philosophy of literature. She is particularly interested in how creative activity and art experience are involved in the development of autonomy and in the understanding and testing of values. She has served as the Director of Warwick's Centre for Research in Philosophy, Literature and the Arts, and is the co-editor of the Blackwell anthology, *The Philosophy of Literature*.

Carlos Augusto Ribeiro is a Researcher at the Instituto de Estudos de Literatura Tradicional (IELT), Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa, and a visual artist. He graduated in Fine Arts-Painting from Faculdade de Belas-Artes de Lisboa (FBAL) and obtained his Masters and PhD in Communication Sciences at Universidade Nova. His research is focused on the relationships between art and environment: speeches of art on earth. Recent activities include Extermínios (Exterminations), an exhibition at the Centro Cultural Emmerico Nunes, Sines (2009), 'Math is accurate even when it is wrong', a dossier of eight images created for the book *Contas X contos X cantos e que* + (ed. Ana Paula Guimarães and Adérito Araújo, Gradiva, 2012); 'Não corto carne, eu corto e retalho dicho', in *Revista Cerrados*, Brasília, 2013.

Acknowledgements

The inspiration for this book stemmed from a series of colloquiums on morality and emotion which I organise annually with IELT – Institute for the Study of Literature and Tradition – my centre at the Faculty of Social Sciences and Humanities, at the New University of Lisbon, Portugal. I am greatly indebted to the Morality and Emotion team for their friendship and competent help. I would also like to thank IELT and the Portuguese Foundation for Science and Technology (FCT) for their financial support.

My sincere thanks also to all the contributors in this volume for their expertise, enthusiasm, and cooperation during the review process, and for having given generously of their time. I extend a warm thanks to Vasco Correia for providing valuable comments on the Introduction. Further acknowledgements are due to the editors at Routledge for their friendly and efficient support, especially Ceri Griffiths, Michael Fenton, John Hodgson, Holly Knapp and Xian Gu.

Last but not least, I owe my deepest gratitude to my family. To Alberto Augusto Oliveira da Silva for his inspiring knowledge and many enjoyable hours of informed discussion and brainstorming, to Rita Graça da Silva for her insights, creative vision and encouragement, and to Maria Anita Marinho Graça for her unconditional love and optimism.

This page intentionally left blank

Introduction

Morality and emotion or 'well, that's another fine mess you got me into'

Sara Graça da Silva

> 'A man's got to know his limitations'.
>
> Harry Callahan, *Magnum Force* (1973)

If you stray from the path you will meet a big bad wolf and he will eat your granny

I vividly recall one episode that happened some years ago to one of my little cousins. She must have been three years old at the time. When her father told her off for picking flowers in a public garden, jokingly telling her she would be in trouble if a policeman caught her, she began wailing in absolute panic: 'Noooooooo, I don't want to go to prison, daddy, don't let them take me, I'm sorry, daddy, I want to go home!' Although our initial response was to laugh at her overreaction, we soon realised she was so wretched with guilt and fear, and so certain that an unspeakable punishment for her wrongdoing was coming her way, that she would not calm down despite our best reassurances. We ended up taking the vicious criminal home to rest.

Understanding morality and emotion, and reactions such as my little cousin's, has occupied scholars since the days of Plato and Aristotle. A book on this relationship is an assumedly ambitious project given the myriad of disciplinary approaches that try to make sense of this conundrum, each with its own particular standpoints and biases. The multidisciplinary framework of the present volume was specifically designed to challenge self-containing disciplinary views on the nature and importance of morality and emotion, encouraging a more integrative analysis that covers various methodological angles and theoretical perspectives.

Over the past few decades, this study has become increasingly interdisciplinary. There have been many solid and dedicated attempts to explain the threads connecting the two concepts, including António Damásio's groundbreaking *Descartes' Error: Emotion, Reason, and the Human Brain* (1994), Jonathan Haidt's *The Moral Emotions* (2003); Jesse Prinz's *The Emotional Construction of Morals* (2007), or Joshua Greene's *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them* (2013). Also noteworthy are the skilfully edited volumes *Moral Psychology: The Neuroscience of Morality: Emotion, Brain Disorders, and Development,* by Walter Sinnott-Armstrong (2007), *Morality and the Emotions*, by Carla Bagnoli

(2011), or *Emotions, Imagination, and Moral Reasoning*, by Robyn Langdon and Catriona Mackenzie (2012), to name just a few. Other influential works have centred on the analysis of specific emotions, such as fear, in *Fear Across the Disciplines* (2012), edited by Jan Plamper and Benjamin Lazier, or pain, in *Pain and Emotion in Modern History* (2012) by Rob Boddice, and Joanna Bourke's latest book *The Story of Pain: From Prayer to Painkillers* (2014).

However, and despite the many attempts to disentangle the relationship between morality and emotion, the extent of this synergy remains, for the most part, as mysterious as in Aristotle's time. I will not engage in a historical survey here as the authors of this book provide excellent background into the ancestry of this debate. While accounts from philosophy, biology and the neurosciences have been legion, particularly from the early 2000s onwards, contributions from other areas, namely from artificial intelligence, literature and the arts, have remained somewhat more guarded, despite influential works such as Marvin Minsky's *The Emotion Machine* (2007). The truth is that specific disciplines tend to provide their own interpretations, and many are either weary of 'contaminations' from other fields, or feel uncomfortable at the prospect of reflecting on the divergences and similarities that exist between different areas on this subject.

Times are a changin'

The reluctance (and appeal) in engaging in cross-disciplinary dialogues is in part explained by the inherent difficulty in finding *a* definition of morality and emotion capable of accounting for the different individual, social, historical, cultural, religious or academic interpretations. Such a consensus simply does not exist. Even in the same field, there is no agreement as to what morality and emotion really stand for. Although etymologically the definition of morality seems straightforward enough (from the Latin *moralitas*, meaning 'manner, character, proper behaviour'), what is meant by proper behaviour and associated moral codes is rather fluid. Likewise, the etymology of the term emotion (from the old French *emouvoir*, meaning 'to stir up', and from the Latin *emovere* meaning 'to remove, expel, to banish from the mind, to shift, displace', according to the OED), anticipates the strength of its eclectic psycho and physiological expressions. We are not all agitated by the same type of occurrences in the same way. Moreover, the same person can respond differently to similar situations depending on a number of factors (moods, thoughts, hormones, past experiences, memories, expectations, etc.).

Historically, the term emotion has shifted across the centuries, 'having won out over passion, affection and sentiment only in the past 200 years' (Frederik, 2009: 206). In his thorough '"Emotion": The History of a Keyword in Crisis', Thomas Dixon astutely notes that terms such as *passions* or *affections,* widely employed in the eighteenth century, were progressively replaced in the nineteenth century by the term *emotions* as a result of an increasing fascination with the self and a tendency to part with religious connotations in favour of a

more secular approach, 'detached from the linguistic worlds of theology and moralism' (Dixon, 2003; Dixon, 2012: 342).

The study of emotions received a great deal of attention in the nineteenth century as a result of the necessity historians like Dixon and others identify of 'articulat[ing] the assumed relationships between physiological processes and mental experiences' (Dixon, 2012: 343; see also Stedman, 2002). Influenced by earlier treatises on expression and emotion such as Charles Bell's *The Anatomy and Philosophy of Expression* (1806) or Alexander Bain' *The Emotions and The Will* (1859), Charles Darwin's seminal work *The Expression of Emotions in Man and Animals*, published in 1872, galvanised the study of emotion in an unparalleled fashion. That being said, inner feeling was not Darwin's main preoccupation. As Janet Browne notes, 'Darwin was more interested in the way the man's body actually worked, than in the theory of perception […] real phenomena were more useful in the fight to establish continuity between human and animal species' (Browne, 1985: 44). Today we know that this continuity is much more profound and intertwined than Darwin could have ever imagined, hence the title's deliberate emphasis on a 'journey to being', and not simply, to being human.

The journey itself is very much ongoing, and has been often hampered by a lack of understanding of its inherent complexity. For example, one of the difficulties when thinking about emotion has to do with its mercurial nature. In her book *Emotions in History – Lost and Found* (2011), Ute Frevert calls attention to this mutability in meaning by elaborating on emotions' constant cycle of loss and renewal:

> Emotions and emotional styles fade away and get lost (like honour or acedia) but [the historical economy of emotions] also witnesses the emergence of new or newly framed emotions. Empathy, sympathy/compassion serve as great examples of emotions that are found and invented in the modern period.
>
> (Frevert, 2011: 12)

Emotions and morality(ies) are deeply variable across time and space. More recently (2014), Frevert *et al.* touched on a similar analysis by presenting a judicious account of emotions' semantic fluctuation using encyclopaedias and lexica, mainly in texts written in English, German and French (2014: 9). Overall, the error seems to stem from a misguided preoccupation in limiting and circumscribing different dimensions into a sole reality instead of accounting for the multiplicity of representations.

Empathy and moral action

Empathy is another concept whose study and relation to moral decision-making has produced a tremendous array of critical literature which is not exclusive to humanity (Bekoff and Pierce, 2009). Prominent researchers like Frans de Waal, for

example, identified pillars of morality that apply to both human and non-human primates, namely reciprocity/prosocial behaviour and empathy/compassion (de Waal, 2010, 2013; de Waal et al., 2014). Research on kin selection, altruism and cooperation initiated in the 1970s has progressively reopened debates on emotion and on the evolution of morality and pro or anti-social behaviour.

In the early 1990s, Daniel Batson formulated his empathy-altruism hypothesis describing how 'empathic concern produces altruistic motivation', which does not necessarily imply self-sacrifice (2010: 2). According to Batson, while empathic emotions are 'other-oriented' rather than self-interested, the association of the latter to selfishness and immorality is flawed: 'to say that A (self-interest) is not B (moral) and that C (altruism) is not A does not mean that C is B. To say that apples are not bananas and that cherries are not apples does not mean that cherries are bananas' (ibid.: 17).

While the ability to read others' expression is crucial to empathetic responses, the hypothesis that empathy is necessary for moral deliberation has divided scholars into various camps. Resisting Batson's and others' formulation (see also Baron-Cohen, 2011), Jesse Prinz defends that empathy is not necessarily needed for moral judgment, and can even be potentially responsible for poor moral discernment: 'maybe empathy is a bad thing. It does not track approbation, and if we use it in that capacity, we would make moral mistakes' (Prinz, 2011a: 228; see also Prinz, 2011b). While Prinz is sceptical about the role empathy plays in moral action, he acknowledges the the key role emotions play in this context. In his sentimentalist theory, he notes that 'emotions co-occur with moral judgments, influence moral judgments, are sufficient for moral judgments, and are necessary for moral judgments, because moral judgments are constituted by emotional dispositions' (Prinz, 2006: 36).

Haidt's intuituinist theory also identifies a particular set of emotions or 'psychological *foundations* upon which cultures construct their moralities' (Graham *et al.*, 2011: 5; Haidt and Joseph, 2004). Haidt describes these as *moral emotions*, 'linked to the interests or welfare either of society as a whole or at least of persons other than the judge or agent' (Haidt, 2003: 583). Whilst the concept of intuition is not new, the originality of Haidt's approach lies in the identification of five particular intuitions that work as 'innate "taste buds" of the moral sense and result in "affectively valenced experiences" such as likes or dislikes', namely: harm/care, fairness/reciprocity, ingroup/loyalty, authority/respect, and purity/sanctity (Haidt and Joseph, 2007: 386).

It is widely recognised that, as a species, we are social beings, and that our 'moral values ground a life that is a social life' (Churchland, 2011: 12). Our gregarious nature brings several challenges to the table when considering how different moralities and emotional responses play out in the gritty dramas of real life. This preoccupation is the driving force behind Prinz's chapter, the first of the present volume. His contribution adds two important layers to the mix by acknowledging that morality is 1) 'highly variable'; and 2) integral to personal identity. By seeking an application of ethics that deals with real moral conflicts at the personal, cultural and political levels, Prinz sets out to dust 'the level

of abstraction' that characterises the 'Grand Theories' of morality, including Aristotle's virtue theory, John Stuart Mill's consequentialism or Kant's duty principle. In his view, these theories 'risk irrelevance if they do not make contact with how morality is actually experienced in human life'.

Certainly, our exposure to socialisation impacts deeply on our (in)actions and on the way we react and perceive the other and ourselves. However, our actions can sometimes go against our best judgment. Vasco Correia and Dina Mendonça both explore this phenomemon by elaborating on akratic action. Drawing on a Belief-Desire theory that affords emotions a key role in decision-making, Correia elaborates on the factors behind 'weakness of will', or akrasia, as well as impulsiveness, whilst suggesting a number of self-control strategies to avoid this seemingly irrational behaviour, such as emotional regulation. Akin to Prinz's difficulty in finding real applicability to the *grand moral theories*, Correia considers Kant's dichotomical paradigm between passions/desires and will/reason to 'fall short of explaining what sort of principle governs the so-called will power (or strength of will)'. On the contrary, in belief-desire theories such as the ones proposed by Mele, the agent's behaviour is explained by 'both a desire and a belief' that motivate us to act. From this perspective, impulsiveness and irrational action typically stem from a cognitive illusion known as 'hyperbolic discounting', which misleads the agent into believing that the smaller, earlier reward (e.g. eat candy) is preferable to the larger, later one (e.g. lose weight). Yet, this preference reversal is only temporary and tends to dissipate as soon as the urging desire is satisfied, which seemingly 'explains why we often regret our indulgencies in hindsight'.

Continuing with the theme of akrasia but steering it into a slightly different angle, Mendoça proposes to examine the role of akratic feelings in creating meta-emotions of puzzlement which, in turn, enhance empathetic responses by promoting a continual search for self-knowledge. Mendonça also identifies the complex and layered nature that characterises most emotional states by pointing out the 'difficulty in separating first order levels of emotion (an emotion: fear) from second order levels of emotion (an emotion about an emotion: sadness about fear)'. According to Mendonça, it is this same difficulty that explains emotional confusion when reading the other, and ultimately, engagement in less cooperative behaviour. Bridging with Prinz's and Correia's conclusions, Mendonça too recognises the 'great potential for discussions about ethics and morality' that the acceptance of akratic feelings as 'a crucial and an inherent part of our emotional world may be one of the ways to better understand the individuality of each one of us, as well as the similarities and differences we share as a species'.

Drawing from insights from anthropology and psychology, Chapter 4, by Augusta Gaspar, touches on a similar understanding by discussing how morality and empathy are the basis of prosocial behaviour, in both humans and other animals. She engages in a meticulous literature review demarcating the different factions, from Batson to Haidt's intuitionist theory or Alan Fiske's social cognition models. Reintroducing Mendonça's acknowledgement of the importance of empathetic responses to the other's situation, Gaspar stresses the

decisive impact of environmental influences, including parental investment and social interaction, in the child's and adolescent's moral development (namely the formulation of a 'inner sense of right and wrong') and Theory of Mind (ToM), understood as the ability to 'infer the full range of mental states (beliefs, desires, intentions, imagination, emotions, etc.) that cause action […] to reflect on the contents of one's own and other's minds' (Baron-Cohen, 2001: 175).

Gaspar also raises important questions regarding the role of *learning* in specific social and cultural contexts, recognising its impact on what we see as morally right and wrong, as well as on our emotional responses, of fear for instance.[1] In this context, she notes that folktales are excellent repositories of information on the transmission of different moral values across cultures.[2] She concludes her chapter by providing an account of a number of relevant studies on empathy beyond the realm of human primates carried out by scholars including de Waal, Goodal, Bekoff, Fiske, and herself. She focuses especially on research on 'consolation, cooperation and altruism' as 'key manifestations of prosocial behaviour'.

Indeed, our prosocial nature demands that we cooperate. Research in cognitive neuroscience and moral psychology suggests that behaving morally helps solve and negotiate social problems (Axelrod and Hamilton, 1981; Damásio, 1994; Ciarameli et al., 2007; Greene *et al.*, 2004; Boyd and Richerson, 2009; Joyce, 2006; Haidt and Joseph, 2004, 2007). As Robin Allott notes, 'morality is a key factor in the success of human groups in competition or co-existence with each other' (Allott, 1991: 455). Darwin himself defined a 'moral being as capable of comparing his past and future actions or motives, and of approving or disapproving of them', and described how natural selection favoured altruistic behaviour as a means to potentiate the group's success and ensure a greater number of offspring (Darwin, 2004: 84).

Advances in the area of artificial intelligence have yielded very interesting insights into the mechanisms underlying moral cooperative and altruistic efforts. In an original addition to the volume, Luís Moniz Pereira takes up issues raised in the previous chapters and proposes an approach 'sans emotions but with ethical discernment' drawing from machine ethics, 'a sprouting interdisciplinary field of enquiry arising from the need of imbuing autonomous agents with some capacity for moral decision-making'. Anchored in Turing's theory of functionalism that sees 'mental states to be multiply realised […] without limiting the class of minds to creatures with brains like ours', Pereira notes that by introducing five cognitive abilities – 'intention recognition, commitment, revenge, apology, and forgiveness' – on computational simulations using techniques from Evolutionary Game Theory (EGT), one can observe the 'emergence of cooperation' at the collective level from a privileged and seemingly unbiased position.

Fiction as a thought experiment

The ability to empathise with the other (or not) brings me back to the importance of the arts as crucial stimuli in emotional responses, in both the creator and in the receptor. The last two chapters explore how art in general, and literature in

particular, represent a privileged source of information on the articulation of morality and emotion by testing a variety of cognitive abilities. A recent study published in *Science* by David Kidd and Emanuele Castano (2013) concluded that reading fiction improves the development of Theory of Mind (ToM) in comparison to reading non-fiction. These findings, which Gaspar also tackles in her chapter, are in line with previous studies that identified a link between reading fiction and an enhanced capacity for empathetic responses (Mar, Oatley, 2008; Oatley, 2011; Johnson, 2012). Linking with Mendonça's earlier elaboration on the way meta-emotions are grounded in values and beliefs, Chapter 6, by Eileen John, sets out to discuss how the values we hold as moral might impact on our emotional responses to fiction, and even on our real-life decisions. She acknowledges different perspectives, contrasting, for instance, *make-believe* approaches which assume that we believe that there are people living those situations, with *thought theories*, which claim that the former fail to reunite the conditions for genuine emotions. Finally, using Saramago's novels *Blindness* and *Seeing*, she demonstrates that we 'respond to fictional characters at least in part as the products of representational activity, as things that manifest choices and possibilities for identifying and presenting what is worth noticing and understanding about human life'.

Identifying and presenting what is worth *noticing* is the main preoccupation behind Carlos Augusto Ribeiro's proposed reading of Edgar Alan Poe's 'The Oval Portrait' and Adolfo Bioy Casares's *The Invention of Morel.* The last chapter extends previous discussions on the representational function of literature by examining the disturbing power of the image in both stories, and particularly how:

> there seems to exist a disturbing contagious law, acting over our heads, which makes us believe in the existence of a system of transfusion between referents and representations, objects and images; and suspect of a perverse relationship between visible and invisible, material and immaterial, vision and blindness.

Through the analysis of this system of transfusion, Ribeiro invites reflections on the morality of the behaviour portrayed by the painter and the witnesses in 'The Oval Portrait', and by the inventor in *Morel's Invention*. In his argumentation, he relates the evil of specific actions with studies on psychopathy and lack of empathy (something that Gaspar also discusses), such as Zimbardo's Stanford experiment in the early 1970s, and with contemporary performances and art installations that test emotional responses in intense sensorial environments.

Many scholars have noted the importance of the visual and the image as crucial instigators of emotional responses. Neuroaesthetics, for example, is a relatively recent but promising field, galvanised by the pioneering research undertaken by the likes of Semir Zeki (1999), Anjan Chatterjee (2013) or Arthur Shimarura (2013), to name just a few. This ties in well with perspectives such as the one proposed by Prinz, which see contemporary science moving into a more Humean, empirical direction, explicitly recognising that our thinking is affected by sensory, bodily and environmental stimuli.

In 2013, I attended a conference on Democracy and Emotions at the Centre for the History of Emotions, at the Max Planck Institute for Human Development, in Berlin, organised by Philipp Nielsen. At the time, a comment from one of the speakers struck a chord with me for both its simplicity and reach. James Jasper, a sociologist at the Graduate Centre of the City University, in New York, pointed out how not so long ago having emotions excluded people from citizenship whereas nowadays they are perceived as a requirement. Additionally, as 'experiences of involvement', emotions have different levels of motivational power (Barbalet, 2011: 36). Some, like anger, for example, prompt us to act more readily than, say, regret or despair.

The question of whether our moral decisions are performed consciously or unconsciously has been a matter of long-standing debate. Whereas many claim that 'conscience is a regulator of conduct in moral domains', others have challenged this assumption (Janoff-Bulman, 2011: 133). Distrusting rationalist models proposed by Kohlberg, for example, so-called 'gut' theorists like Prinz or social intuitionists like Haidt *overall* defend conscious thought processes as 'post hoc justifications' on previous judgments that were mainly driven by automatic intuitions or 'gut reactions' (Haidt, 2001; Prinz, 2004).[3] Contrary to Haidt's and Prinz's view, Paul Bloom and David Pizarro have argued that 'fast and automatic moral intuitions are actually shaped and informed by prior reasoning' (Pizarro and Bloom, 2003: 193). A thorough elaboration on the different sides and perspectives regarding this subject would be a book on its own, I suspect, especially when there is yet so much to discover about how the brain, neural maps and the body are interconnected or how the *Self Comes to Mind* (Damásio, 2010). For now however, it will suffice to say that debates concerning consciousness, its meaning and expression, are far from consensual.

Understanding what is moral and what is not, as well as the role of emotions in decisions and behaviour, has and will continue to engage scholars from various disciplines. Although advances in neuroimaging techniques and behavioural psychology have provided compelling evidence confirming that emotions play a key role in the morality and rationality of our decisions (May, 2014), and despite the fact that there have been numerous experiments on moral judgments in lab environments with real subjects (or computational 'subjects', as we have seen), there is arguably neither a bigger nor more perplexing lab than real life (Greene and Haidt, 2002; Eskine *et al.*, 2011; Inbar, Pizarro and Bloom, 2012).

From a very early age, most individuals display a deep concern for questions of right and wrong, innocence and guilt, and try to make sense of them, often unsuccessfully. Just ask my cousin, the mobster. The present volume represents a humble contribution to this ongoing effort by assembling expert perspectives from varied scientific backgrounds. As a whole, it intends to provide an impression of the nature and variety of research that has recently been carried out in the study of the topic whilst showing that it is possible (and necessary) to find common cross-disciplinary threads in order to help shed new light on the puzzling and fascinating ways in which morality and emotion are mutually imbricated.

Notes

1 The motifs covered in folktales reflect themes that have fascinated humankind from time immemorial. An example of this is the relation between fear learning and the formation of long-lasting memories, as well as the role of social environments in fear regulation – something that might resonate with many aspects of our day-to-day life (LeDoux, 1996, 2012; Raio and Phelps, 2015).

2 The study of morality in folktales is fascinating to many. Renowned folklorist Stith Thompson commented on how 'the interest of the reader or hearer is always carried along by the interplay of contrasting forces: the good and the evil, the clever and the stupid, hero and villain, faithful and unfaithful' (1946: 108). Questions of trust and distrust, for example, are rampant in tales, with countless stories of faithless sisters, mothers, and wives driven by unnatural and amoral motivations. The preference for not cooperating with close kin is certainly unexpected, but not that surprising if we consider issues such as sibling rivalry, jealousy and competition for the progenitors' attention. Disloyalty by progenitors towards their offspring, however, is perhaps harder to understand. Interestingly, some evolutionary psychologists refer to a 'Cinderella effect' or 'Cinderella complex', which consists of less parental investment in those who are not genetically related to us in favour of biological descendants (Daly and Wilson, 1999, 2005).

3 This intuitionist reasoning is thoroughly explained in Haidt's and Bjorklund's chapter entitled 'Social Intuitionists Answer Six Questions about Moral Psychology', in W. Sinnott-Armstrong's (2008) excellently edited volume *The Moral Psychology, Volume 2: The Cognitive Science of Morality: Intuition and Diversity*. Cambridge, MA: MIT Press, 181–217.

References

Allott, R. (1991). 'Objective Morality'. *Journal of Social and Biological Structures*, 14(4), 455–471.

Axelrod, R. and Hamilton, W.D. (1981). 'The Evolution of Cooperation'. *Science*, 211, 1390–1396.

Bain, Alexander (1859). *The Emotions and the Will.* London: John Parker.

Barbalet, J. (2011). 'Emotions Beyond Regulation: Backgrounded Emotions in Science and Trust'. *Emotion Review*, 3, 36–43.

Baron-Cohen, Simon. (2001). 'Theory of Mind in Normal Development and Autism'. *Prisme*, 34, 174–183.

Baron-Cohen, Simon. (2011). *The Science of Evil: On Empathy and the Origins of Cruelty.* New York: Basic Books.

Batson, Daniel C. (2010). *Altruism in Humans.* Oxford: Oxford University Press.

Bekoff, Marc and Pierce, Jessica (2009). *Wild Justice: The Moral Lives of Animals.* Chicago, IL: University of Chicago Press.

Bell, Charles (1872). *The Anatomy and Philosophy of Expression: As Connected with the Fine Arts.* London: Henry G. Bohn.

Bourke, Joanna (2014). *The Story of Pain: From Prayer to Painkillers.* Oxford: Oxford University Press.

Boyd, R. and Richerson, P. J. (2009). 'Culture and the Evolution of Human Cooperation'. *Philosophical Transactions of the Royal Society (B)*, 364: 3281–3288;

Browne, Janet (1985). 'Darwin and the Expression of Emotions'. In *The Darwinian Heritage*, ed. David Khon. Princeton, NJ: Princeton University Press.

Chatterjee, Anjan (2013). *The Aesthetic Brain.* Oxford: Oxford University Press.

Churchland, Patricia (2011). *Braintrust: What Neuroscience Tells Us about Morality.* Princeton, NJ: Princeton University Press.

Ciaramelli, E., Muccioli, M., Ladavas, E. and di Pellegrino, G. (2007). 'Selective Deficit in Personal Moral Judgment Following Damage to Ventromedial Prefrontal Cortex'. *Social Cognitive and Affective Neuroscience,* 2: 84–92.

Cova, F., Deonna, J. and Sander, D. (2013). 'The Emotional Shape of our Moral Life: Anger-related Emotions and Mutualistic Anthropology'. *Behavior Brain Sciences,* 36(1): 86–87.

Daly, Martin and Wilson, Margo (1999). *The Truth about Cinderella: A Darwinian View of Parental Love.* New Haven, CT: Yale University Press.

Daly, Martin and Margo Wilson (2005), 'The 'Cinderella' Effect: Elevated Mistreatment of Stepchildren in Comparison to Those Living with Genetic Parents'. *Trends in Cognitive Sciences,* 9: 507–508.

Damásio, António (1994). *Descartes' Error: Emotion, Reason, and the Human Brain.* New York: Grosset/Putnam.

Damasio, António (2010). *Self Comes to Mind*. New York: Pantheon

Damásio, António and Carvalho, G.B. (2013). 'The Nature of Feelings: Evolutionary and Neurobiological Origins'. *Nature Reviews Neuroscience*, 14(2): 143–152.

Darwin, Charles (1998). *The Expression of Emotions in Man and Animals*, ed. Paul Ekman. London: HarperCollins.

Darwin, Charles (2004). *The Descent of Man and Selection in Relation to Sex,* ed. A. Desmond. Harmondsworth: Penguin.

de Waal, Frans (2010). *The Age of Empathy: Nature's Lessons for a Kinder Society.* London: Souvenir.

de Waal, Frans (2013). *The Bonobo and the Atheist: In Search of Humanism Among the Primates*. New York: WW Norton.

de Waal, Frans, Churchland, Patricia S., Pievani, Telmo and Parmigiani, Stefano (2014). *Evolved Morality: The Biology and Philosophy of Human Conscience.* Leiden: Brill.

Dixon, Thomas (2002). '"Emotion": The History of a Keyword in Crisis', *Emotion Review*, 4: 338–344.

Dixon, Thomas (2003). *From Passions to Emotions: The Creation of a Secular Psychological Category.* Cambridge: Cambridge University Press.

Eskine, K. J., Kacinik, N. A. and Prinz, J. J. (2011). 'A Bad Taste in the Mouth: Gustatory Disgust Influences Moral Judgment'. *Psychological Science*, 22, 295–299.

Frederik, Shane (2009). 'History of Emotion'. In The *Oxford Companion to Emotion and the Affective Sciences,* ed. David Sander and Klaus Scherer. Oxford: Oxford University Press.

Frevert, Ute (2011). *Emotions in History – Lost and Found.* Budapest: CEU Press.

Graham, Jesse, Brian A. Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter. H. Ditto (2011). 'Mapping the Moral Domain'. *Journal of Personality and Social Psychology* 101(2): 366–385.

Greene, J. and Haidt, J. (2002). 'How (and Where) Does Moral Judgment Work?' *Trends in Cognitive Sciences*, 1, 6(12): 517–523.

Greene, J. D., Nystrom, L. E., Engell. A. D., Darley, J. M. and Cohen, J. D. (2004). 'The Neural Bases of Cognitive Conflict and Control in Moral Judgment'. *Neuron*, 44(2): 389–400

Gross, David M. (2006). *The Secret History of Emotion: From Aristotle's Rhetoric to Modern Brain Science*. Chicago, IL: University of Chicago Press.

Haidt, J. (2001). 'The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgment'. *Psychological Review*, 108: 814–834.

Haidt, J. (2003). 'The Moral Emotions'. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of Affective Sciences*. Oxford: Oxford University Press (pp. 852–870).

Haidt, J. and Joseph, C. (2003). 'Elevation and the positive psychology of morality'. In: *Flourishing: Positive Psychology and the Life Well-Lived*, ed. C.L. Keyes and J. Haidt Washington, DC: American Psychological Association, 275–289.

Haidt, J. and Joseph, C. (2004). 'Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues'. *Daedalus*, 133 (4): 55–66.

Haidt, J. and Joseph, C. (2007). 'The Moral Mind: How Five Sets of Innate Intuitions Guide the Development of Many Culture-specific Virtues, and Perhaps even Modules'. *The Innate Mind*, 3: 367–392.

Inbar Y., Pizarro D. A. and Bloom P. (2012). 'Disgusting Smells Cause Decreased Liking of Gay Men'. *Emotion*, 12: 23–27.

Janoff-Bulman, R. (2011). 'Conscience: The Do's and Don'ts of Moral Regulation'. In: *The Social Psychology of Morality: Exploring the Causes of Good and Evil*, ed. by M. Mikulciner and P. Shaver. Washington DC: American Psychological Association, 131–148.

Johnson, Dan R. (2012). 'Transportation into a Story Increases Empathy, Prosocial Behavior, and Perceptual Bias Toward Fearful Expressions'. *Personality and Individual Differences,* 52(2): 150–155.

Joshua May (2014). 'Does Disgust Influence Moral Judgment?' *Australasian Journal of Philosophy*, 92 (1): 125–141.

Joyce, Richard (2006). *The Evolution of Morality.* Cambridge, MA: MIT Press.

Kagan, Jerome (2007) *What is Emotion? History, Measures, and Meanings*. New Haven, CT: Yale University Press.

Kidd, David and Castano, Emanuele (2013). 'Reading Literary Fiction Improves Theory of Mind'*, Science*, 342 (6156): 377–380.

Kohlberg, L. (1984). *The Psychology of Moral Development: The Nature and Validity of Moral Stages*. New York: Harper & Row.

LeDoux J. E. (1996) *The Emotional Brain*. New York: Simon & Schuster.

LeDoux J. E. (2014). 'Coming to Terms with Fear'. *PNAS*, 25, 111(8): 2871–2878.

Mar, Raymond and Oatley, Keith (2008). 'The Function of Fiction is the Abstraction and Simulation of Social Experience'. *Perspectives on Psychological Science*, 3: 173–192.

Oatley, Keith (2011). 'In the Minds of Others'. *Scientific American Mind*, 22(6): 62–67.

Oerlemans, Ono (2002). *Romanticism and The Materiality of Nature*. Toronto: University of Toronto Press.

Plamper, Jan (2015). *The History of Emotions: An Introduction*. Oxford: Oxford University Press.

Prinz, J. (2004). *Gut Reactions: A Perceptual Theory of the Emotions*. Oxford: Oxford University Press.

Prinz, J. (2006). 'The Emotional Basis of Moral Judgments'. *Philosophical Explorations*, 9 (1): 29–43.

Prinz, J. (2011a). 'Against Empathy'. *The Southern Journal of Philosophy*, 49 (Spindel Suppliment), 214–233.

Prinz, J. (2011b). 'Is Empathy Necessary for Morality?' In: *Empathy: Philosophical and Psychological Perspectives* (1971), ed. by P. Goldie and A. Coplan. New York: Oxford University Press, 211 – 229.

Pizarro, D. A .I and Bloom P. (2003). The Intelligence of the Moral Intuitions: Comment on Haidt (2001). *Psychological Rev*iew, 110(1): 193–196.

Raio, C. M. and Phelps, E. A. (2015). 'Observational Fear Learning'. In: *Brain Mapping: An Encyclopedic Reference*, ed. by A.W. Toga. Oxford: Elsevier, 137–141.

Richerson, P. and Boyd, R. (2005). *Not by Genes Alone: How Culture Transformed Human Evolution.* Chicago: University of Chicago Press.

Shimarura, Arthur (2013). *Experiencing Art: In the Brain of the Beholder.* Oxford: Oxford: University Press.

Stedman, Gesa (2002). *Stemming the Torrent: Expression and Control in the Victorian Discourses on Emotions, 1830–1872.* Aldershot: Ashgate.

Thompson, Stith (1977). *The Folktale.* Berkeley, CA: University of California Press [1946].

Zeki, Semir (1999). *Inner Vision: An Exploration of Art and the Brain.* Oxford: Oxford University Press.

1 Emotions, morality, and identity

Jesse Prinz

Moral theory and human life

Morality is grounded in emotions. That was the conclusion of the eighteenth-century philosophers known as sentimentalists. The judgement that something is morally good or bad, they claimed, consists in positive or negative feelings. This is a controversial view, but it has come back into vogue in recent years, and it has become the subject of intense psychological testing. Here I will argue that recent empirical research can be used to support and extend the sentimentalist view of morality. I will also explore two implications: morality is highly variable and also integral to personal identity. Moral variation was recognized in the eighteenth century, though sentimentalists tried to resist it in various ways. The link between morality and identity is a more recent insight, but it may be the most important lesson that we can extrapolate from sentimentalist theories. This link, I will argue, allows us to shed light on moral conflicts that arise in human life. Philosophical work in ethics has often operated at a level of abstraction that makes real-word application difficult. The link between morality, emotion, and identity offers a possible remedy.

Grand ethical theories

Morality is one of the central topics in philosophy. Many of the most important figures in the history of the field have advanced moral theories. The most celebrated theories consist of grand principles of guidelines dictating what it is to act ethically or to conduct a virtuous life. These theories are characteristically presented as universal: governing all human conduct. They are also presented as timeless, abstracting away from historical and geographical contexts. Immanuel Kant even suggested that morality is like mathematics: a set of immutable truths, derived *a priori*. On the other hand, the architects of these grand theories have also appreciated that morality depends on human psychology. Aristotle tries to derive moral guidelines from a theory of human nature, John Stuart Mill begins with the nature of human happiness, and Kant derives his moral system from an analysis of human agency, focusing on our ability to see ourselves as acting from reason. There is, I believe, a tension between the universalizing and abstract

ambitions of moral theory, and this focus on psychology. If we grant that morality has psychological foundations and let the study of human psychology inform our inquiry, we may come to see morality in a different light. Psychology tells us that morality is passionate, personal, and parochial. Grand theories may have a place, but they risk irrelevance if they do not make contact with how morality is actually experienced in human life. Grand theories shed little light on the moral conflicts that divide people and nations, and this failing undercuts their value, both as explanations of human behaviour and as tools for improvement. Here, I want to explore morality as we live it. I will claim that morality is grounded in our emotions, and that these emotions are shaped by culture and history. I will also claim that, as a result, morality is linked to identity, and this link is crucial for grasping how morality plays out in personal and political spheres.

To begin, I want to elaborate on the contention that grand ethical theories do not make adequate contact with human life. In the history of Western ethics, three grand theories have had vastly more impact than any others. I will briefly introduce these and then indicate why they may be inadequate guides to real moral problems. The first approach I will consider is virtue theory, which was most influentially formulated by Aristotle in his *Nicomachean Ethics*. Unlike modern ethical theories, which focus on norms for behaviour, Aristotle focuses on norms for character. Instead of asking 'How should people act?', he begins with the question, 'How should people be?' By cultivating good character, he says, we can arrive at good action, and we can also lead lives that qualify as living well or flourishing. The details of Aristotle's account introduce a number of controversial commitments, which his followers have not always accepted. He derives his list of virtues using a doctrine of means, according to which virtuous traits lie midway between excesses and deficits. Courage is a virtue, he says, because it lies between Cowardliness and Recklessness. Though elegant, it is dubious how far this doctrine can go. For example, one might think it is a virtue to be loyal, but is it really a vice to be too loyal? Likewise for kindness, wisdom, empathy, and a sense of justice. Aristotle also makes the contentious and demanding claim that one cannot have any virtuous trait without having them all, and he adds that virtue can only be pursued by the affluent, and cannot be assessed until the end of life. One can reject these commitments, while maintaining the core thesis that ethics is fundamentally concerned with the cultivation of good character. Virtue theories were popular among Medieval European philosophers, and have much currency today. Character-based approaches have also been developed within the Confucian tradition in China.

The second grand theory is Consequentialism, which is associated with the British moralists, Jeremy Bentham and John Stuart Mill. Consequentialist theories define the morally right action as the one with the best (or the one that tends to have the best) consequences. Good consequences are measured in different ways by different theories. Bentham and Mill emphasize pleasure, which they claim is intrinsically good, and thus the best actions are those that maximize pleasure. Some consequentialists find the focus on pleasure too limiting: there are other goods that may be important. Pleasure can be fleeting,

and superficial. Instead of pleasure, consequentialists can measure goodness in terms of preference satisfaction.

The third grand theory owes to Kant. He rejects the focus on consequences advocated by Bentham and Mill, and says we should act from duty. A duty is something that applies universally to all, regardless of individual preferences. For Kant, we should do only that which we could coherently command all others to do. This is one formulation of his 'categorical imperative'. On another, perhaps distinct, formulation Kant says we should always treat others as ends, rather than as means. He calls this the 'principle of humanity'. There are technical problems with both formulations. The first delivers bizarre results unless we can come up with a set of principled constraints on the appropriate level of abstraction for formulating commands. I might want to go to art school, but I certainly would not command everyone one to do that; if everyone went to art school, there would be no one to take on other jobs that make life (and art!) possible. It does not follow that going to art school is immoral. The principle of humanity looks too demanding on the face of it: whenever someone performs a service for us, we are using that person as a means. One might avoid this problem by interpreting the principle as an injunction to respect people's autonomy. That, however, may be impossible if it turns out human autonomy is a fiction, as sceptics about free will are inclined to believe. Followers of Kant try to cope with these worries, and they maintain his emphasis on the idea that moral mandates can be derived from universal features of the human will.

In introducing these theories, my goal is not to explain them fully, much less to review the arguments for and against. Rather, I want to recall the gist of each theory (in both traditional and more up-to-date versions), in order to make a more general point about grand ethical theories. That point concerns the challenge of putting such theories to work in the real world.

Can ethical theories be applied?

Consider the moral problems that arise in human life. These can be culled from daily headlines. Consider violent international conflicts, civil wars, ethnic rivalries, and sectarian violence: Israel and Palestine, Russia and Ukraine, India and Pakistan, Armenia and Azerbaijan, Tutsi and Hutu, Dinka and Nuer, Turks and Kurds, Hindu and Muslim, Buddhist and Muslim, Shia and Sunni, secular vs. religious, communist and capitalist, and so on. Grand ethical theories all have some resources for condemning violence: violence fails to exhibit virtue, reduces pleasure, and disrespects autonomy. But these theories offer neither remedy nor diagnosis for such violent conflicts. If parties to these battles were to study the classics of Western ethics, they would make little progress. Usually, both sides see the other as blameworthy, and violent reprisal as a justified wrong. Once a violent rivalry has begun, there are urgent practical questions about how to broker peace, and grand ethical theories are not particular suited to this task.

Violent conflicts between groups also have analogues within national boundaries, even when there is no civil war. Consider the battle between

Mexican drug lords and the Mexican government. Could the drug lords be criticized using grand ethical theories? Drug selling is an exercise of autonomy, it can increase pleasure in users, and it is a courageous defiance of laws. Or consider the American carceral state: 1 in 100 U.S. citizens is in prison, disproportionately black and poor, often as punishment for non-violent offences. Do grand theories help us see whether or why this is unjust? Do they shed any light on mass incarceration or point to a solution? Such theories are stated at such high-levels of abstraction that it is hard to know how to even begin putting them to work in the real world.

This is equally true when it comes to public policy. Consider morally contested laws concerning wealth redistribution, corporal punishment, marriage, and abortion. Much work in applied ethics tries to bring grand theories to bear on these issues, but the same theory can deliver different verdicts. Does corporal punishment reduce happiness overall or increase it? If there were a clear answer (which seems unlikely), would that settle the question of whether it should be tolerated? Do bans on gay marriage violate autonomy? Presumably yes, but Kant used his framework to condemn homosexuality, and also masturbation, restricting sexuality to procreative ends. Likewise, Aristotle has been interpreted as opposing abortion (when his outmoded embryology is updated), while also supporting infanticide. Bentham and Mill favoured free markets, but their ideas were adapted by reformers such as Robert Owen and J. A. Hobson to defend socialism. Finding a consistent message on public policy in grand ethical theories is difficult. Diverging sides of a policy debate will each claim that their position has good consequences and respects autonomy. All parties believe themselves to be virtuous, and thus assume that virtuous people would favour their perspective.

One obstacle for applicability is bad psychology. Leading ethical theorists recognize that there is a relationship between morality and human psychology, but they often build on psychological assumptions that are dubious, or make recommendations that are in tension with how human minds work. Aristotle says that we are by nature rational, and then characterizes virtues as governed by reason. But human beings are not perfectly rational: we are petty, hypocritical, impulsive, impatient, selfish, and passionate. Temperaments also differ from person to person, and there are cultural differences in views about which traits are constrictive of flourishing (e.g., do we flourish more when we are independent or interdependent?). Virtue Theory demands that we alter our predispositions in fairly dramatic ways. Consequentialists traditionally characterize human beings as hedonists, who work to increase pleasure, or at least to satisfy preferences. In reality, we often act arbitrarily, or out of habit, inculcation, laziness, compulsion, and caprice. We get fixed in routines, because repetition comes more naturally than change. We are self-destructive, rash, and fickle. We are ignorant of our desires, and bad at forecasting what will make us happy. Kant recognizes that each of us has personal urges and inclinations that may be incompatible with the universalizing demands of his theory. He nevertheless instructs us to bracket off those inclinations and adopt a more universal perspective. This sounds noble,

but it may be difficult, or even impossible: if we could bracket off preference, moral apathy might follow.

Unrealistic psychology and unrealistic demands make grand ethical theories utopian. It also deprives them of explanatory purchase on core aspects of human life. I noted that these theories cannot account for the moral conflicts that dominate headlines. By focusing on idealized conceptions of how we should be, these theories say too little about how we actually are. This suggests that we ought to look for ethical theories that make more effort to accurately characterize human psychology.

Normative versus descriptive?

I just suggested that grand ethical theories – like those of Aristotle, Mill, and Kant – do a poor job accounting for the moral conflicts that arise in the actual world. Against this, it will surely be objected that these theories are not intended for such explanatory work. Their aim is not to characterize how things are, but rather to specify how they should be. To use the standard jargon, these theories are normative, not descriptive. My plea for theories that are more psychologically accurate might appear to run afoul of this fundamental distinction. A theory that describes moral psychology well may provide no guidance about how we ought to act. After all, we cannot derive an ought from an is. Thus, such a theory would be even worse off than the grand theories I have been considering. It would provide no advice. It would account for the conflicts of human life without offering any solutions.

I grant that a descriptively adequate theory would not tell us how to act, but I want to nevertheless suggest that such a theory can contribute in important ways to normative inquiry, and would have advantages over the grand ethical theories. Three points deserve attention. First, and most obviously, any normative theory should be consistent with human psychology. A theory that makes unrealistic demands will violate the stricture that 'ought' entails 'can'. All the grand theories I have considered run this risk. Aristotle restricts virtue to a select view. Mill and Bentham reduce moral deliberation to a hedonic calculus in which we are asked to maximize ends regardless of the means, and Kant asks us to ignore personal preferences and rely on dispassionate reason alone. We must be saints, calculators, or zombies on these accounts.

Second, the failure of grand theories to explain social woes may also limit their capacity to provide solutions. In medicine, correct diagnosis is important for finding remedies. Accounts that do not provide realistic theories of human psychology offer little insight into why, for example, sectarian violence is so widespread. It is no surprise, then, that such accounts offer no special insight into how such violence can be addressed.

Third, it is important to accept that grand ethical theories may not be true. Each aspires to find some firm, universal foundation for morality, but they make dubious assumptions, and none enjoys anything approximating consensus support. The best minds in ethical theory pick sides and offer little hope for

adjudication. We cannot afford to wait for the right moral theory to magically emerge from centuries of dispute. We must operate under the assumption that even if there were a single true grand moral theory, we do not currently know what it is, and we are unlikely to find out.

This last point raises a crucial question: what can we do about moral conflict if we cannot rely on grand ethical theories? From the perspective of philosophy, this may look like an insurmountable problem. But, in the real world, it is a problem we face every day, and we look for practical solutions. Instead of relying on grand theories, we try to find ways for competing sides to get along in the absence of any over-arching truth. Doing this well would benefit from an accurate understanding of moral psychology.

One might put the point by saying that the distinction between normative and descriptive theories is based on a mistake. It is based on the premise that normative theories must be universal, and thus descriptive theories cannot play a normative role. Suppose we replace this notion of normativity with another one, grounded in practical reason. On this alternative conception, normativity – the investigation of what we ought to do – comes down to questions about how we can make human life go better in the real world. In cases of conflict, all sides tend to agree that things could be better. The exception would be cases where one side has a clear upper hand and is thriving at the expense of others. In that situation, the side that has been exploited or oppressed will recognize that things could be better. Thus, in every case, at least some parties involved feel dissatisfied, and they engage in efforts to improve their lot. This, I am suggesting, is a kind of normativity. It is not a transcendental precept about how to behave, but rather a deep aspiration, grounded in human experience, to live more comfortably. This kind of normativity does not eschew descriptive projects. It depends on a clear understand of how things are, how we want things to be, and how to bring these two into better alignment. Of course, the 'we' here can refer to different groups. Each of us has our own goals, and thus the normative project is also a project of coping with diversity. How do we (for each we) achieve our goals when some of those goals conflict with the goals of others? Normative questions in the real world are, to that extent, questions of social coordination.

Morality, emotion, and culture

The emotional turn

I have been suggesting that a moral theory for the real world must begin with a descriptive moral psychology. To understand and address moral conflicts, we must investigate the psychological processes that underlie moral decision-making and behaviour. While many philosophical ethicists have invested their energy in grand theories, others have been more concerned with the psychological basis of moral judgements. I want to begin with one strand in that tradition, and then describe how recent work in the cognitive sciences has added new detail and support.

The historical strand I want to take up begins with the British moralists in the eighteenth century. This, of course, was the time of Enlightenment. One consequence of the new outlook was a secularization of moral theories. Hobbes and Locke had already done much to secularize moral thinking in England during the nineteenth century, but their concerns were mostly with political questions: the legitimacy of sovereignty in the case of Hobbes, and the nature of natural rights in the case of Locke. At the turn of the eighteenth century, Locke's friend, Anthony Ashley Cooper, the third Earl of Shaftesbury, began to shift attention from the political to the psychological. In 1711, Shaftesbury (as he is known), published a compendious, though unsystematic work called *Characteristics of Men, Manners, Opinions, Times*, which includes an account of how people arrive at moral judgements. The account made central reference to 'sentiments' or 'affections' (what we now call emotions). More exactly, Shaftesbury identifies moral judgements with second-order sentiments: sentiments directed towards other feelings or thoughts that occur in one's self or in another person. For example, one might feel an inclination to be dishonest, and then, reflecting on that inclination, arrive at a feeling of disapproval. This would be the judgement that dishonesty is wrong.

The equation between moral judgements and sentiments was subsequently taken up by other authors, including a number of Scottish philosophers: Francis Hutcheson, David Hume, and Adam Smith. The tradition is now known as sentimentalism. Each author in the tradition has distinctive views. Hutcheson took up sentimentalist ideas, and posited a moral sense, analogous to vision, hearing, or touch, which uses emotion to recognize right and wrong. The term 'moral sense' had been used by Shaftesbury, but Hutcheson develops the analogy in more detail. Hume then rejects it. He agrees that moral judgements are based on sentiments, but he denies that these derive from a special sense, and says, instead that morality is an extension of non-moral sentiments that we acquire through learning. By nature, we have benevolent attitudes towards our near and dear (what Hume calls a natural virtue), but, though socialization, we come to extend our 'fellow-feelings' to strangers. We learn to take up the 'general point of view', feeling disapproval when anyone is poorly treated even if that person is neither friend nor kin.

Smith takes up where Hume left off, developing a more complete story of how we arrive at moral judgements. He begins with the notion of sympathy, also noted by Hume, which consists in our capacity to experience feelings that we see vividly displayed by others. In some cases, he notes, we fail to sympathize. These are cases where an emotion displayed by another runs contrary to how we ourselves would respond. Failure to sympathize is tantamount to a kind of disapproval, whereas successful sympathizing is a kind of approval. For Smith, moral judgement is constituted by the degree of sympathy we experience when we consider a situation from the perspective of an impartial observer.

In some ways, these sentimentalist theories resemble the grand ethical theories considered earlier. Shaftsbury advances a general theory of what is it to be good. Something is good to the extent that it contributes to the existence,

propagation, or well-being of the category to which is belongs; for example a human behaviour is good if it contributes to humankind. A human being is said to be a good or virtuous person if she or he makes those contributions 'primarily and immediately' – which is to say, not out of calculated self-interest or duress. In emphasizing the motives and traits behind good behaviour, and not just mentioning outcomes, Shaftesbury recalls Aristotle's emphasis on character. Hutcheson, in contrast, places emphasis on outcomes; and says the moral worth of an action is proportionate to the number of persons whose happiness is thereby increased. This anticipates Bentham and Mill. Smith then anticipates Kant in his emphasis on impartial observers; the good can be determined by adopting an objective perspective. The crucial difference between these authors and their counterparts is the effort they make to describe the psychological processes underlying moral judgement and their emphasis on emotions.

Hume stands out in the group in that he offers comparatively few normative claims. He does not advance a grand ethical theory. He does not derive a list of universal virtues, specify the most desirable consequences, or develop a universally applicable decision procedure. He gestures in all of these directions, but they never occupy a central position in his work. In this respect, Hume's project is more descriptive. His primary interest is in characterizing how moral judgements are made. He is least vulnerable to the charge of over-reaching. The others present encompassing normative theories that, in their abstraction and generality, would be difficult to apply to real-world problems.

If we follow Hume in foregoing the normative ambitions within the sentimentalist tradition, we can focus attention on the psychological claims put forward by these authors. We can ask whether they are right that moral judgements have a basis in our sentiments and, if so, how in particular our sentiments contribute. In the eighteenth century, such questions were broached by means of introspection and speculations. The authors pronounce on the psychology of morals with great confidence, but they provide little insight into how their conclusions are obtained. In the intervening centuries, we have made progress in psychology and can now move beyond the limits of introspection and test theories of how moral judgements actually arise.

Empirical evidence for sentimentalism

Over the last two decades, there has been a growing interest in empirically investigating the role of emotions in moral judgement. Here, I will briefly review some of that evidence (see also Prinz, 2007a). Recent research offers support for sentimentalism, while also adding details that went unnoticed in the eighteenth century.

Before the 1990s, psychological research on morality had focused on styles of moral deliberation. During these early years, emotions were rarely discussed. One of the most influential authors in empirical moral psychology was Lawrence Kohlberg. He had used interview methods to argue for a sequence of three main stages in moral development. The first stage includes reasoning based on

perceptions of character (e.g., 'anyone who does that is bad!'), and thus relates to the Aristotelian tradition. The second stage encompasses reasoning about outcomes and thus relates to Consequentialism. The third stage culminates with principles about our universal duties as persons, and thus relates to Kant. It is a central feature of Kohlberg's account that development recapitulates the history of ethics, though he manages to skip sentimentalism in his story. Though elegant, the neat correspondence between Kohlberg's stages and the history of Western ethics was also a source of suspicion. Critics quickly realized that there is little evidence that people generally reach the culminating Kantian stage. They also found that people often move back and forth between alleged stages, or use mixed styles of reasoning. Eventually, Kohlberg's views drifted from popularity for lack of evidence.

The most immediate successor to Kohlberg was the 'domain theory' developed by Eliot Turiel (1983) and collaborators. Turiel approaches moral psychology by examining how the domain or moral rules differ from other kinds of rules, especially social conventions. Three differences are identified: moral rule violations are considered more serious than conventional rule violations; moral rules are said to hold independent of authority, and moral rules are said to justified by appeal to empathy. Domain theorists do not emphasize emotion, but it is easy to see how their characterization of the moral domain might lend itself to a sentimentalist analysis. The final feature – appeals to empathy – is obviously compatible with sentimentalism, and it aligns with the theory of Adam Smith. The first feature, seriousness, is also easy for sentimentalists to explain. Why are moral violations regarded as serious? Because they elicit strong emotions. The second feature, authority independence, is a bit trickier, but it has also been explained in emotional terms. James Blair (1995) argues that there are strong associations between moral rules and emotions, and, as a result, whenever we imagine a moral violation, we find it upsetting, even if we are told that some relevant authority has deemed the violation permissible.

Much more can be said about these theories, but my goal here is not to review the literature on them, but to show how emerging dissent and reinterpretation of the data set the stage for an emotional turn in empirical research. In the 1990s, researchers began to explore the possibility that moral judgements are based on emotion. Blair (1995) contributed to this shift by arguing that emotional deficits can explain the moral insensitivity in criminal psychopaths. This suggests that moral competence requires emotions. Jonathan Haidt then began to provide evidence that people consult their emotions when making moral decisions. In some work, he and his collaborators induced disgust and found that it amplified perceptions of wrongness (Wheatley and Haidt, 2005; Schnall *et al.*, 2008). In other work, he showed that perceptions of wrongness can remain strong even when people cannot provide reasons for their negative judgements, suggesting that emotions, rather than reasons, are the final arbiters (Haidt, 2001).

Haidt and his collaborators also sought to update sentimentalist models by identifying specific emotions that contribute to moral judgement. Much of the early research focused on disgust, but a survey study suggested that anger and

contempt might also play important roles. In particular, evidence suggested that people associate disgust with crimes in which bodies are mistreated (e.g., sexual crimes), anger with crimes that involve violations of justice or rights (e.g., property theft), and contempt with crimes that threaten the social order (e.g., disrespect of the elderly). Follow-up work has confirmed that disgust and anger play these different roles (Seidel and Prinz, 2012*a*). Follow-up work on contempt has broadened the role of that emotion to cover cases of incompetence, such as when one tries to look smart and fails (Hutcherson and Gross, 2011). Such research suggests that anger, disgust, and contempt can all occur in response to moral transgressions, but each responds to a somewhat different (though often overlapping) range of cases.

Other work explores self-directed emotions. When we make moral judgements about the wrongdoings of others, anger, disgust, or contempt are likely to arise. But what if we make moral judgements about our transgressions? Here, guilt and shame seem to be the dominant emotions. Guilt is mostly likely to occur when we judge that our actions harmed another person, as when we physically hurt someone, and shame is most likely to occur when we judged that we used our bodies in ways that are regarded as wrong, as in the case of inappropriate sex acts (Prinz, 2011a). This wrongdoing can elicit different responses depending on whether the perpetrator is one's self or another person.

Another class of emotions seems to arise when we consider good moral behaviour. Stories about noble behaviour induce feelings of elevation (Haidt, 2003). Related findings have been obtained using emotion induction techniques: listening to uplifting music makes one more likely to judge that helpful actions are good and obligatory (Seidel and Prinz, 2012b). There may also be a self/other distinction when it comes to positive moral judgements, just as there is in the negative case. Others' noble actions are elevating, but what about our own good deeds? There is evidence that we respond to our own actions with pride (Etxebarria *et al.*, 2014).

This work on emotion differentiation enriches the traditional sentimentalist approach. It suggests that moral judgements are not based on generic feelings of approbation and disapprobation, but rather on a range of specific emotions. When I say that it was morally bad or morally good to perform a certain action, that statement expresses an emotion I am feeling, but the emotion depends on the case. Was I the one doing the action, or was it someone else? What kind of action was it? Moral judgements are constituted by different emotions depending on who did what to whom.

Work on emotional differentiation also bears on a debate that divided Hutcheson and Hume. Hutcheson believed in a moral sense, while Hume argued that our moral capacity derives from more general emotional dispositions that arise outside of the moral domain. The evidence linking moral judgements to emotions such as anger, disgust, contempt, elevation, and pride is more consistent with Hume's view, since each of these emotions can arise in in non-moral contexts. Even guilt and shame have non-moral applications: one can feel guilty about breaking one's diet, and ashamed about poor performance in

an exam. It seems that morality makes use of emotions that are not necessarily moral.

So far, in reviewing the empirical literature, I have ignored one emotional construct that was important to some eighteenth-century sentimentalists: sympathy. Recall that Adam Smith makes empathy the centrepiece of his theory. He says that moral judgements consist in the empathic responses. More accurately, to judge that something is good or bad is to feel the sympathy (or lack of sympathy) that an impartial observer would feel in response to the feelings of a person who had performed an action under consideration. Does this view find support in the contemporary empirical literature?

There is a considerable amount of work on sympathy. We now use the term 'empathy' for what Smith had in mind: feelings that are congruent with the feelings saliently displayed by others. Much of that work explores the role of empathy in moral motivation (e.g. Batson and Shaw, 1991). There is no work explicitly addressing the conjecture that empathy is a correlate of moral judgement, but there are reasons for doubt.

For one thing, there seem to be certain kinds of moral judgements that do not route through empathy. The classic example is judgements having to do with justice. We can recognize that certain things are unjust, and hence wrong, without thinking about how they make people feel. In fact, there is research suggesting that empathy and justice can actually come into conflict. Empathy promotes preferential treatment (we want to give special advantages to those with whom we identify), and justice opposes preferential treatment. Induction of empathy can make people insensitive to justice (Batson *et al.*, 1995). Using economic games, Leliveld *et al.* (2012) have shown that individual differences in empathy do not moderate perceived injustice, but simply effect how we compensate those who have been mistreated. Thus, empathy does not contribute to seeing that unfairness is wrong, but only to helping those who are adversely affected. There is also evidence that people who are low in empathy are more likely to make consequentialist as opposed to deontological (that is, roughly, Kantian) decisions in moral dilemmas (Duke & Bègue, 2015). That suggests that consequentialist reasoning is also independent of empathy. Deontological reasoning may not depend on empathy either. In some dilemmas, it reflects personal distress about causing harm rather than empathy for the victim (Sarlo *et al.* 2014).

Such considerations cast doubt on Smith's conjecture that empathy (or sympathy) is a component of moral judgement. Empathy may contribute to our concern for others' welfare and our motivation to help, but it is not necessary for making moral judgements (for more discussion, see Prinz, 2011b). Emotions such as anger, shame, and elevation are better candidates for this role.

In summary, empirical findings confirm key predictions of sentimentalism: our capacity to make moral judgements depends on emotions, and the strength of our judgements varies with the strength of the emotions we are experiencing. This suggests that moral judgements are grounded in emotions. Recent work supplements eighteenth-century theories by adding needed evidence and

by identifying the specific emotions involved. Empirical findings can also test competing versions of sentimentalism, and cast doubt on certain claims (e.g., Hutcheson's moral sense theory and Smith's sympathy account), while affirming the main thesis that moral judgements are emotional in nature.

Where moral emotions come from?

I have just been describing how psychological research can confirm and correct sentimentalist theories of morality. I want to end this section with another line of correction. The eighteenth-century sentimentalists all wondered where our moral sentiments come from. Shaftesbury and Hutcheson thought they were innate. Hume thought they were learned extrapolations of innate fellow-feelings. Smith falls somewhere in between these views, emphasizing an innate endowment, while also recognizing the impact of 'custom'. Crucially, even Hume and Smith see the role of learning as heavily constrained. All these authors think that our moral sentiments owe largely to human nature. They also hold the culturally inculcated divergence in our sentiments can be adjudicated (as we see in a moment). I think this outlook overestimates moral convergence, and underestimates the impact of culture.

It cannot be denied that our emotional dispositions are partially a consequence of biology. This is especially clear in the case of basic emotional responses that we share with non-human animals: fear of physical dangers, disgust at spoiled foods, sexual pleasure, and so on. When it comes to our more social emotional responses, such as love, respect, or envy, things are more complicated. Even granting that these are innate emotions, their application is heavily informed by culture. Does love lead to exclusive long-term partnerships? Does respect entail rigid social hierarchies? Should we envy wealth or despise it? In different cultural contexts, these questions receive different answers. Cultural beliefs and values can influence the application of emotions. Indeed, this is even true for the more basic cases. Should we fear casual brawls or fight recreationally? Is it disgusting to eat yoghurt, dogs, or intestines? What body-types are most sexually exciting? Human emotions are always culturally influenced.

The emotions underlying moral judgement are perhaps even more susceptible to cultural influence. Much has been made of the fact that certain aspects of morality have biological roots. For example, cross-species research on economic games suggests that apes and monkeys get upset when their conspecifics do not reciprocate in certain circumstances. But this tendency (which differs across species and experimental settings) places only a weak constraint on human behaviour. Cultures vary in what they regard as fair exchanges. Is the principle of reciprocity consistent with huge wealth disparities? Is it consistent with slavery and indentured servitude? Does it absolve us of responsibility to those who cannot reciprocate, such as people with profound disabilities? Cultures vary in their answers to such questions.

Strictures against killing, cruelty, incest, and selfishness vary widely and wildly (Prinz, 2007b). There are no moral universals. We may seem to find

universals when we state principles abstractly (e.g., 'incest is wrong'), but these abstractions are interpreted differently from place to place (e.g., is marriage between first cousins or half-siblings incest?). Many cultures will overlap, of course, and certain practices will be rare (e.g., parent/child marriages). But we cannot uncover 'natural' norms that would be universally recognized. Our biological dispositions are more like broad frameworks that must get filled in by culture to be applicable. Human life is cultural life. We live in constructed worlds. Everything around us is a result of ingenuity and invention, and it is the nature of our species to flexibly adapt to our circumstances. Our emotions are culturally tuned, and that means members of different cultures will generally have somewhat different values.

These values are shaped over historical time. The factors that have an impact are varied. Values might be changed by past environmental conditions, economic systems, political organization, power struggles, or external influence. Let me give a quick example of each. Attitudes towards infanticide tend to vary with environmental conditions. In places where population growth is not sustainable, there will be greater tolerance to infanticide, and girls will be more likely victims than boys, because they contribute more to population growth. Attitudes towards slavery tend to vary with economic systems. Hunter-gatherer societies cannot enforce or productively use slaves, but societies with large-scale agriculture can. Thus, slavery became widespread after the Neolithic revolution but dissipated with the industrial revolution, which switched to a wage model. Attitudes towards torture seem to vary with political organization. As nations become more democratic, torture is used less in judicial contexts. An absolute monarch can assert authority using torture, but a government elected by the people does not require assertions of authority to retain power, since the people are readily willing to grant authority to themselves. Democratic societies are more tolerant of torture towards outsiders, however, or those who are regarded as infiltrators, who threaten the democratic system. Power struggles have contributed to changing attitudes towards gender equality. The global wars of the twentieth century sent many young men to battle, diminishing the workforce in new industrial economies. Factories, including those that manufactured war machines, needed workers, and women entered the labour market as never before. It was harder for women to enter professions that were less economically vital or less dependent on large numbers – medicine, academia, politics – but economic independence paved the way for suffrage and reproductive rights, which have allowed for slow but steady empowerment, and changing gender norms. Foreign influence can be credited with moral change as well, along with the economic opportunities of foreign trade. One example is foot-binding in China. After 1,000 years, this practice came to a quick end in the twentieth century, as China came into greater political and economic contact with the West. Other examples are easy to multiply. Crimes of passion are more tolerated in societies where moral norms must be enforced by codes of honour rather than external policing; gay marriage is more tolerated in post-industrial societies where procreation is costly; communism arose from the

power struggles between social classes after the rise of industrial capitalism, and so on. Every deeply held value is a product of history.

In contemporary contexts, the historicity of values has an interesting consequence. Our societies are large and pluralistic. So the same nation, city, or town may include people with different cultural backgrounds. We each have our own histories. Many of us have multiple histories, which correspond to our different familiar lines and different places in which we have lived. As a result, many different value systems co-exist in the same place, or even within the same person. As a result, moral variation has become a central aspect of human life. I will return to this point below.

The picture that emerges is what philosophers call descriptive moral relativism: as a matter of fact, people have different moral values. Descriptive moral relativism is often contrasted with a more controversial view: normative moral relativism, which says that there is no single morality which deserves to be called the correct or true morality. One could be a descriptive relativist but not a normative relativist. One could admit, in other words, that values vary, while still insisting that only one set of values is correct.

Curiously, this seems to have been the position of both Hume and Smith. They astutely recognized that values differ across cultures, but they held out hope that such differences are superficial, and that really there is a universal set of values given to us by human nature, which can be used to settle disputes across cultures. Hume makes his case in a dialogue, which he published at the end of his *Enquiry Concerning the Principles of Morals.* The dialogue begins with a litany of observations about the morals of classical Athenian society. The Athenians, it is alleged, tolerated sibling marriage, homosexuality, infanticide, parricide, suicide, treachery, and torture. This suggests that morality is descriptively relative. Hume then asks whether we should conclude that there is no single true morality. Do Athenian values have equal claim to truth over the values favoured during the Scottish Enlightenment? He delivers a negative verdict. All people, he claims, share the same 'higher' principles, and simply draw different conclusions. For example, he surmises that the Greeks tolerated homosexuality because they believed that such relationships fostered friendship, sympathy, mutual attachment, and fidelity – values that are universal according to Hume. He regards this as a kind of understandable mistake. Greeks were right to value these things, but wrong to think that homosexual relationships are a good (or acceptable?) way to obtain them. He also remarks that some variation is a consequence of circumstances. For example, military valour is a useful virtue in a warring society, but not in a peaceful one. We can settle which values are right by using reason to deduce which values are more useful for obtaining our higher principles under a given set of conditions.

Smith offers a similar response to moral variation. He dedicates section V of his *Theory of Moral Sentiments* to this topic. Smith's central contrast is between the values of 'polite' societies (e.g., eighteenth-century Scotland) and 'savage' societies (e.g., indigenous peoples of the Americas). In some cases, Smith thinks the contrasts are exaggerated. For example, Europeans are horrified by the Native American

practice of binding and reshaping the scull, but this is no worse, he claims, than the European practice of forcing women to 'squeeze the beautiful roundness of their natural shape' into corsets (233). In other cases, he credits cultural variation to differences in circumstance. He stereotypes Native Americans as indifferent to each other's suffering, brazen in the face of death, and preoccupied with self-restraint. This is a result of the harsh conditions under which they live. Constant danger, hunger and hardship requires self-interest, fearlessness, discipline, and intolerance for excess. In such conditions, people cannot afford to indulge in strong emotions. In Europe, however, people have the luxury to express pains freely, and let themselves be overcome by love. Smith sees these tendencies as especially prevalent in France and Italy. Citing an example from the Abbé Du Bos, he says an Italian expresses more emotion on receiving a small fine than an Englishman would on receiving a death sentence. But, Smith notes, the English are becoming more expressive, because increasing prosperity allows such excess. These differences do not reflect an irresolvable moral divide across cultures, but rather reflect the way in which values adapt. The right values are those that are most suitable to the situation in which one lives. This opens up an avenue from rational criticism. We can denounce a society for having values that are not required by circumstance. Smith gives the example of Greek infanticide. They inherited this practice, he surmises, from barbarous times, when leaving a child to die might have been the only way to avoid starvation or allow escape from an enemy. In classical Greece, there was no such excuse.

In summary, Hume and Smith both provide ways of explaining moral divergence. In some cases, differences are acceptable applications of the same basic values to divergent conditions, and in others, differences are mistaken inferences or intolerable holdovers from previous times. In all these cases, we can say there is a single set of basic values that can determine whether a given practice is right or wrong. Thus, Hume and Smith avoid normative relativism. But their argument does not hold up under scrutiny. First, it is unclear what the stock of foundational values consists in. These are said to be our natural values, but humans, by nature, construct cultures introducing different forms of life. Hume mentions fidelity, but this is too abstract to be applicable outside of specific cultural contexts. To whom should we be faithful? To friends, family, lovers, partners in trade, political leaders, or group-members? Note that these can come into conflict, and nature cannot settle the ranking. Even if there were a most natural application of fidelity, it would have little bearing on the value of this attitude in cultural settings. Hume also mentions utility, but one can ask, utility for whom? Is a value useful if it benefits those who possess it? Or those in power? Or the group over all? Or the species? And we can ask, useful for what? Survival, happiness, and productivity are all candidates with different implications. Note too that whatever answer we pick will offer little guidance in moral debates. For example, two societies may come up with different rules that are equally useful when it comes to group success overall.

Smith offers little more help. He suggests that values are best when they suit our circumstances. First, note that this is actually a concession to normative

relativism, not a response to it. Smith effectively grants that, in different circumstances, there may be different moral requirements. This concession may not go far enough. What, after all, does it mean to say that a set of values is fit? Does it mean those values increase survival? Whose survival? To what degree? And what about well-being? Would a value that increased an individual's prospects be right for that individual? In which case, might an eighteenth-century Scott adopt the values that Smith (dubiously) attributes to Native Americans on the ground that selfish indifference to others is self-serving? Smith also implicitly assumes moral universal in his arguments, rather than establishing that such universals exist. He denounces Greek infanticide on the grounds that there was no need for it, but why assume that necessity is a precondition for justification? This itself is a moral principle (take a life only when necessary), and it begs the question to assume it in the context of a moral debate. Most societies tolerate some avoidable death (capital punishment, euthanasia, suicide, risky professions, collateral damage in war, and so on).

Hume and Smith rightly recognize that sentimentalism entails a kind of relativism. If values have an emotional basis, then values will vary across time and space, because emotions can be shaped by culture. They both try to limit the implications of this variation by suggesting that moral differences can be adjudicated or rationally justified. That is probably an overly optimistic view. In any case, we should at the very least accept descriptive relativism. This will be important for the next topic I will examine: the relationship between morality and identity.

Morality and identity

From feeling to being

For eighteenth-century sentimentalists, moral values are things that we possess. The metaphor of possession implies that values are like inherited property. They are things that we come to own, and they are, to that extent, external to us. This implies that values can be lost without greatly affecting their owners. They can be traded in for other values or abandoned. This way of looking at values is misleading. Sentimentalism implicitly suggests an alternative view about our relationship to values. It implies that values are not things we possess, but are rather part of who we are – part of identity. The British moralists did not draw this conclusion, but their sentimentalist theories raise this possibility.

The topic of personal identity is hotly disputed in philosophy. The human mind has many features but only some seem important to identity. Suppose I have a fleeting pain in my foot; that pain does not seem to be part of what makes me the person who I am, and, when the pain passes, I still seem to be the same person. Philosophers have proposed various theories of which of my current mental states and traits are part of me (synchronic identity) and what makes me the same person over time (diachronic identity). John Locke (1690/1975) proposed that memory is the key to diachronic identity. Each of us becomes a

coherent person over time because past and present are linked by memory. A related idea is that identity has to do with narratives; we weave the events of our lives into stories (Schechtman, 1996). Some Kantians prefer to think of identity in terms of agency; the self is constituted by those aspects over which I can exert rational control (Korsgaard, 1989).

I do not think it is constructive to think of these theories as competing with each other. Each theory rightly captures things that people consider important to identity. Questions about what makes me should not be answered by positing some deep metaphysical fact. Such questions are best answered by specifying what I and others in my social environment regard as important. Identities are things that we construct, not natural kinds stitched into the fabric of the physical universe. Memory, narrative, and agency all matter to identity. What I want to suggest now is that morality matters too, and it may even be more important than these other dimensions.

The link between morality and identity can be derived from the kind of sentimentalist theory that I have been presenting. For sentimentalists, moral values are grounded in emotions. The values I have are constituted by the moral judgements I am disposed to make, and those are constituted by my emotional dispositions. Emotional dispositions are linked to identity because they are analogous to personality traits. Personality is characteristically defined in terms of temperament. Traits of temperament – such as irritability, neuroticism, and extroversion – are emotional dispositions. Irritability, for example, is the disposition to get annoyed. As emotional dispositions, moral values can be regarded as part of our temperament. Temperament, or personality, is something that we tend to consider an aspect of identity personality as part of what makes us who we are, both to ourselves and others. Thus, morality is part of our personality.

Aristotle acknowledges this when he says that a moral theory should concern itself with cultivating good character. The idea that values are part of identity is implicitly captured by the fact that we sometimes use the phrase 'a person's character' to refer to their moral dispositions. For Aristotle, these dispositions take the form of virtues, and contemporary virtue ethics contrast character-based ethics with ethics based on principles of action. Sentimentalism exposes this as a false dichotomy. Principles of action function like character traits when they are implemented through emotional dispositions.

The emotional basis of morals also connects with identity in other ways. Emotions are states that we consciously feel. Emotional dispositions can be described as our ordinary ways of feeling about things. When a person's emotional reactions depart from their ordinary dispositions, we tend to say the person is not acting like herself. From a first-person perspective, when my emotional responses are off, I do not feel like myself. Emotions are also connected to how we act. Morals provide our normal ways of behaving in the world. This too might be regarded as part of identity. In addition, emotions can be regarded as the indicators of what matters to us. Things that do not elicit emotions are regarded with indifference. Things that elicit emotions are felt to matter. It is plausible that my identity is partially a function of what matters to me.

There has been a recent empirical effort to confirm that moral values are regarded as important to identity. In one series of studies, Strohminger and Nichols (2014) asked participants to think about the consequences of various kinds of deficits for identity. For example, they ask people to imagine a friend who is losing various faculties through the ravages of age: imagine your friend loses his memory, or language, or cognitive abilities, or perhaps there is a loss of personality. For each of these, participants are asked to what extent would your friend remain the same person after the onset of the deficit. Among these, they also ask what would happen if moral values were lost. Their key finding in this study, and in four others like it, is that loss of moral values is regarded as a major blow to the self. When people imagine a friend whose values have changed, they judge that there is a strong sense in which the friend is no longer the same person. Moreover, in all five studies, moral change is judged to be a bigger blow to the self than any other trait or faculty that was investigated.

Nichols and I obtained similar results (Prinz and Nichols, in press). We conducted a series of studies that directly compared moral identity to the theories of identity mentioned above: memory, narrative, and agency. All matter for identity, but moral continuity was judged to be much more important to identity than these other things.

Shared identity and signalling

A further link between morality and identity can be derived from the discussion of historicity, above. Moral values are products of history. The values we have depend on our cultural heritage. Outside of philosophy, heritage is widely recognized as important to identity. When people try to explain who they are to others, they often mention where they come from (e.g., where they were born, their parental nationalities, their religious background, and so on). Morality is part of that. When we reveal information about our heritage, we simultaneously convey information about our values. Each cultural trajectory to which we belong has an associated set of moral values. Because culture is part of identity, morality is as well.

One consequence of this is that moral identity can be shared with others. Leading philosophical theories of personal identity focus on individualist traits: memories, personal narratives, and agency. Morality is a collective aspect of identity. Members of cultural groups tend to share values in common. In fact, many of the most salient ways of classifying people carry strong moral associations: religion, region (rural vs. urban), political party, ethnicity, national heritage (especially if one belongs to a cohesive minority group), and even age and gender each correlate with certain kinds of values. A demographic profile will often suffice for accurately guessing someone's values. The aspects of social identity we care about most tend to have moral significance.

Moral cohesion is not restricted to groups with overt moral agendas, such as political parties and religious denominations. Consider communities of taste. Often broad moral orientation can be inferred from factors such as musical preferences, interest in fine art, favourite films and novels. If someone likes hip-

hop or punk rock, they might be more likely to be liberal than someone who likes country music or opera. More avant garde taste can correlate with more progressive political views. On the other hand, in some cultural settings, taste for high culture indicates high social class, which can correlate with conservatism. In addition, we draw inferences about values from what cars people drive, what clothes they wear, and what food they eat. In a recent study, I confirmed that all of these are used to infer moral values in an American sample (see Prinz, forthcoming).

Of course, the associations are contingent and local: there are exceptions, and they do not apply transnationally. They are fairly robust within a community, however, and that is important. There is extensive empirical evidence that we do not like to associate with people whose values differ from our own (Skitka *et al.*, 2005; Leach *et al.*, 2007). We use salient aspects of taste to signal our moral identities to others. By forming communities of taste, we increase the probability that we will fraternize with those who have similar moral identities.

The social nature of identity relates back to sentimentalism. Because morality is grounded in emotion, it is upsetting to be confronted with those whose values depart from our own. Members of different moral group are less likable, less predictable, and, it turns out, less cooperative (Iyengar & Westwood, 2015). Thus, we construct social worlds that divided into groups of morally similar individuals, and we make those boundaries more salient by forming associations between moral and non-moral traits.

Upshot: human life revisited

These observations about moral groups bring us back to where we began: the misalignment between grand ethical theories and human life. Philosophers often defend ethical theories that are too abstract to explain and resolve human conflicts. The world divides along boundaries of religion, class, political orientation, ethnicity, regional identity, and nation. Sentimentalism and the idea of moral identity can help diagnose these divisions.

The form of sentimentalism presented here makes sense of conflict in the following way. There are multiple moralities; these differ from each other and there is no universal moral principles to adjudicate such conflicts. People identify with their moral values and form groups on that basis. Moral groups who come into contact will disagree over policy, and each will feel like its own values are correct, regarding others as morally degenerate, evil, or confused. When moral groups come into conflict, both emotions and identity will be at stake, so conflicts will be heated and existentially threatening. These features will exacerbate the more general tendency to favour in-groups over out-groups, potentially resulting in violence.

Unlike grand ethical theories, sentimentalism predicts that there will be moral conflicts, dividing people along demographic lines. It may also be helpful in trying to find remedies. Grand ethical theories offer universal frameworks for resolution. That is not especially helpful: these theories are difficult to apply in specific cases and their universality is hard to defend. Indeed, were one to

criticize extant moral groups using the framework of a grand ethical theory, that would be equivalent to proposing a new moral group, and inter-group conflict would be the inevitable result. Instead, we should begin with the premise that there is no moral framework to settle differences, and work to find ways to allow for peaceful co-existence.

Here are some strategies suggested by this framework. First, governments that include morally diverse populations should strive for proportionality; winner-takes all solutions threaten the integrity of moral minorities. Second, power should be localized; within small geographic regions, there tends to be greater moral homogeneity. Third, there is an imperative to allow mobility; those whose moral views differ from people in power will be better off if they can migrate to like-minded communities. Fourth, relativism should be taught; danger is greatest when one group thinks it is in possession of the moral truth. Finally, in cases where a group tries to impose its will on others, we need solidarity; we should help persecuted moral groups even when we do not share their values, because moral imposition is a threat to us all.

These guidelines are no panacea, to be sure, but they may be more promising than the advice we get from grand theories. Also, they are revisionary, not self-congratulatory. We should worry when ethical theories conform too closely to prevailing values, since that is a sign that we are using philosophical rhetoric to confirm our own contingent prejudices. The foregoing guideless suggest that some of the most popular Western, liberal, academic moral projects are problematic. If relativist sentimentalism is true, we should be suspicious of globalization, the spread of secular democracy, universal human rights, big government, and confidence in reason. All of these ideals underestimate the intractability of moral diversity and the extent to which expansion of moral communities threatens the identities of moral minorities. The links between morality, emotion, and identity give rise to conflicts. The solution is not, *per impossible*, to sever those links but to respect them.

Acknowledgments

I am grateful to Sara Graça da Silva for ongoing encouragement, extraordinary patience, and helpful feedback. This chapter would not have come to fruition without her help.

References

Aristotle (350 BCE/1985). *Nicomachean Ethics*. T. Irwin, trans. Indianapolis, IN: Hackett Publishing Company.

Batson, C. D., & Shaw, L. L. (1991). 'Evidence for altruism: toward a pluralism of prosocial motives'. *Psychological Inquiry*, 2, 107–122.

Batson, C. D., Klein, T. R., Highberger, L., & Shaw, L. L. (1995). 'Immorality from empathy-induced altruism: when compassion and justice conflict'. *Journal of Personality and Social Psychology*, 68, 1042–1054.

Bentham, J. (1780/1996). *Principles of Morals and Legislation*, J. H. Burns & H. L. A. Hart (Eds.). Oxford: Oxford University Press.

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

Duke, A. A., & Bègue, L. (2015). 'Drunk Utilitarian: blood alcohol concentration predicts utilitarian responses in moral dilemmas'. *Cognition*, 134, 121–127.

Etxebarria, I., Ortiz, M. J., Apodaca, P., Pascual, A., & Conejero, S. (2014). 'Antecedents of moral pride: the harder the action, the greater the pride?' *The Spanish Journal of Psychology*, 17, E52.

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

Haidt, J. (2003). 'Elevation and the positive psychology of morality'. In C. L. M. Keyes & J. Haidt (Eds.), *Flourishing: Positive Psychology and the Life Well-lived* (pp. 275–289). Washington DC: American Psychological Association.

Hume, D. (1739/1978). *A Treatise of Human Nature*, P. H. Nidditch (Ed.). Oxford: Oxford University Press.

Hume, D. (1751/1998). *An Enquiry Concerning the Principles of Morals*, T. L. Beauchamp (Ed.). Oxford: Oxford University Press.

Hutcherson, C. A., & Gross, J. J. (2011). 'The moral emotions: a social-functionalist account of anger, disgust, and contempt'. *Journal of Personality and Social Psychology,* 100, 719–737.

Hutcheson, F. (1738/1994). *An Inquiry into the Original of our Ideas of Beauty and Virtue*. In R. S. Downie (Ed.), *Philosophical Writings*. London: J. M. Dent.

Iyengar, S., & Westwood, S. J. (2015), 'Fear and loathing across party lines: new evidence on group polarization'. *American Journal of Political Science*, 59, 690–707.

Kant, I. (1785/1998). *Groundwork of the Metaphysics of Morals*, M. J. Gregor, Trans. Cambridge: Cambridge University Press.

Kohlberg, L. (1969). 'Stage and sequence: the cognitive-developmental approach to socialization'. In D. A. Goslin (Ed.), *Handbook of Socialization Theory and Research*. Chicago: Rand McNally.

Korsgaard, C. (1989). 'Personal identity and the unity of agency: a Kantian response to Parfit'. *Philosophy and Public Affairs*, 18, 101–132.

Leach, C. W., Ellemers, N., & Barreto, M. (2007). 'Group virtue: the importance of morality (vs. competence and sociability) in the positive evaluation of in-groups'. *Journal of Personality and Social Psychology*, 93, 234–249.

Leliveld, M. C., van Dijk, E., & van Beest, I. (2012). 'Punishing and compensating others at your own expense: the role of empathic concern on reactions to distributive injustice'. *European Journal of Social Psychology*, 42, 135–140.

Locke, J. (1690/1975). *An Essay Concerning Human Understanding*, P. H. Nidditch (Ed.). Oxford: Oxford University Press.

Mill, J. S. (1863/2001). *Utilitarianism*. Indianapolis, IN: Hackett.

Prinz, J. J. (2007a). *The Emotional Construction of Morals*. Oxford: Oxford University Press.

Prinz, J. J. (2007b). 'Is Morality Innate?' In. W. Sinnott-Armstrong (Ed.), *Moral psychology, vol 1: Evolution of morals* (pp. 367–406). Cambridge, MA: MIT Press.

Prinz, J. J. (2011a). 'Sentimentalism and Self-directed Emotions'. In A. K. Ziv, K. Lehrer, and H. B. Schmid (Eds.), *Self-evaluation: Affective and Social Grounds of Intentionality* (pp. 138–154). Dordrecht: Springer.

Prinz, J. J. (2011b). 'Against Empathy'. *The Southern Journal of Philosophy*, 49, 214–233.

Prinz, J. J. (forthcoming). *The Moral Self*. New York: Oxford University Press.

Prinz, J. J., & Nichols, S. (in press). 'Diachronic identity and the moral self'. In J. Kiverstein (Ed.), *Handbook of the Social Mind*. London: Routledge.

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). 'The CAD triad hypothesis: a mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity)'. *Journal of Personality and Social Psychology*, 76, 574–586.

Sarlo, M., Lotto, L., Rumiati, R., & Palomba, D. (2014). 'If it makes you feel bad, don't do it! Egoistic rather than altruistic empathy modulates neural and behavioral responses in moral dilemmas'. *Physiology & Behavior*, 130, 127–134.

Schechtman, M. (1996). *The Constitution of Selves*. Ithaca, NY: Cornell University Press.

Schnall, S., Haidt, J., Clore, G., & Jordan, A. (2008). 'Disgust as embodied moral judgment'. *Personality and Social Psychology Bulletin*, 34, 1096–1109.

Seidel, A., & Prinz, J. J. (2012a). 'Sound morality: irritating and icky sounds influence opinion in divergent moral domains'. *Cognition*, 127, 1–5.

Seidel, A., & Prinz, J. J. (2012b). 'Mad and glad: musically induced emotions have divergent impact on morals'. *Motivation and Emotion*, 37, 629–637.

Shaftesbury (1711/2000). *Characteristics of Men, Manners, Opinions, Times*, L .E. Klein (Ed.). Cambridge: Cambridge University Press.

Skitka, L. J., Bauman, C. W., & Sargis, E. G. (2005). 'Moral conviction: attitude strength or something more?' *Journal of Personality and Social Psychology*, 88, 895–917.

Smith, A. (1759/2002). *The Theory of Moral Sentiments*. Cambridge: Cambridge University Press.

Strohminger, N., & Nichols, S. (2014). 'The essential moral self'. *Cognition*, 131, 159–171.

Turiel, E. (1983). *The Development of Social Knowledge: Morality and Convention*. Cambridge: Cambridge University Press.

Wheatley, T., & Haidt, J. (2005). 'Hypnotically induced disgust makes moral judgments more severe'. *Psychological Science*, 16, 780–784.

2 Weakness of will and self-control

The role of emotions in impulsive behaviour

Vasco Correia

Introduction

According to the standard definition, 'weak-willed' (or akratic) action is free, intentional action contrary to one's better judgment (Davidson, 2001; Mele, 2012; Stroud, 2010). Some of the most common examples of this phenomenon include the smoker who is unable to quit, the dieter who yields to temptation, the procrastinator who puts off urgent tasks, and the person who cheats on his or her spouse despite a genuine intention to remain faithful. As Davidson observes, 'in such cases we sometimes say [the agent] lacks the willpower to do what he knows, or at any rate believes, would, everything considered, be better' (ibid., 21). And here lies the paradox and the conundrum of weakness of will: the intemperate person typically knows that option A (e.g., lose weight) is better, all things considered, than option B (e.g., eat candy) – or, in utilitarian terms, that A would maximize her well-being or utility – and yet, for some reason, chooses to do B instead.

This inconsistency between practical judgment and action is what raises the problem of irrationality. While there is nothing irrational about the decision to smoke cigarettes, for example, if one considers that the pleasure of smoking is worth the risks it entails, it is arguably irrational to continue to smoke if one decides, after mature deliberation, that it would be better to stop smoking. According to Aristotle, this conflict between judgment and action is precisely what distinguishes the 'weak-willed' person (*akrates*) from the outright 'depraved' person (*akolastos*): whereas the latter chooses the wrong course of action because his or her judgment is wrong, and does not regret it; the former does so despite knowing that it is wrong, and therefore tends to regret it (Aristotle, 2000: 132). The depraved person who lives a life of excesses and vice may be considered *immoral* by other members of society, but provided he or she acts in accordance with his or her own judgment, preferences and values, there are no grounds to deem his or her conduct *irrational*. The weak-willed person, on the other hand, acts contrary to what he or she believes to be in his or her best interest, and it is this inconsistency that seemingly renders the conduct irrational.[1]

Weakness of will remains both a practical and a theoretical problem. From a practical perspective, as Mele points out, 'it is an impediment to the achievement

of certain of our goals, to the execution of certain of our projects and intentions' (Mele, 1987: 3). Although intemperance can serendipitously bring about beneficial outcomes, it seems reasonable to assume that most people would rather stick to their good resolutions and resist the temptations that undermine them. The ability to do so roughly corresponds to Aristotle's definition of self-control (*enkrateia*). In a minimal sense, at least, self-control can be defined as the ability to conduct oneself in accordance with one's stable preferences in the face of the temptation to do otherwise. In other words, an agent is said to be self-controlled if he or she does not act contrary to his or her better judgment. The million-dollar question, in that regard, is of course: how can self-control be acquired? In which ways can we enhance the consistency between our actions and our decisions?

One cannot solve the practical problem of self-control without addressing the theoretical problem of weakness of will in the first place; that is to say, without explaining why people often act contrary to what they judge better. The methodological assumption here is that the normative question 'How is self-control possible?' presupposes the descriptive question 'How is weakness of will possible?' In other words, only a correct understanding of the phenomenon of weakness of will can pave the way for the development of effective methods of self-control. This is why I will begin by revising the traditional account of weakness of will (second section), which hinges on the notion of 'willpower' and tends to describe impulsiveness as the outcome of a battle between the will and the desires. I argue that this view is mistaken and fails to account for the causes of impulsiveness. In particular, it falls short of explaining what the agent's willpower (or 'strength of will') depends on, and also why it appears to be 'stronger' in some individuals than in others. This, in turn, will lead me to question the very notion of Will understood as a free and immaterial faculty of decision.

In the third section, I show that weakness of will (or akrasia) can be fully understood – and the alleged 'paradoxes of irrationality' dissipated – from the standpoint of a Belief-Desire theory of action, which highlights the decisive role of emotions in the process of reviewing decisions. According to this approach, weakness of will is better understood as the result of a conflict between competing desires, rather than a conflict between the so-called willpower and the desires. Drawing on Ainslie's (2001, 2005) theory of 'hyperbolic discounting', I argue, in the fourth section, that in many cases weakness of will is prompted by a specific cognitive bias that affects the agent's practical judgment and leads to a temporary preference reversal. Finally, in the fifth section, I contend that, if this model of irrational behaviour is correct, the most effective way to improve self-control is what decision-theorists call 'precommitment', i.e. self-imposing constraints on future options in order to be less likely to yield to small, immediate rewards (e.g. smoke a cigarette, eat candy, etc.), and more likely to stick to larger, later rewards (e.g. feel better, live longer, etc.). I then succinctly describe some of the self-control strategies that can plausibly enhance the rationality of our choices.

The myth of willpower

What the ancient Greeks called *akrasia* – literally 'lack of strength' – is traditionally explained as a result of a 'weakness' of the agent's Will. At first glance, this account may seem straightforward and even intuitive: whenever an agent decides to do A in light of all available relevant reasons but cannot resist the temptation to do B, it is presumed that his Will is too 'weak' to stick to the initial decision. This explanation of akratic behaviour was introduced by the Stoics and later developed by the likes of Aquinas, Augustine, Descartes, Kant, and more recently Holton (2003). Descartes, in particular, defines the 'strength of the soul' (*la force de l'âme*) as the ability to resist the passions that are opposed to the rational judgment, and claims that 'any soul, however weak, can if well-directed acquire absolute power over its passions' (Descartes, 1989: 47). In a similar vein, Kant goes as far as suggesting that 'to be subject to affects and passions is probably always an illness of the mind, because both affect and passion shut out the sovereignty of reason' (Kant, 2006: 149).

In the article 'How is weakness of will possible?', Davidson rightly observes that this account tends to conceptualize 'the Will' (or willpower) as a subjective faculty or power torn between the recommendations of Reason and the urges of Desire:

> Here there are three actors on the stage: reason, desire, and the one who lets desire get the upper hand. The third actor is perhaps named "The Will". It is up to The Will to decide who wins the battle. If The Will is strong, he gives the palm to reason; if he is weak, he may allow pleasure or passion the upper hand.
>
> (Davidson, 2001: 35)

The advantage of this model is that it renders agents morally responsible – and therefore accountable – for their irrational behaviour. Thus, when a person acts contrary to her better judgment, it makes sense to blame her for not fully exerting her willpower or for 'not trying hard enough'. For example, if Linda lights a cigarette a few days after deciding to quit smoking, it is her (weak) willpower that is to blame. And if Thomas sleeps with another woman despite his intention to remain faithful to his wife, it must be that his willpower was not strong enough to resist that temptation. At any rate, it is the person's will that is to blame in cases of self-indulgence, or to praise in cases of self-control.

The difficulty with this paradigm, however, is that it falls short of explaining what sort of principle governs the so-called willpower (or 'strength of will'). It remains unclear, in particular, why it is that the willpower is said to be 'weaker' in some individuals than in others. Why is Linda's willpower too weak to allow her to quit smoking, for example, whereas her friend Thomas managed to quit without too much difficulty? To qualify Linda as 'weak-willed' and Thomas as 'strong-willed' is not a consistent criterion, given that the same individual can be strong-willed with respect to some decisions and weak-willed with respect

to others. Thomas, for example, was strong enough to resist the temptation of smoking, but not the temptation of sleeping with another woman. This leads us to a second question. How come the same person's willpower appears to be 'weak' on some occasions and 'strong' on others?

Holton (2003) claims to have an answer to this question. According to him, the faculty of willpower is 'something like a muscle' in the sense that it can be developed through hard work and exercise. '[The] strength of will', he writes, 'is standardly achieved by its exercise' (ibid., 67). The meaning of the metaphor is easy to grasp: the more often agents succeed in exerting self-control, the easier it becomes to maintain their resolutions in the future. Like the athletic person who can lift heavy weights almost effortlessly, the strong-willed agent is able to resist strong temptations and stick to pondered decisions without even trying too hard. Yet, suggestive as this image may be, it does not explain what the 'strength of will' consists of *non-metaphorically*, nor does it account for the question of why some agents' willpower is stronger than others. Holton surmises that the agent's willpower may ultimately depend on his 'motivation to employ that faculty' (ibid., 60). But the question then becomes: what does the motivation to employ the faculty of willpower itself depend on? As Elster points out, the idea that the willpower is a sort of 'mental muscle' may be misleading, insofar as it purports to describe in physical terms a metaphysical instance such as the Will (Elster, 2007: 13).

In *The Concept of Mind*, Ryle goes one step further and labels both as a 'myth' and as a 'dogma' the assumption that there is a Will: 'I hope to refute the doctrine that there exists a Faculty, immaterial Organ, or Ministry, corresponding to the theory's description of the *Will*' (Ryle, 1949: 61–62). Although we commonly use the notion of Will as if it were self-explanatory, Ryle warns that it is a purely metaphysical concept that does not refer to any tangible reality. Much like the notion of soul, he writes, the notion of Will is 'just an inevitable extension of the ghost in the machine' (ibid., 62). Perhaps it is also worth noting that the ancient Greeks did not even have a notion of Will similar to ours. It is Cicero (2002) who later introduces the Latin word *voluntas* to translate the Greek term *boulesis*.[2] However, this translation is questionable, insofar as the term *boulesis* refers in principle to a 'rational desire', i.e. a specific type of desire (*orexis*) among others. In fact, both Aristotle (2010) and Plato (2004) put it on a par with other types of desire, such as the 'physical desire' (*epithumia*) and the 'emotional urge' (*thumos*). Rather than a 'free will' differentiated from (and often opposed to) the desires, such as Descartes and Kant would have it, the *boulesis* is described by Aristotle and Plato as a rational desire that competes with less rational desires in the process of deliberation. Nowhere, at any rate, do they conceive the existence of a separate and immaterial power such as the Will. Consequently, as Mele observes, what Aristotle and Plato called *akrasia* would not be correctly translated as 'weakness of will' (Mele, 2012: 2). The Greek term *akrasia* simply meant 'lack of power' (*kratos*), in the sense 'lack of power over oneself' or 'lack of self-control', without reference to the notion of Will understood as a supreme and undetermined power of decision-making. As

we have seen, the later image corresponds to the Stoic model, which Descartes, Kant and other Christian philosophers later developed. Be that as it may, according to Ryle 'this traditional dogma is not only not self-evident, it is such a welter of confusions and false inferences that it is best to give up any attempt to re-fashion it' (ibid., 61).

To sum up, the Stoic model of the 'Will' fails to account for the phenomenon of akrasia. It does not explain why the so-called willpower is stronger in some individuals than in others. It does not explain why one and the same individual appears to be 'strong-willed' on some occasions and 'weak-willed' on others. And it falls short of defining the notion of Will in non-metaphorical terms. What exactly do we refer to when we use the word Will (or willpower)? What does it mean to say that a person's willpower can be 'strong' or 'weak'? And what does the 'strength of will' depend upon, i.e. by virtue of what mysterious principle does it tend to oscillate?

So long as we consider the problem of akrasia in the light of what Ryle calls the myth of willpower, it remains difficult to understand why we often act contrary to our better judgment. Like St Paul, we are condemned to remain in a state of perplexity in the face of our own self-indulgence: 'My own behaviour baffles me. I find myself doing what I hate, and not doing what I really want to do!' (St Paul, Romans: 7.15). Perhaps more importantly, this approach has proven sterile when it comes to providing solutions to the problem of intemperance. Insofar as the rationality of the agent's behaviour is said to depend ultimately on his or her 'strength of will', and insofar as it remains unclear what the strength of will itself depends on, the only advice the intemperate person can be given is 'You must be stronger!' or 'You must try harder!', to use common expressions, although she already knew this at her moment of weakness.

The Belief-Desire theory of action

Although the Stoic account of the Will remains highly influential, not only among philosophers but also in common-sense psychology, there is an alternative account of the Will which seems able to overcome these difficulties and fully explain irrational action. This account was developed by the likes of Hobbes, Spinoza and Hume, who all insisted that it is an illusion to conceive the will as an immaterial faculty of choice opposed to the 'carnal' emotions and desires. Instead, they proposed to consider the Will as something very tangible and specific, namely: the sum of individual desires and other emotions that arguably determine us to act. In this view, the so-called Will is in fact reducible to the set of *volitions* (desires and emotions) that constitute the agent's motivation to act. As Spinoza so often stresses in *Ethics*, the Will as an absolute and undetermined power is a construction of people's imagination. The only instances capable of causing us to act are our desires, our fears, our ambitions, our hopes, our angers, our jealousies, and so forth, which simultaneously *move us* and *make us move*: 'In the mind there exists no absolute faculty of willing or not willing. Only individual volitions exist' (Spinoza, 2001: 88).

More recently, the proponents of the *belief-desire theory of action* – Davidson (1985), Mele (2001), Searle (2001), Lewis (2000) and Goldman (1976), among others – have developed a similar model, although these authors also highlight the role of beliefs (or judgments) in the process of deliberation. According to this view, it is both a *desire* and a *belief* that are needed to explain the agent's action. For example, Linda's motivation to quit smoking depends not only on her desire to quit, but also on beliefs such as 'Smoking may cause cancer' or 'Non-smokers tend to live longer'. There are of course many versions of this theory, but according to Humean accounts, desires have a more prominent role than beliefs in the causation of action. As Lewis observes, 'a Humean thesis about motivation says that we are moved entirely by desire: we are disposed to do what will serve our desires according to our beliefs' (Lewis, 2000: 42). In particular, a relevant belief may be decisive inasmuch as it is liable to affect the intensity of a specific desire. My desire to eat an exotic fruit, for example, utterly vanishes if someone informs me that it is in fact poisonous. But an evaluative belief such as 'I ought to stop smoking' is causally insufficient to offset the desire to smoke. Only the opposite motivation, such as the fear of cancer (or the desire to live longer) can effectively lead a smoker to quit his habit. 'Reason alone can never produce any action', Hume explains, 'Nothing can oppose or retard the impulse of passion, but a contrary impulse' (Hume, 1985: 462).

The most significant implication of this model is that it construes motivational conflicts in terms of an opposition between competing desires, rather than an opposition between the Will, on the one hand, and the desires, on the other hand (Cicero, Descartes, Kant, Holton, etc.). This hypothesis, in turn, provides a perfectly clear explanation of why people often lack self-control: an agent is typically 'weak-willed' with regard to the initial decision to do A if the motivation to do A is weaker than the competing motivation to do B. Thus, for example, if Linda is unable to refrain from smoking despite her resolution to quit smoking, it is not because her willpower is too 'weak' – whatever that means – but simply because her fear of cancer (or desire to live longer, etc.) is weaker than her desire to smoke. Maybe one day her fear of cancer will be stronger than her desire to smoke, in which case Linda will eventually stick to her resolution and be praised as 'strong-willed'. But until then, there is little chance that her behaviour will change. She may, of course, *claim* that she is determined to quit smoking, and sincerely believe that she will succeed, but that does not mean that her motivation is effectively sufficient. After all, we can be mistaken about our motivation just like we can be mistaken about the strength of our muscles. Likewise, the fact that Thomas cheated on his wife simply translates the fact that his desire to sleep with someone else was stronger – albeit for a brief moment – than his desire to remain faithful to his wife. At any rate, the struggle within the agent's mind takes place between competing desires, not between the alleged 'willpower' and the desires. This is why Hume avoids the expression 'strength of will', preferring to use instead the term 'strength of mind', which he defines as 'the prevalence of the calm passions above the violent' (Hume, 2007: 24).

If this hypothesis is correct, there is nothing paradoxical or incomprehensible about the phenomenon of lack of self-control. Assuming that our actions are neither determined by the inclinations of the will, as the Stoic model claims, nor by the judgments of reason alone, but by the host of desires, fears and other emotions that together constitute our motivation to act, it is barely surprising that we sometimes act contrary to what we judge better. One thing is to have good *reasons* to do A despite the temptation to do B, quite another is to effectively *desire* A more than B. As Bird suggests, once we cease to assume that our judgments have the ability to determine our actions, it no longer seems paradoxical that people sometimes act contrary to their better judgment:

> Typically, then, weakness of will consists in acting in accord with the preponderance of my desires, when the preponderance of good reasons for acting suggests acting differently. There is no paradox, for there is no reason why my desires should match my good reasons for acting.
>
> (Bird, 1994: 31)

Hume and Locke have also insisted at length on this point: it is not enough to *judge* that a given course of action is the best all things considered, nor is it enough to (claim to) *want* that option more than anything else; one must above all *desire* it more than any alternative option at the moment of choice. Locke writes: 'the greater good, though apprehended and acknowledged to be so, does not determine the will until our desire, raised proportionally to it, makes us uneasy in the want of it' (Locke, 1997: 234).[3]

One may also question, from this standpoint, whether it is always *irrational* to be intemperate with regard to a given resolution. Granted, the 'weak-willed' agent fails to act in accordance with what he or she judged better initially. But, presumably, this is the case because his or her preferences have evolved between the moment of decision and the moment of action. After all, as the old Latin proverb has it, 'Many things fall between the cup and the lip' (Multa cadunt inter calicem supremaque labra). At the moment of action, the agent's choice seems to be in conformity with his or her (new) preference. To that extent, it would be difficult to argue that the agent's action transgresses the utilitarian criterion of rationality, according to which actions are rational if they maximize the fulfilment of the one's preferences and goals. On the contrary: according to standard rational choice theory, 'people do reveal their preferences through their actual choices' (Sen, 1986: 66). Regardless of what an agent claims he or she wants or thinks best to do, if he or she chooses A instead of B one must assume that, at that point in time, he or she effectively prefers A to B. One may thus consider the agent's action to be rational, to the extent that it is consistent with his or her present preferences.

Furthermore, it may occasionally happen that the agent's transgression of the initial decision proves serendipitously beneficial. For example, Thomas' infidelity may have led him to realize that he and his wife were no longer happy in their marriage and to reconsider his options in a new light. Assuming

that Thomas' behaviour was neither inconsistent with his preferences nor detrimental to his well-being, why should it be considered irrational?

Weakness of judgment

In other cases of akrasia, however, there seems to be an inconsistency between the agent's choice and the agent's preferences *at the very moment of action*. This is particularly noteworthy in cases of backsliding, when the agent knows *by experience* that a given option is less advantageous than the initial decision, yet cannot help but repeating the same mistake. The smoker who relapses and the person who is unable to follow a diet are paradigmatic examples of this phenomenon. Unlike Thomas, who chooses to transgress his initial decision without being certain that the outcome will be ultimately negative, Linda *knows* in advance that she would be a happier person if she managed to quit smoking. In fact, she is seemingly aware of this at the very moment she lights another cigarette. In such cases, the agent's behaviour may be deemed irrational, inasmuch as he or she chooses the least advantageous option *in full awareness*.

Yet, how is this possible? Assuming that Linda genuinely believes she would be better off without smoking, and assuming that her motivation not to smoke is effectively stronger than her motivation to smoke, one would expect her behaviour to be consistent with her resolution. How come she intentionally chooses to do otherwise? Is the desire to smoke an irrational impulse that compels her to do something she knows she will regret?

Hume provides us with an answer to this question. In his view, desires and emotions are not irrational per se, but only insofar as they are associated to irrational judgments: 'A passion must be accompanied with some false judgment in order to its being unreasonable; and even then it is not the passion, properly speaking, which is unreasonable, but the judgment' (Hume, 1985: 463). According to this hypothesis, the irrationality we tend to ascribe to our desires, and subsequently to the actions that such desires motivate, stems in fact from the irrationality of our judgments. More specifically, Hume is referring to the *evaluative judgments* through which we assess the expected value of each available option, i.e. the foreseeable levels of pleasure or pain that each option seems to entail (either at a mental or physical level). Assuming, with Hume, that 'it is from the prospect of pain or pleasure that the aversion or propensity arises towards any object' (ibid., 661–462), it follows that an irrational appraisal of the expected hedonic value of a given option may result in an irrational desire of that option, which in turn leads to an action that is contrary to the agent's self-interest.

Some decision-theorists contend that this phenomenon is due to a particular cognitive illusion – known as the 'hyperbolic discounting bias' – which may be described as a sort of 'myopia' regarding future preferences (Ainslie, 2001; Elster, 2007; Strotz, 1956; for a review, see Loewenstein *et al*., 2003). Loosely speaking, the idea is that people tend to overrate the instrumental value of immediate options and, conversely, to underrate the instrumental value of future options.

Drawing on a series of empirical studies, Ainslie argues that '[people] tend to prefer smaller, earlier rewards to larger, later ones temporarily, during the time that they're imminent' (Ainslie, 2001: 38). Moreover, Ainslie's research shows that the curve describing the evaluation of future goods proportionally to their delay is *hyperbolic*, rather than exponential, which implies that 'the smaller reward is temporarily preferred for a period before it is available' (ibid., 32). Thus, when the motivation to defer gratification is insufficient, agents end up giving in to temptation and choosing an immediate small reward (e.g. a cigarette, a piece of cake) rather than a future greater reward (e.g. a longer life, an elegant silhouette).

To that extent, in many cases of weakness of will, the problem is not that the agent acts contrary to his or her own judgment, as Davidson (1980, 1985) claims, but more exactly that he or she acts on the basis of a *temporarily biased judgment*, presumably due to the influence of a strong desire. As Elster (1999) points out, Davidson's mistake is to assume that the akratic agent knowingly chooses to act contrary to his or her better judgment, which would indeed entail some form of synchronic inconsistency between the agent's attitudes. Instead, akrasia seems to involve a mere *diachronic inconsistency* between the agent's attitudes, and more specifically a temporary reversal of his or her preferences. For example, after deciding to start a diet, John enters a restaurant at Time 1 with the certainty that he prefers the greater long-term reward of losing weight to the immediate, smaller reward of eating a dessert. Yet, Elster explains, when John finishes the main course, the imminent availability of the smaller reward may affect his evaluative judgment, and for a short moment the reward of eating a dessert may appear superior to the later reward of losing weight:

> As the meal progresses, a preference reversal occurs at time $t^{\star}$, and when the waiter asks him, at Time 2, whether he wants to order dessert he answers in the affirmative … He is not, however, acting against his better judgment *at the time of ordering dessert*.
>
> (ibid., 430)

This hypothesis explains why we so often regret our indulgencies in hindsight. The reason we often come back to our senses right after having succumbed to temptation is now obvious: once the urging desire is satisfied, it ceases to affect our practical judgment (given that desires fade away as soon as they are satisfied). After eating the dessert – say, at Time 3 – John's craving dissipates and no longer distorts his assessment of the two available options. He is then able to see distinctly that it would have been in his best interest to abstain from the immediate pleasure of eating dessert. And this is why he bitterly regrets his choice, for he now realizes that it was not worth it. Interestingly, this interpretation of akrasia fits well with Aristotle's analogy between the agent who lacks self-control (*akrates*) and 'the person who is intoxicated or asleep': they both seem to temporarily 'forget' something they usually know (Aristotle, 2000: 125).

Finally, this model also accounts for the fact that most people are sometimes impulsive and 'weak-willed', *but not always*. After all, most people frequently abstain from immediate, smaller rewards for the sake of delayed, larger rewards. Most people wake up every day early to go to work, for example, and put away money for their retirement instead of spending it all at once. Like the Ant in La Fontaine's tale, the *Homo economicus* spends a vast amount of time 'sacrificing' immediate pleasures for the sake of long-term gains. The problem with this principle of rationality is that it only works when the delayed rewards are significantly larger than the immediate rewards. In such cases, the motivation to secure the long-term goals suffices to offset the desire for immediate gratification. When the delayed reward is only slightly superior to the immediate gratification, however, the bias of hyperbolic discounting may create the illusion that the nearest reward is greater. To be sure, the illusion is only temporary, but it is often enough to lead the agent to transgress his or her good resolution.

If this analysis is correct, akrasia is better understood as a 'weakness of the judgment' rather than a 'weakness of the will'. The intemperate agent typically chooses to reconsider his or her initial decision due to a biased evaluative judgment that leads to a temporary preference reversal. The problem is neither that the Will is too 'weak' nor that the agent acts contrary to his or her own judgment, but that he or she acts on the basis of an irrational appraisal of the available options. To that extent, it seems reasonable to conclude that practical irrationality typically stems from cognitive irrationality, and more specifically from what psychologists call 'motivated reasoning', i.e. the influence of desires and other emotions on judgment, reasoning and decision-making.

Self-control by precommitment

What are the implications of this account of intemperance (akrasia) when it comes to the problem of self-control (enkrateia)? It goes without saying that any efficient method of self-control needs to be based on a correct understanding of the causes of intemperate action. We have seen, in particular, that merely appealing to a hypothetical willpower (or 'strength of will') may not be the most effective strategy to improve one's ability to stick to resolutions. In contrast, if we accept the hypothesis that intemperance stems either (1) from a sheer lack of motivation or (2) from a cognitive illusion and a judgment bias, it becomes possible to take precautions in order to enhance self-control.

It is important to grasp, however, at the risk of sounding redundant, that self-control ultimately depends on the agent's motivation. According to the Desire-Belief theory, at any rate, a decision cannot be successfully maintained if the agent does not desire it more than any competing alternative. If Linda decides to give up smoking but is not motivated enough to do so – i.e. if her desire to be healthier (or her fear of cancer, etc.) is not stronger than her desire to smoke – no magical 'strength of will' or 'effort of willpower' will be of any help. It is a myth to believe that our motivation depends directly on our voluntary efforts:

we cannot decide to desire something (or not to desire it), just like we cannot decide to fall in love at will or stop fearing something by snapping our fingers. Desires and emotions *are* what we call the Will and, consequently, it makes little sense to say that they can be 'determined by the Will' or that our voluntary efforts can enhance our motivation. Linda may well *think* that she would be better off without smoking, all things considered, but she will find herself unable to act accordingly so long as her motivation to quit is not stronger than her desire to smoke. Self-control depends on people's motivation, but people's motivation does not depend on them. Hence Hume's famous statement that 'Reason is, and ought only to be the slave of passions' (Hume, 1985: 462), or Spinoza's similar claim that 'a man is necessarily always subject to passions' (Spinoza, 2001: 167). Reason can help us find effective means to our goals, but we do not get to decide which goals we happen to desire the most.

On the other hand, when impulsiveness is not due to a lack of motivation, but to a temporary lack of discernment (hyperbolic discounting bias), there are numerous strategies one can adopt to prevent it. A plausible way of counteracting this phenomenon is to self-impose specific constraints meant to mitigate the effects of judgment bias in future actions. Rather than being optimistic about one's own ability to stick to resolutions and resist temptations, it may be wiser to predict the possibility that a judgment bias might occur at some point, leading in turn to a preference reversal and to an impulsive action. According to this view, self-control strategies should take into account people's propensity to fall prey to judgment biases in contexts of uncertainty and forestall their impact on the process of decision-making. In a sense, one may say that such strategies are designed to coerce the agent's *future self* into honouring his or her *present self*'s decisions for the sake of promoting long-term well-being.

This method of self-control is what decision-theorists call 'precommitment'. It consists in deliberately eliminating or imposing restrictions on future options.[4] As Jones points out, 'a major mechanism for dealing with likely future lapses in self-control is to establish binding rules that prohibit the unwanted behaviour' (Jones, 2001: 46). The classical example of precommitment is the Homeric episode in which Ulysses instructs his crewmen to tie him to the mast, thereby allowing him to hear the Sirens' alluring songs without taking the risk of running his ship onto the rocks. Precommitment thus requires the agent's acknowledgement of his own propensity to yield to irrational impulses, along with the notion that one cannot maximize long-term well-being without resorting to self-imposed constraints. A more common example of self-control by precommitment is the individual who manages to avoid overeating sweets at home simply by not having any sweets at home. The prospect of having to go out to buy sweets is often enough to dissuade the impulse to give in to the craving. Likewise, the indebted consumer who continues to purchase on credit is often advised to cut up his credit cards with a pair of scissors. And, similarly, some countries allow pathological gamblers to sign a self-exclusion agreement from casinos, which effectively keeps them away from gambling whether or not they later succumb to temptation.

But self-control by precommitment does not always entail the elimination of future options. In certain cases, it may be more fruitful to impose a sanction on the tempting alternative. Elster gives a convincing example of this strategy:

> If I begin saving for Christmas but find myself taking money out of my savings account instead of keeping it there … I may put my savings into a high-interest account that carries a penalty for early withdrawal, thus combining premium and penalty.
>
> (Elster, 2007: 238)

In a study about the efficiency of precommitment against procrastination, Arieli and Wertenbroch (2002) showed that this sort of strategy effectively reduces people's tendency to procrastinate. Interestingly, they write, 'people are willing to self-impose deadlines to overcome procrastination, even when these deadlines are costly' (Arieli/Wertenbroch, 2002: 221). Furthermore, their study indicated that precommitment to deadlines is successful not only in reducing procrastination but also in helping students achieve better grades. In particular, students who chose to be penalized at the rate of 1 per cent of the grade for each day late had on average better marks than their peers. As Arieli later observed, these results are interesting in that they suggest that 'although almost everyone has problems of procrastination, those who recognise and admit their weakness are in a better position to utilise available tools for precommitment and by doing so, help themselves overcome it' (Arieli, 2009: 116).

Another way of counteracting irrational impulsiveness is what psychologists call 'emotional regulation'. We have noted that emotions are involuntary states, in the sense that it is not 'up to us' to feel (or not feel) something, but it is also true that we can control our emotions indirectly, to a certain extent, either by focusing on the relevant stimulus or by manipulating the external conditions (Levenson, 1994). Given that irrational preference reversals often originate in the influence of emotions on the process of decision-making, emotional regulation can be an effective tool for self-control. Thus, for example, the person who wants to lose weight may adopt the strategy of buying groceries shortly after a meal, when his or her cravings are weaker and therefore less likely to affect his or her judgment. Likewise, the employee who doubts he will have enough courage to ask his boss for a rise can deliberately try to recall all the injustices he endured at work in an attempt to enhance the motivation to confront him (Skinner, 1953: 236). And the mother who dreads the consequences of feeling overly angry with her child can mitigate that feeling by focusing on the good moments they spent together (Mele, 2001: 106).

In each of these examples of self-control, the key aspect is that the agent's strategy relies on the acknowledgement that his or her judgment may be affected by emotional biases. There are of course many more ways of preventing impulsive behaviour and of ensuring that one's 'future self' sticks to the good resolutions. The point to be made here is that the most fruitful methods to promote rational action are likely to be those that take into account the human

propensity to act irrationally and include devices meant to lock oneself into a desired course of action, even if it implies manipulating one's 'future self'.

Conclusion

This chapter was an attempt to show how a correct understanding of the phenomenon of impulsiveness can help us come up with effective ways to prevent it, or at least to mitigate its effects. This meant demystifying the myth of willpower and the illusion that people have the ability to enhance their own motivation by an act of will. This traditional approach construes self-control as a victory of the Will over the emotions, but we have seen that it struggles to define the Will without metaphors, and to explain what mysterious principle governs the so-called 'strength of will'. As Ledoux points out, the persistence of this approach may be due to the fact that 'Christian theology has long equated emotions with sins, temptations to resist by reason and willpower in order for the immortal soul to enter the kingdom of God' (LeDoux, 1996: 24).

Instead, Belief-Desire theories of action suggest that the Will is reducible to the plethora of emotions that constitute our motivation to act. On this account, the battle does not take place between the Will and the emotions, but rather between competing emotions: between the desire to smoke and the fear of cancer, for example, or between the desire to eat candy and the desire to look good, or between the desire to be unfaithful and the fear of the consequences. There is nothing obscure or unintelligible about the explanation of action in terms of an agent's motivation (his or her emotions). Unlike the Will, the Soul and other 'ghosts in the machine' (Ryle, 1949: 62), emotions leave traces on brain scans and are present in other species. That is to say that it is possible to account for their role in human behaviour without resorting to purely metaphysical notions. Spinoza expresses this idea very elegantly when he observes that the best way to counteract impulsiveness is not by fighting emotions, but by improving our knowledge of how emotions operate: 'There is no remedy within our power which can be conceived more excellent for the affects than that which consists in a true knowledge of them' (Spinoza, 2001: 232).

Once we understand that we are either weak-willed or self-controlled depending on our motivation to do what we judge better; and that our motivation, in turn, does not depend on our own direct, voluntary efforts, it becomes apparent that the best way to promote self-control is by imposing constraints on future options and by treating our 'future self' as a rebel employee that needs constant supervision. Granted, this image does not sound as glorious and illustrious as Kant's ideal of autonomy, but maybe that is because it is grounded on what we know about human nature, rather than on what we wish it would be.

Notes

1 Rationality is an elusive concept, as Ainslie (2005: 645) points out, insofar as 'there is no hard and fast principle that people should follow to maximize their prospect of reward'. However, there is one minimal requirement that all rational agents are expected to meet, namely: the requirement to act in accordance with their own principles of rationality. Davidson (1985: 346) conceptualizes irrationality as the failure to meet this requirement: 'We should limit ourselves to [consider as irrational] cases in which an agent acts, thinks or feels counter to his own conception of what is reasonable; cases in which there is some sort of inner inconsistency or incoherence'.
2 For a fuller analysis of the origin of the notion of will, see C. Kahn (1988) and J.-B. Gourinat (2002).
3 In fact, the proponents of this model interpret the phenomenon of intemperance as evidence that the traditional notion of Will is fundamentally flawed. Hobbes (1996: 44) writes: 'The definition of the Will, given commonly by the Schools, that it is a rational appetite, is no good. For if it were, then could there be no voluntary act against reason'. See also Spinoza, *Ethics*, III, Prop. 2, *scholium*.
4 For a fuller account, see Elster (2007), Chapter 13.

References

Ainslie, G. (2001). *Break-down of the Will*. Cambridge: Cambridge University Press.
Ainslie, G. (2005). 'Précis of Breakdown of Will'. *Behavioural and Brain Sciences*, 28, 635–673.
Arieli, D. (2009). *Predictably Irrational*. London: Harper Collins.
Arieli, D. and Wertenbroch, K. (2002). 'Procrastination, Deadlines, and Performance: Self-control by Precommitment'. *Psychological Science*, 13, 3, 219–224.
Aristotle (2000). *Nicomachean Ethics*, trans. R. Crisp. Cambridge: Cambridge University Press.
Aristotle (2010). *Rhetoric*, trans. W. Roberts. New York: Cosimo, Inc.
Bird, A. (1994). 'Rationality and the Structure of Self-deception'. *European Review of Philosophy: Philosophy of mind,* vol. 1, Stanford, CA: CSLI Publications, 19–38.
Cicero (2002). *Cicero on Emotions: Tusculan Disputations 3 and 4*, trans. M. Graver. London: University of Chicago Press.
Coplan, A. and Goldie P. (eds.) (2011) *Empathy: Philosophical and Psychological Perspectives.* Oxford: Oxford University Press.
Davidson, D. (1985). 'Incoherence and Irrationality'. *Dialectica*, 39, 345–354.
Davidson, D. (2001). 'How is Weakness of Will Possible?' In Davidson, *Essays on Actions and Events*. Oxford: Oxford University Press [1969], 21–42.
Descartes, R. (1989), *The Passions of the Soul*, trans. S. Voss. Indianapolis, IN: Hackett Publishing Company.
Elster, J. (1979). *Ulysses and the Sirens*. Cambridge: Cambridge University Press.
Elster, J. (1999). 'Davidson on Weakness of Will and Self-deception'. In L. E. Hahn (ed.), *The Philosophy of Donald Davidson*. Open Court: La Salle, 425–441.
Elster, J. (2007). *Explaining Social Behaviour*. Cambridge: Cambridge University Press.
Goldman, A. (1976). *A Theory of Human Action*. Princeton, NJ: Princeton University Press.
Gourinat, J. B. (2002). 'L'apparition de la notion de volonté dans le Stoicisme', in P. Saltel (ed.), *La volonté*, Paris: Ellipses, 49–59.

Hobbes, T. (1996). *Leviathan*, Cambridge: Cambridge University Press.
Holton, R. (2003). 'How is Strength of Will Possible?' In S. Stroud and C. Tappolet (eds.), *Weakness of Will and Practical Irrationality*. Oxford: Oxford University Press, 39–67.
Hume, D. (1985). *A Treatise of Human Nature*. London: Penguin Books.
Hume, D. (2007). *A Dissertation on the Passions*. Oxford: Oxford University Press.
Jones, B. (2001). *Politics and the Architecture of Choice*. Chicago, IL: University of Chicago Press.
Kahn, C. (1988). 'Discovering the Will: from Aristotle to Augustine'. In J. Dillon and A. Long (eds.), *The Question of 'Eclecticism'*. Berkeley, CA and London: University of California Press, 234–261.
Kant, E. (2006). *Anthropology from a Pragmatic Point of View*, Robert B. Louden (ed.), Cambridge: Cambridge University Press, 2006.
LeDoux, J. (1996). *The Emotional Brain*. New York: Simon & Schuster Paperbacks.
Levenson, R. (1994). 'Emotional Control: Variations and Consequences'. In P. Ekman and R. Davidson (eds.), *The Nature of Emotion*, Oxford: Oxford University Press, 273–280.
Lewis, D. (2000), *Papers in Ethics and Social Philosophy*. Cambridge: Cambridge University Press.
Locke, J. (1997). *An Essay Concerning Human Understanding*. London: Penguin Books.
Loewenstein, G., Read, D. and Baumeister, R., (eds.) (2003). *Time and Decision*. New York: Russell Sage Foundation.
Mele, A. (1987). *Irrationality*. Oxford and New York: Oxford University Press.
Mele, A. (2001). *Autonomous Agents*. Oxford and New York: Oxford University Press.
Mele, A. (2012). *Backsliding*. Oxford and New York: Oxford University Press.
Plato (2004). *Republic*, trans. C. Reeve, Indianapolis, IN: Hackett Publishing Company.
Ryle, G. (1949). *The Concept of Mind*, London: Penguin Books.
Searle, J. (2001). *Rationality in Action*. Cambridge, MA: MIT Press.
Sen, A. (1986). 'Behaviour and the Concept of Preference'. In J. Elster (ed.). *Rational choice*. New York: New York University Press, 60–81.
Skinner, B. (1953). *Science and Human Behavior*. New York: The Macmillan Company.
Spinoza, B. (2001). *Ethics*. Chatham: Wordsworth Editions.
Strotz, R. H. (1956). 'Myopia and Inconsistency in Dynamic Utility Maximization'. *Review of Economic Studies*, 23, 165–180.
Stroud, S. (2010). 'Is Procrastination Weakness of Will?' In C. Andreou and M. White (eds.) *The Thief of Time: Philosophical Essays on Procrastination*, Oxford, Oxford University Press, 51–67.

3 Emotions and akratic feelings

Insights into morality through emotions

Dina Mendonça

Introduction

The fact that emotions are crucial for morality has been increasingly stated in the philosophical, psychological and cognitive science literature. Horberg *et al*., for example, point out how recent scientific research shows that emotions have input in complex moral judgments (Greene and Haidt, 2002; Haidt, 2001), and how the embodiment effect of emotions contributes to moral judgment (Hoberg *et al*., 2011: 237). Other examples include Jesse Prinz's argument that emotions are necessary for morals (Prinz, 2006), Nussbaum's description of how moral emotions influence a person's prescriptive conception of fairness (Nussbaum, 2001), and Justin D'Arms and Daniel Jacobson's extensive writings on how values are grounded on sentiments (D'Arms and Jacobson 1994, 2000, 2005, 2006). It is by now clear that any analysis of morality and ethics[1] requires the inclusion of emotions.

In this chapter, I show that the role emotions play in morality is also connected with their ambivalence and their obscure nature. Emotions are ambivalent because sometimes they are the source of certainty and foundation for security in action, and at other times they are the source of error, deception and self-deception. As Peter Goldie has written, 'we are inclined to say that emotional experience can sometimes tell us things about the world that reason alone will miss' (Goldie 2004: 249). And we inevitably add that 'on the other hand, we are inclined to say that our emotions can and do profoundly distort our view of things: in anger or jealousy, for example' (ibid.). The reason why their ambivalence is increased by their obscure nature is twofold. First, the connections between emotions and the words is not simple and straightforward as the words we have to describe them fall short of grasping their full nature because '[e]motions are processes, changing over time, and emotion labels cannot capture the complexity of this psychological motion' (Ellsworth *et al.* 2006: 585). Second, their value and identity is not immediately given with their occurrence. Most likely this lack of immediacy in meaning occurs because of their Janus-faced nature. That is, '[t]hey tell us something about the world, and they tell us something about ourselves' (De Sousa 2007: 323). Acknowledging emotions' difficult nature reminds us that they are far from being as simple

and compact as the names we have for them (Dewey 1934: 48), and provides the ground for recognizing that some of our emotional experiences force the recognition of a meta-emotional level.

The chapter will argue that the obscure nature of emotions is a catalyst for our effort to understand emotional episodes better, recognizing their meta-emotional level, enhancing and refining empathic behaviour, and promoting the need for continual self-knowledge (Mendonça 2013). Acknowledging this will reinforce the role of emotions as unique repositories of information on morality. The combination of emotions' obscure nature with their reflective character allows us to explore the ways in which this symbiosis creates a new platform for empathy. The chapter begins by establishing the existence of akratic feelings (Mele, 1989), and how they give rise to a meta-emotional feeling of puzzlement. Afterwards, I show how the meta-emotional platform afforded by these feelings impacts on empathic processes. It will become clear how both akratic feelings and meta-emotions (Mendonça, 2013) provide the ground for a deeper sense of empathy and how they promote a continual move for self-knowledge, further solidifying the possibility for empathy. While the empathic connection is made stronger as people struggle to understand their own and other people's feelings, there is no point at which people can finally say they are fully transparent.

Akratic feelings

Akratic feelings are emotions and sentiments[2] that make no sense within a specific emotional landscape, and are felt as unexplainable both from the subject's point of view as well as from a second person's perspective. The term 'akratic' refers to a specific action, which, despite having been done in a free and voluntary fashion, is also simultaneously contrary to the subject's best judgment.

The problem of akratic action was first identified in Ancient Greek times. It can be found in Plato's dialogue *Protagoras*, which construes it as an illogical moral concept given that no one does something bad intentionally (358d), and in book VII of Aristotle's *Nicomachean Ethics* (1146b5–1147b20), where it is described as a weakness of the will found in moments of diminished cognitive and intellectual capacity. In contemporary philosophy, we continue to find akrasia widely discussed. It is, in conjunction with the problem of self-deception, a major challenge for any theory of rationality (Correia, 2010: 275). Philosophers that have engaged in this debate include Donald Davidson, who circumscribed the limits of akratic action by showing that the subjects make a judgment on a subset of possible considerations (Davidson, 1980), Amélie Rorty, who distinguished between different kinds of akratic action such as akrasia of direction, interpretation, irrationality or character (Rorty, 1980), Richard Holdon, who argued that it is possible to act contrary to our best judgment and not suffer from weakness of the will (Holton, 1999), and Alfred Mele, who has undertaken a deep analysis of the way in which akrasia is connected to the rest of all the other mental concepts that underlie action, and who offered a detailed conceptual analysis of akrasia in the history of philosophy

(Mele, 1986a, 1986b). Mele also drew an interesting parallel between akratic action and akratic emotions (Mele, 1989).

In a paper entitled 'Akratic feelings', Mele argues that akratic feelings are possible and explains and identifies their occurrences. According to Mele, akratic feelings are similar to akratic actions for they are emotional entities that go against the subject's best judgment. Akratic emotions should not be taken as emotions that we do not want to feel as, for example, the impatience with strangers when we are standing in line, or the irritation we may feel towards someone we love when they tells us truthful and important things that are hard to accept. These are not akratic feelings because, though we may not want to feel them, we recognize a certain reasonability of their occurrence, regardless of the variation of the self-control we have in expressing them. However, sometimes, we have emotions that surprise us not only because they are indifferent to our will and self-control but also because they seem to make absolutely no sense given the circumstances. These are the emotions that Mele identified as akratic feelings, and which 'must, like akratic actions, be at odds with a decisive better judgment of the subject' (Mele, 1989: 279). For example, someone may feel romantic jealousy for a person with whom there is no relationship nor desire for such relationship or, to use the example given by Mele, someone may feel simply a sensation of inadequacy in finding out a lover's betrayal, or even a sense of pleasure or relief in a partner's betrayal.

As Mele points out that just as akratic actions are distinct from compelled action, akratic feelings also assume this trait. Consequently, they are subject to control in a way that compelled feelings are not. Just like we distinguish 'the heroin addict from the thrill-seeker akratically using the drug for the first time, on the grounds that the former cannot help herself' (ibid., 278), we also distinguish feelings of anger when someone is screaming at us after a day of excessive caffeine intake and a week with two or three hours of sleep from the anger which occurs when the day has just begun after a well-rested night, or from the anger that is felt when someone is saying something apparently nice and polite. The difference between compelled emotions and those that are not compelled observed in the aforementioned examples regarding anger can be found in other emotions, even if at times it is harder to identify. Nevertheless, the distinction is subtle and only in repeated cases that become problematic are emotions recognized as compelled in a clear and unquestionable way. For example, it is clear that the fear felt by someone who suffers from arachnophobia consists of a compelled emotion as opposed to a fear expressed by someone who does not suffer from such a condition. However, the boundaries of what is considered within our control remain problematic since some people overcome their phobias while others do not. Therefore, the distinction between what can be seen as being part of a person's control in what concerns emotional experience, and what is labelled as compelled, is decisive for placing boundaries on what can be considered an akratic feeling. Despite these underlying problems, the fact that we can find clearly compelled emotions and a difference of degree in emotional control validates the existence of akratic feelings (ibid., 287).

Akratic feelings and meta-emotions

One of the important traits of akratic feelings is the way in which they allow a meta-emotional level of emotional experience. When we experience an akratic feeling, we also feel confused, puzzled and sometimes surprised by our own feelings, unearthing a meta-emotional level within our emotional world. Let me exemplify: imagine that someone feels perplexed by being sad after winning a game. Here, the sense of confusion does not appear in a simple and sequential fashion as it happens when someone loses a game or an object of great emotional value. In this last case, one may feel clueless about when and how the object was lost, while in the first example one is puzzled about the emotional state (sadness) promoted by the event (winning the game). That is, emotions can be layered instead of sequential (Pugmire, 2005: 174). When someone is perplexed (second-order emotion) about his or her own sadness (first-order emotion), emotions are layered. However, emotions are sequential when we are first clueless and then sad about the loss of an object because both emotions are first-order emotions. It seems to be expected to be able to feel happy following a victory in a game, and thus the example given also qualifies as an akratic feeling. It is the non-compulsive nature of akratic emotions that forces the meta-emotional level of puzzlement experienced.

If akratic emotions and feelings are dismissed as irrational and accidental, they lose their strength and, subsequently, can easily be ignored, bypassing the meta-emotion puzzle they promote. In addition, it is also hard to identify meta-emotions even though their existence is undeniable. Meta-emotions are hard to identify because their phenomenology is sometimes mixed with the first-order emotions phenomenology, and the vocabulary used to describe emotions is equal to both first- and second-order emotions.

Meta-emotions

In 'Paradox of Fiction', Susan Feagin points out the shortcomings of our vocabulary for emotional experience. She claims that the difficulty in distinguishing the meta-emotion from emotions of a first-order level is that the vocabulary available to describe first- and second-order emotional experience is the same. Consequently, it is hard to separate a first blush from a blush about a blush (Feagin, 1995: 208). Feagin concludes that the difficulty does not diminish the crucial importance of the distinction, stressing there are two types of embarrassment in the two blushes. The fact that the relationship between emotions and our vocabulary for them is not straightforward is not novel. However, Feagin's reflection adds another difficulty: the words we have for emotions may refer to different levels of emotional experience and, consequently, intensify the complexity of this connection.

Philosophers and psychologists have been increasingly exploring the phenomenon of meta-emotion (Jäger and Bartsch 2006; Mitmansgruber *et al.* 2009; Mendonça 2013; Jäger and Bänninger-Huber 2014; Norman and

Furnes, 2014), and established that their importance lies partly in the way they mould the first-order emotional experience. In 'Mediating with Heart in Mind: Addressing Emotion in Mediation Oractice', Jones and Botcker describe how '[m]eta-emotions color or influence the primary emotions being experienced' (Jones and Botcker, 2001: 240).

There are further reasons that explain the difficulty in separating first (an emotion: fear) from second-order level of emotion (an emotion about an emotion: sadness about fear). First, the second-order level may be confused with the first-order level when the meta-level reinforces the phenomenological manifestations of the first-order emotion. This happens when someone is sad about being sad, or embarrassed about being embarrassed, or happy about feeling happy. Second, the meta-emotion may demand a different action from the person who experiences both levels, and even an opposite action from the first-order emotion. This means that in addition to the fact that it may be hard for someone who is sad to acknowledge that there is a second-order level of sadness, it can also be hard to identify the second-order level of emotion when it is different from the first-order emotional level. Jones and Botcker give the example of someone who is angry and then embarrassed about being angry, and point out that the difficulty lies in acting in a strategic and adequate way in line with both emotional feelings (Jones and Botcker, 2001: 239). That is, when these two levels are different, they may demand contradictory actions and experiences, causing emotional confusion and making it hard to identify the different layered emotions. Jones and Botcker further explain that since meta-emotions are grounded in the values and beliefs we have, and are taught and determined by culture, our emotional landscape includes settled relationships between first- and second-order emotions. Therefore, people who have been taught not to get angry will feel shame in the experience of the emotion of anger, and may find ways to cover and disguise the first-order level emotion in such a way as to make the second-order emotion invisible. Likewise, they may feel shame at the possibility of anger making shame a prominent emotion and turning the threat of the first-order emotion invisible. Thus, argue Jones and Botcker, meta-emotions are especially problematic for mediation given that people are normally less aware of their meta-emotional processes. Furthermore becoming more aware of our meta-emotional processes may be a crucial way to change perspective and adopt a more collaborative way to deal with a situation of conflict (ibid.).

Clearly, the regulative nature of meta-emotions can block and limit first-order emotions, for example, when we feel shame about an inadequate emotion such as joy at someone else's misfortune. Meta-emotions can positively contribute to moral judgment by helping to regulate and refine first-order emotions. For instance, guilt about jealousy can help master and overcome it. Nevertheless, the recognition of the existence of meta-emotions raises many further questions about their nature. It is neither clear if meta-emotional processes happen at certain specific moments or if they are always present, nor if all emotion types are capable of being felt at the two emotional levels (Mendonça, 2013: 393).

With regard to moral considerations, the fact that both akratic emotions and meta-emotions seem to be one item makes matters more obscure. However, I want to argue that it is in the connection between akratic and meta-emotion where we can mostly appreciate their contribution to morality and ethics by considering the empathetic outlook they can promote. When we experience akratic feelings, we are puzzled by our own emotions and this sense of puzzlement about ourselves enables a renewed sense of equality with others when we do not fully understand their emotions. That is, when we recognize akratic feelings these give place to meta-emotions of puzzlement and enable a meta-emotional level where we can be surprised about our own emotions. When this happens, we are able to have empathetic experiences with subjects in a different situation that also created a sense of surprise. In addition, because this specific empathetic process reinforces the emotional experience that gives rise to it, it also encourages the maintenance of the meta-level of ongoing wonder about our own feelings and emotions. That is, since the path for empathy is created by a sense of surprise, the more we empathize with others, the more we focus our attention in the mysterious nature of our emotions. The more we become amazed, the more we are capable of empathizing with others which in turn also increases our ability to be astonished.

Empathy

Empathy can mean many different meanings (Coplan and Goldie 2011; Gallagher 2012). We may feel empathy when we see someone sobbing uncontrollably even if we ignore his or her reasons. We can also feel empathy for someone when we know that someone close to them is dying, even if we do not see that person expressing any sign of sadness or grief. Despite the multiplicity of meaning of the term, all types of empathetic processes can be understood as a general capacity to understand the other, and this is developed since childhood (Gallagher, 2012: 169).

The comparison of different modes of empathy can be useful. Goldie distinguishes between two types of empathy: 1) an 'empathetic perspective-shifting', which requires a modification of the perspective in which a person intentionally and consciously imagines what it is like to the other person, and what he or she thinks, feels and decides, 2) an 'in-his-shoes perspective-shifting', in which we place ourselves in the other-person situation and imagine what we ourselves would do, feel and think if we were in their place (Goldie, 2011). Goldie explains why these two types of empathy are normally undifferentiated and points out that in simple cases the outcome of both is the same, given that most individuals share common generalities about how they feel, think and act, for example when someone close to them dies. Nevertheless, Goldie adds that, if we move beyond the basic and more general cases and consider the details of the situations, then the distinction between these two different types becomes important (ibid., 308). Hence, he argues that ignoring the distinction may bring out errors in empathetic processing, and open the possibility for

misunderstanding others. Consequently, we should aim for 'in-his-shoes perspective-shifting' as the approach enabling a more accurate sense of empathy.

Consider the establishment of empathy with people who are much younger. For example, imagine a toddler having a tantrum because he or she does not want to go to bed. We may or may not feel empathy towards him or her but we will want to be able to communicate with the child as to end the tantrum and succeed in putting him or her to bed. By establishing this distinction, Goldie wants us to recognize that we have difficulty imagining what it is like for the child to have to go to bed and how that feels. Even if we had similar tantrums, it is unlikely we would remember them with enough detail. If we do not acknowledge the differences that age difference bring about, we may end up acting in inappropriate ways when trying to help the toddler. At least in some situations, it is crucial to acknowledge that others are not like ourselves for a more effective communication.

Additionally, Goldie's distinction suggests that by becoming more aware and self-aware, we are also more capable of adopting the 'in-his-shoes perspective-shifting' and this, in turn, provides more opportunities for self-knowledge and increases our potential to empathize with others. That is, empathy is a lifelong discovery of others and our own selves. There are many reasons for the fact that emotion fosters an ongoing progress of both self-discovery and empathy. First, it seems, we are not the best judges of our own emotional states. Some of our emotional experiences are lived without full awareness that they exist, and of their meaning (Haybron, 2007).[3] However, in 'Real Emotion', Pugmire describes that we have a tendency to think about our emotions through a biased perspective about ourselves. He writes:

> Choice will center on emotions that promise advantage of power (e.g. pity), moral advantage (e.g. forgiveness, and above all, righteous anger) or that reassuringly affirm desirable personal qualities (e.g. compassion, remorse). […] The trapping of one emotion can serve to mask another: righteous indignation rather than envy or spite; zealous commitment to a cause that offers a feeling of belonging or of transcendence of the commonplace and the compromised; pity instead of disdain; solicitous concern as opposed to prurient fascination. Notice that the masking emotions tend at once to resemble and to deny the masked emotions.
>
> (Pugmire, 1994: 114)

According to Pugmire, we adopt emotions that can camouflage others, and which feel more desirable for the evaluation of our own person. Pugmire's take reinforces Goldie's indication that at times we cannot know what others would do in certain situations, for we also know little about what we would do, feel and think. There is a sense in which we are as mysterious to ourselves as others are to us, and the emotional area is the place in which we mostly feel the opaqueness of our own persons. The suggestion above does not undermine the sincerity of the subject who observes. As Pugmire describes, the subject who

experiences a certain emotion of righteousness indignation disguising envy is truly sincere about his or her emotion, and believes in its authenticity even if it can be pointed out how it holds an illusionary perspective about the individual's relationship to the world (Pugmire, 1994: 108). Empathetic processes are not only ways to know others but are crucial for self-discovery, and both empathy and self-knowledge reinforce each other and cannot be taken in isolation.

When we find that someone else feels perplexity in face of a certain akratic feeling, we may not understand what the other person feels in the first-order level of emotional experience (that is, at the level of the emotion) but if we have had the sharpness to identify akratic feelings in ourselves, we would know how it feels on the second order of emotional experience (that is, on the level of the emotion about an emotion) because we too have been perplexed by our own feelings. Just as we do not understand ourselves totally, we also do not understand the other, and our own opacity and that of others shifts from being a problem to becoming a shareable event. Likewise, the opacity of others also becomes a shareable event in that there is a common measure of being perplexed about a certain type of feelings. This move offers the possibility of an inversion of emotions concerning the other that holds great impact for moral and ethical discussions.

Sharing the experience of perplexity about our own emotions

In 'Emotions as Amplifiers: An Appraisal Tendency Approach to the Influences of Distinct Emotions upon Moral Judgment', Horberg *et al*. show how specific emotions influence judgments of self-other similarity or dissimilarity, stating that, '[o]verall, compassion promoted feelings of self-other similarity, whereas pride promoted feelings of dissimilarity from others' (Horberg *et al*., 2011: 240). The qualitative empathetic move granted by akratic emotions and the meta-emotional process enables a platform in which one can feel compassion towards oneself and pride for others in a inversion of the most immediate and easily identified amplification of emotions.

This point in which we feel as opaque as others makes it easier to compare and contrast differences and similarities. For example, imagine that we know someone who had a violent episode of jealousy. We can check if we would or would not feel jealousy in such an instance. However, we would have difficulty in verifying if the way in which we would feel jealousy is similar to the person we observed. But if we imagine that this person feels perplexity about her or his own episode, it is more plausible to imagine a sense of perplexity about an emotional episode (which does not require it to be about the same first-order emotion of jealousy). Feeling perplexed by our own feelings forces us to accept a certain type of ignorance that enables a more humble type of self-knowledge and a less misguided and less prone tendency to self-delusional processes.

The akratic feeling requires the first-person authority. Only when someone recognizes that an emotion feels akratic can there be the recognition of opacity, that is, we do not need to establish the similarity or equality of the first-order

emotional event for this empathetic level, but there needs to be the first-person authoritative recognition of being surprised about one's feelings for an empathetic understanding. Thus, it is fundamental that the subject sees him or herself as incomprehensible as it is this recognition of self -opacity that enables the comparison and contrast with others. The phenomenological literature acknowledges this proposition.

Thompson, for example, writes about the way in which we experience ourselves as another through empathy, and are able to access a perspective about our own person in its totality by going beyond the singular first-person perspective (Thompson, 2001: 8). Thompson adds that Edith Stein developed a similar conception of empathy with her concept of 'reiterated' (Thompson, 2001: 19). This concept aims to grasp the way through which we understand the other's perspective and the way in which we obtain an empathetically connection with ourselves such that we move beyond what we realize about ourselves, and obtain a much deeper perspective than the insights gained from more elementary empathetic processes. Thomson concludes that our personal identity is inseparable from our ability to understand the other and from our ability of self-awareness (Thompson, 2001). Thus, personal identity includes seeing oneself as an entity-empathetically-grasped-by-another such that 'one's sense of self-identity, even at the most fundamental levels of embodied agency, is inseparable from recognition by another, and from the ability to grasp that recognition empathically' (Thompson, 2001: 20–21).

Conclusion

At this point, we do not know how frequently akratic feelings occur. Because they are uncomfortable, and we tend to search for reasonableness of feelings, they are often dismissed and ignored. Nevertheless, accepting them as a crucial and an inherent part of our emotional world may be one of the ways to better understand the individuality of each one of us, as well as the similarities and differences we share as a species. This has a great potential for discussions about ethics and morality. Not only do akratic feelings offer the possibility of renewal of empathetic processes that enlarge the knowledge we have of ourselves and of others, they also reinforce the importance of feelings for evaluating and judging situations. As Thompson writes, 'emotions, as value feelings, make possible the evaluative experience of one-self and the world, and therefore are the very precondition of moral perception, of being able to 'see' a situation morally before deliberating rationally about it' (Thompson, 2001: 24).

The power of emotions is unique in the way they provide insights into morality and ethics for it is grounded on the recognition of our deep ignorance about ourselves and others, and on the complex relationships between various emotional entities and levels. There is, hopefully, much more to discover about our ignorance. Future work might be able to fully embrace the Deweyan pragmatist challenge, and by taking up philosophy as the general theory of education, examine 'where acceptance or rejection makes a difference in practice'

(Dewey 1916: 338), and further explore the consequences for ethical education. Hopefully, it will also do justice to the statement of the philosopher Ronald de Sousa who, in an insightful paper on the education of emotion, concludes, 'if we cease to think of our emotions as inevitable in just that way, we are also more likely to view them as open to modification, and to enlist them as instruments of freedom rather than tools of self-oppression' (De Sousa, 1990: 445).

Acknowledgements

This work would not have been possible without the financial support of Fundação para a Ciência e a Tecnologia (SFRH/BPD/102507/2014), and IFILNOVA for necessary backup for research activity. I would also like to thank Sara Graça da Silva for organizing the present volume, as well as the ongoing support of Professor João Sàágua.

Notes

1 I follow Gibson's take that Morality searches for considerations that give moral judgments claim to moral truth and looks for the demands of what is morally right with very precise employment of reasoning attaining principles such as, for example, the categorical imperative. Ethics targets the cultural grounds of human action and experience exploring the values of a community (Gibson 2011). The two are complementary: morality without ethics risks being too detached from human nature 'issuing demands that cannot be squared with human nature or that are destructive of the very practices – friendship, for example – that give our lives meaning (Gibson 2011: 81).

2 I will mainly refer to emotions throughout the paper and assume that the terms feelings, emotions, sentiments are all affect-laden mental states that are object directed as opposed to moods, which lack object and can attach themselves to a variety of objects.

3 Though the subject of unconscious emotions is still currently under debate, there are good reasons to consider their existence as well as their absurdity (Hatzimoysis, 2007). For the present purpose, it is sufficient to point out that we mask some uncomfortable and undesired emotions with others.

References

Coplan, A. (2012). 'Understanding Empathy: Its Features and Effects', in Amy Coplan and Peter Goldie (eds.), *Empathy. Philosophical and Psychological Perspectives*, Oxford: Oxford University Press, 3–18.

Coplan, A. and Goldie, P. (eds.) (2011) *Empathy: Philosophical and Psychological Perspectives.* Oxford: Oxford University Press.

Correia, V. (2010). 'Os limites da racionalidade: auto-engano e acrasia'. *Disputatio*, III, 28, May, 275–291.

Davidson, D. (1980). 'How is Weakness of the Will Possible?' *Essays on Actions and Events*. Oxford: Oxford University Press, 21–42.

D'Arms, J. and Jacobson, D. (1994). 'Expressivism, Morality and the Emotions'. *Ethics* 104, 739–763.

D'Arms, J. and Jacobson, D. (2000). 'Sentiment and Value'. *Ethics*, 110(4): 722–748.

D'Arms, J. and Jacobson, D. (2005). 'Two Arguments for Sentimentalism'. *Philosophical Issues*, 15: 1–21.

D'Arms, J. and Jacobson, D. (2006). 'Anthropocentric Constraints on Human Value'. *Oxford Studies in Metaethics* (ed. R. Shafer-Landau). Oxford: Oxford University Press, 99–126.

Decety, J. and Jackson, P.L. (2006). 'A Social-Neuroscience Perspective on Empathy'. *Current Directions in Psychological Science*, 15(2): 54–58.

De Sousa, R. (1987). *The Rationality of Emotions*. Cambridge, MA: MIT Press.

De Sousa, R. (1990). 'Emotions, Education and Time'. *Metaphilosophy,* 1990, 21, pp. 434–446.

De Sousa, R. (2007). 'Truth, Authenticity, and Rationality'. *Dialectica,* 61, Issue 3 (September): 323–345.

Dewey, John (1916). *'Democracy and Education' The Middle Works*, 1899–1924. Southern Illinois University Press, 1985, vol. 9.

Dewey, John (1934). *'Art as Experience' The Later Works, 1925–1953*. Chicago, IL: Southern Illinois University Press, 1986, vol. 10.

Ellsworth, Phoebe C. and Tong, Eddie M.W. (2006). 'What Does It Mean to be Angry at Yourself? Categories, Appraisals, and the Problem of Language'. *Emotion,* 6(4): 572–586.

Feagin, S. (1995). 'The Pleasures of Tragedy', in A. Neill and A. Ridley (eds.) *Arguing about Art,* New York: McGraw-Hill, Inc., 204–217.

Gallagher, S. (2012). 'Neurons, Neonates and Narrative: From Embodied Resonance to Empathic Understanding', in A. Foolen, U. Lüdtke, J. Zlatev and T. Racine (eds.), *Moving Ourselves, Moving Others*, Amsterdam: John Benjamins, 167–196.

Gibson, J. (2011). 'Thick Narrative', in N. Carroll and J. Gibson (eds.), *Narrative, Emotion, and Insight.* University Park, PA: The Pennsylvania State University Press, 69–91.

Goldie, P. (2004). 'Emotion, Reason and Virtue' in D. Evans and P. Cruse (eds.), *Emotion, Evolution, and Rationality*, Oxford: Oxford University Press, 249–267.

Goldie, P. (2011). 'Anti-empathy' in A. Coplan and P. Goldie (eds.), *Empathy. Philosophical and Psychological Perspectives*, Oxford: Oxford University Press, 302–317.

Greene, J. and Haidt, J. (2002). 'How (and Where) Does Moral Judgment Work?', *Trends in Cognitive Science*, 6: 517–523.

Haidt, J. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment', *Psychological Review*, 108: 814–834.

Hatzimoysis, A. (2007). 'The Case Against Unconscious Emotions', *Analysis*, 67(296): 292–299.

Haybron, Daniel M. (2007). 'Do We Know How Happy We Are? On Some Limits of Affective Introspection and Recall', *Noûs*, 41(3): 394–428.

Holton, R. (1999). 'Intention and Weakness of Will', *The Journal of Philosophy*, 96(5): 241–262.

Horberg, E.J., Oveis, C. and Keltner, D. (2011). 'Emotions as Amplifiers: An Appraisal Tendency Approach to the Influences of Distinct Emotions upon Moral Judgment', *Emotion Review,* 3(3) (July): 237–244.

Jäger, C. and Bartsch, A. (2006). 'Meta-emotions', *Grazer Philosophische Studien*, 73: 179–204.

Jäger C. and Bänninger-Huber, E. (2014). 'Looking into Meta-emotions', forthcoming in *Synthese*, online 26 November 2014, DOI 10.1007/s11229-014-05.

Jones, T. and Botcker, A. (2001). 'Mediating with Heart in Mind: Addressing Emotion in Mediation Practice', *Negotiating Journal*, 3: 217–244.

Koven, N. S. (2011). 'Specificity of Metaemotion Effects on Moral Decision Making'. *Emotion*, 11(5): 1255–1261.

Mele, A. (1986a). 'Incontinent Believing', *Philosophical Quarterly*, April, 36(143): 212–222.

Mele, A. (1986b). 'Is Akratic Action Unfree?', *Philosophy and Phenomenological Research,* June 1986, 46(4): 673–679.

Mele, A. (1992). 'Akrasia, Self-Control, and Second-Order Desires', *Noûs*, September, 26(3): 281–302.

Mele, A. (1997). 'Real Self-Deception', *Behavioral and Brain Sciences*, March, 20(1): 91–102.

Mele, A. (1989). 'Akratic Feelings'. *Philosophy and Phenomenological Research*, Vol. 50, No. 2 (Dec.), 277–288.

Mendonça, D. (2013). 'Emotions about Emotions', *Emotion Review*, 5(4) (October): 390–396.

Mitmansgruber, H., Beck, T.N., Höfer, S. and Schüßler, G. (2009). 'When You Don't Like What You Feel: Experiential Avoidance, Mindfulness, and Metaemotion in Emotion Regulation', *Personality and Individual Differences*, 46: 448–453.

Norman, E. and Furnes, B. (2014). 'The Concept of 'Metaemotion': What is there to Learn From Research on Metacognition?', *Emotion Review*. Published online 13 October 2014, doi: 10.1177/1754073914552913.

Nussbaum, M. (2001). *Upheavals of Thought: The Intelligence of Emotions*. Cambridge and New York: Cambridge University Press.

Prinz, J. (2006). 'The Emotional Basis of Moral Judgments', *Philosophical Explorations*, 9(1) (March): 29–43.

Pugmire, D. (1994). 'Real Emotion', *Philosophy and Phenomenological Research*, March, LIV(1): 105–122.

Pugmire, D. (2005). *Sound Sentiments. Integrity in the Emotions*, Oxford: Oxford University Press.

Rorty, A. (1980). 'Where Does the Akratic Break Take Place?', *Australasian Journal of Philosophy*, 58: 333–346.

Thompson, E. (2001). 'Empathy and Consciousness', *Journal of Consciousness Studies*, 8(5–7): 1–32.

4 Morality and empathy *vs* empathy and morality

A quest for the source of goodness in phylogenetic and ontogenetic contexts

Augusta Gaspar

> Evolution has not left the important events of birthing and the ensuing nurturance and bonding either to chance or to the vagaries of individual learning.
>
> (Jaak Panksepp, *Affective Neuroscience*: 248)

Morality across cultures and as a universal human trait

Morality is a cross-cultural feature of humanity and a strong pillar of prosocial behaviour. Another major force keeping society together is empathy. Across borders and cultures, human communities have organized and supported social life on structures that bear astounding similarities among themselves. First and foremost, they all have myths of creation and share a belief in supernatural beings. Therein follows the organization of religious beliefs into a sensible religion, another universal trait of societies, which provides hope, consolation for grief, moral values and guidelines for proper conduct. Although morality is by no means identical across cultures and religions, moral values have common denominators – they all place a high price in the respect for gods, leaders and ancestors, in the care for one's family, or in not taking the life of a community member. This cross-cultural and all-time tendency to create a moral building to support civilization has been coined a Human Universal (Brown, 1991).

The origin of morality has been addressed by many theories. Albeit their close association, moral values do not necessarily stem from religion nor does religion stem from them. And although all humans share – far back in time – a common cultural evolution, cultures have diverged towards conflicting values, such as the place of women or homosexuals in society, not to mention towards different gods and myths of creation. What perhaps has not changed so much, binding all humans, are our most basic emotional responses which are at the very basis of widely accepted morality theories. Assuming that morality stems from emotional reactions that are common to the human brain across the entire geographical distribution of humans means that one has to accept that human morals cannot be arbitrary caprices of cultures. That moral behaviour originates in emotion, and thus in biology, is a view shared by many authors, from legendary ethologists such as Frans de Waal and Mark Bekoff,

to renowned psychologists such as Daniel Batson and Jonathan Haidt. Solid and compiled cultural studies within anthropology have documented that moral rules revolve around a limited number of themes, such as sexual behaviour, division of resources, or loyalty and respect towards authority figures (Brown, 1991; Shweder *et al.*, 1997). This could certainly suggest that they might have evolved throughout the cultural evolution of populations, as a cognitive strategy to regulate behaviour within communities and thus maintain social order and the *status quo*. Populations share common ancestry along the many features of cultural evolution, such as language, just as they share genetic ancestry. Thereby, their values must have homologous structures.

Richard Shweder and his collaborators (1997) posited three common denominators of morality across cultures: 1) *autonomy* (which emphasizes free will, and principles that allow people to tell right from wrong), 2) *community* (stressing interdependence, commitment and duty to one's community), and 3) *divinity* (establishing that the universe is ruled by the divine). For example, the latter proposes that cultures have established that humans have souls that belong to a divine being, whereby morality is the guardian of the immaterial soul from the constant threat of contamination from bodily things, which derives from the fact that it is contained in a body that has impure *drives* and commits disgusting, wrongful acts.

Jonathan Haidt has attempted to reconcile Shweder's three-domain model with the view that emotion drives moral choices, envisaging a way out of the biological/cultural dichotomy. He unfolded two of Shweder's domains into four, and equated divinity with purity/sanctity, later expanding into a model with five major foundations of morality, namely: in-group loyalty, authority/respect, fairness/reciprocity, harm/care and purity/sanctity (Haidt and Joseph, 2004). He also proposed that within these realms, people quickly decide what is right and wrong based on a 'gut-feeling', and afterwards pick a moral norm to support the moral option. This feeling is automatic but people take some time organizing their thoughts and articulating why that given action is immoral (Haidt *et al.*, 2000). This conclusion is based on the presentation of specific problems to people during interviews whereby people make quick judgements of right and wrong but are unable to provide reasons to support them (or otherwise provide delayed reasons). Examples of such problems include harmless violations of taboos – for example, whether it would be acceptable to eat a family pet that has been killed in a car accident or if it would be acceptable for a brother and sister to have voluntary protected sex. People were quick at providing judgements that the action was wrong, and only afterwards began searching for plausible reasons, often, as Haidt eloquently describes, introducing elements of harm, such as stating that eating dog meat would make a person sick (Haidt *et al.*, 1993). Shweder and Haidt (1993) formulated a theory of 'cognitive intuitionism' to explain these findings, where they posited that the human mind is set to respond to certain moral goods, not drawing them from principles but by a process of organizing and detecting patterns. Haidt *et al.* (2000) coined this intuitive, emotional judgement 'moral dumbfounding'.

Haidt's social intuitionism was strongly influenced by Alan Fiske's relational models theory (Fiske, 1991, 1992). Fiske proposed that morality springs from four distinct types of social cognition that seem to be common to all human cultures: 1) *communal sharing*, which involves kindness, kinship, and empathic concern for close others; 2) *authority ranking*, which describes the processes by which power and rank regulate access to resources but also requires that superiors protect their subordinates; 3) *equality matching*, which involves tit-for-tat reciprocity and the sense of fairness associated with egalitarianism, and 4) *market pricing*, in which ratio values of goods and services must be computed and aggregated across transactions. Fiske went into great detail on how processes of group identification are established within group bonding, authority and dominance, suggesting some hardwired mechanisms in the brain selected throughout human evolution that direct us to generate these societal processes and prompt us for such similar moral yields across societies. He also identified three of these types of morality generating social cognition in non-human primates. I will return to this particular discussion in the third section. Steven Pinker (2011) pointed out that in human history moral convictions have caused more harm than good – by driving people to religious wars, genocides and all sorts of horrible punishments. In the light of western contemporary views, all that suffering was inflicted because of what now seems to be very insubstantial moral justifications, or at least moral misdemeanours even within the cultures that deemed them important (e.g. homosexuality, chastity issues, disobedience and heretic talk). This volatility of moral values inspired Pinker to propose a refinement of the concept of morality. By distinguishing morality from a 'moral sense' we can sort out human universal predispositions to create overarching principles applicable to behaviour (moral sense), from strictly cultural codes (morality). The former refrain all societies from chaotic, amoral conduct; the latter meet more specific needs within each society.

Prevailing views today are based on accumulated empirical evidence favouring a dynamic interplay between biological, cultural and interpersonal factors in the development of human morality. One path taken explores where emotions may lead us regarding moral sense. As Dina Mendonça proposed in the previous chapter, meta-emotions may play a crucial role in morality via a sense that one is mirroring what others' feel, and someone else is mirroring what one has felt before. Neuroscience provided a substrate for this mirroring ability by revealing the existence of a mirror neuron system that enables motor and emotional simulation in the brain (e.g. Gallese, 2001, 2004), a strong vicarious experience. Mendonça also argues that different emotions provide content and meaning to morality eliciting situations and moral options. Furthermore, as the second section will show, our first-hand experiences (either emotional or not) play a major role in the development of cognitive empathy – i.e. the ability to 'see' through somebody else's eyes, understand their thoughts, feelings and actions. The emotional experience fuels cognitive empathy, which in turn, drives individuals to act in prosocial ways that are considered moral. It is important to distinguish that empathy comprises two major dimensions: emotional empathy

and cognitive empathy. The former includes primal emotional phenomena such as emotional contagion and empathic distress, which involve little or no conscious cognitive activity.

By definition, prosocial behaviour entails positive, friendly voluntary actions that benefit another individual or group of individuals (Eisenberg and Mussen, 1989). It includes offering assistance, helping, comforting, and ultimately, altruism, a special form of prosocial conduct that can result in extremely high costs for the individual who acts, and in extreme gains for the receiver of the act.

In the following pages, I set to address two key questions: 1) the interchangeable use of empathy and morality as synonyms of goodness, which is confusing and does not allow us to advance explanations of altruism, and 2) the underpinnings of prosocial acts that are likely to be convergent end products of different mental processes and distinct remote evolutionary causes. Both ontogenetic and phylogenetic perspectives provide important input to these questions, helping us to disentangle what is driven primarily by emotion from what is not.

Moral development in humans has been scientifically studied since early in the twentieth century. First approaches were highly influenced by behaviourism and later by social learning theory, and saw morality as the acquisition of moral norms by virtue of punishment and reward or through social observation and imitation, respectively. Morals were also thought to be context-dependent. On the contrary, biologically framed perspectives view morality as a default towards which children move throughout their ontogeny, whereby some basic predispositions are revealed as the child matures and is exposed to varied situations and developing opportunities.

Views on the ontogenetic development of morality

Making a case for culture

Different as they may be, many psychological approaches to the development of morality (*sensus* moral sense), voice a common denominator: the great emphasis put on environmental influences. Early twentieth-century views within psychology emphasized the 'acquisition' of morality by children. In this view, departing from a state of amorality, the child acquired notions of *right* and *wrong, good* and *bad* from supervising adults (parents, teachers) by means of punishment, value teaching, or moral metaphors such as those found in fables and other traditional tales for children. In the absence of these active 'injections' of morality, children were thought to fall into the kind of savagery portrayed in *Lord of the Flies* (Golding, 1954) and postulated by Durkheim (1925), who argued that children only distinguish between right and wrong if they are taught by an external authority – a parent, or a teacher – and provided they are taught discipline, sense of duty and sense of justice.

Hartshorne and his colleagues (Hartshorne and May, 1928; Hartshorne *et al.*, 1930; cit. by Liebert, 1979) studied large samples of children and adolescents (about 11,000) in the 1920s and 1930s in various experimental situations where

there was a temptation to commit a misdemeanour. They found that where children thought that adults were not observing them or could not have known what they did, children leaned towards cheating and lying, as they assumed that their acts would not be revealed. This occurred especially in the younger ones, and suggested that children's acts were largely dependent on context and not on universal rules of conduct. They concluded that knowing the moral rules *per se* (for example, the Ten Commandments or the Scout law) did not ensure moral conduct.

Playing the devil's advocate, one could argue that perhaps moral values are never 'poured in' enough, given that most adults never seem to reach the highest stages of morality predicted by child moral reasoning theories (e.g. Kohlberg, 1969, 1981; Gibbs *et al.*, 2007) – which is puzzling, *at least from an evolutionary stand point*. I will discuss this in the fourth section. Overall, much evidence has been amassed in support of the role of culture and situational variables in the emergence of antisocial behaviour.

Albert Bandura (Bandura *et al.*, 1961) revealed the powerful effect of imitation in the genesis of aggressive behaviour, whereby children copied the violent *vs* nonviolent behaviour of research assistants. The mimicked behaviour itself played a role in the construction of *schemas* (cognitive structures that enable further processing of information), suggesting that social modelling influenced both thoughts and behaviours related to what is right and wrong.

In the 1970s, researchers such as Grusec, Moore, Eisenberg, Hoffman and Mussen, began carrying out experiments with children that also highlighted the role of modelling in social conduct, particularly in prosocial behaviour (for a review see Eisenberg and Mussen, 1989). These researchers have shown that emotional experiences play a major role in prosocial behaviour and that the effect of models decreases with age, suggesting that the child internalizes norms on appropriate models quite early on and mostly before school age.

But important as modelling is, it is not the only factor operating. Several experiments have highlighted that young children are much more likely to imitate prosocial conduct when their parents or teacher (or otherwise experimental model) are warm, nurturing, attentive adults, and even more if they are the same sex as the child in question, than when they are neutral and more distant, or of the opposite sex (Eisenberg and Mussen, 1989). The effects seem to be long lasting. Eisenberg and Mussen also reviewed studies showing that older children and adolescents are sensitive to modelling by adults and peers who are perceived as powerful (e.g. prize winners), and that the effect of real conduct is more effective than manifesting concern and intention to act (altruistically for example).

In line with this, and after a series of studies with children, Staub (2003) concluded that in order to engage in prosocial behaviour, children need adult guidance and information. For example, he found a connection between orientation towards specific helpful actions and proneness to become helpful later in life; being debriefed about the specific consequences of one's acts yielded a similar outcome.

Despite their importance, social models could not be held accountable for all the antisocial conduct performed by children and adults. Not only does the child actively choose whom to imitate and what to act upon, but he or she also carries predispositions and develops schemas that will have enormous weight throughout his or her moral development.

Jean Piaget's (1932) perspective on the moral development of the child was closely linked to his view on cognitive development, entailing a gradual shift from complete moral *heteronomy* towards moral *autonomy*. The first and departing extreme – *heteronomy* – is characterized by a morality that resides outside the child's conscience, based on parents' morals and on the desire to be accepted by adults who are unilaterally and ultimately respected. It is also driven by the fear of punishment – the child has a sense that rules are sacred and immutable. This stage extends until the child is about 8 or 9 years old. In Piaget's theory, moral *autonomy* begins at about 9 to 11 years old and becomes the adult mode of moral reasoning, characterized by thinking moral values independently of whether others agree. Concurrent with formal operations and the ability to abstract concepts, from this point on, adolescents are thinking in terms of ethical principles and moral values, and confronting the application of these to real problems. Although cultural values give content to this moral sense, the child and the adolescent are active seekers and choosers in this process. This general view is still quite pervasive in western culture and in folk psychology.

Lawrence Kohlberg (1969, 1981) expanded Piaget's theory by formulating a six-stage model of moral development. Both agreed that children became increasingly autonomous and less self-centred in their moral judgements, moving towards an increasingly sophisticated social perspective, and that they did so very actively. Kohlberg's theory has been criticized as paradoxical because the growing ability to understand conventions, values and moral principles does not render a full development of morality, with only a meagre 2 per cent of the adult population estimated to reach the highest moral stages, known as post-conventional (Fowler, 1981; Gilligan, 1982). Kohlberg posited that throughout development, children navigate from an egocentric perspective towards a social perspective-taking, and from shallow values that reflect obedience and conformity to rules, towards an individual adoption and deep understanding of values and their worth in relation to norms and arbitrary conventions. Placement in a particular stage depended not on the specific solution to a moral dilemma but on the arguments and moral reasons provided by the individual in support of the chosen way out of the dilemma. Whilst this universally claimed path reflected maturity and cognitive development, it also showed distinct qualities that made it a separate construct from intelligence and reasoning. In his view, the process of moral development entailed specificities provided by culture, which could either spur moral reflection and social perspective-taking, or moderate them.

Other researchers have found support for the role of interactions in enhancing morality. For example, discussing moral topics with adolescents seems to be a particularly effective way of enhancing their moral reasoning, as adolescents are

able to find flaws in their own and others' moral arguments (e.g. Berkowitz and Gibbs, 1985; Walker *et al.*, 2000; cited by Vail 2011). Adolescents' moral reasoning and moral conduct is also influenced by their involvement in religion and community, particularly by engaging in solidarity activities within networks of shared values with adults and older peers (King and Furrow, 2004; see Vail, 2011 for a review). William Damon (1988) emphasizes the opportunities social play and interaction with peers offer to bolster morality, starting at the age of 3, by providing the contexts for sharing, perspective-taking, empathic concern and learning from other's behaviour. Although Kohlberg did not dwell on empathy issues, the increment in moral judgement that results from experience, either direct or indirect, finds a parallel in empathy development. Several studies today support the notion that empathy development can be assisted by exposure and participation and even by indirect experience.

Elisabeth Paul (2000) assessed empathy using a human-oriented and an animal-oriented questionnaire, and found that whilst both forms of empathy were moderately correlated, animal-oriented empathy was mostly related to having a pet at home or having had one during childhood, while human-oriented empathy was related to currently having a child or children at home. Gaspar and colleagues found an identical correlation (Emauz *et al.*, 2016) and confirmed that animal-oriented empathy is predicted by having a household pet during childhood (Emauz *et al.* in revision; Gaspar *et al.*, in press 2016).

The complexity of children's social interactions and the opportunities to develop their Theory of Mind (ToM) abilities along the intense maturation of relevant brain structures (e.g. prefrontal cortex), especially through 5 to 6 to 10 years old, plays a major role in empathy development (Decety and Michalska, 2010; Decety and Svetlova, 2012). So do aspects of family interaction early in childhood, including the quality of parental care and the family environment (e.g. Zahn-Waxler and Radke-Yarrow, 1990), the quality of the attachment the child establishes with her parents (e.g. Schore 2001; Decety and Svetlova 2012), the parents' actions (modelling of prosocial behaviour) to which the child is especially sensitive before 6 years old (Bandura *et al.*, 1961; Hoffman, 1975), and positive affect parenting styles, which have been connected to the prevention of antisocial conduct (Webster-Stratton, 1998).

Furthermore, influences over empathy predispositions also include the simulation of experiences in the mind of an observer while watching someone else's experience, which is enabled by the mirror neuron system (Iacoboni *et al.*, 2005), as noted in the first section of this chapter. Exposure to movies and other forms of vivid display of another's emotion seems promising as an enhancer of empathy. There are recent reports of increasing empathy towards specific targets as a result of experiments using movies (e.g. Blasco and Moreto, 2012; De Vied *et al*., 2009) or literary texts as stimuli (Kidd and Castano, 2013), when comparing the effects of the literary texts with more popular literature. The authors of the literary study, suggested that exposure to literary fiction might have engaged participants in processes of understanding the behaviour of characters based on their personalities and contexts, whereas popular fiction (and even day-to-

day life) exposes individuals only to outcomes that are interpreted based on stereotypes and common predictions. Hence, whereas the former promotes empathy, the latter does not.

Available comparisons across the life span have generally indicated a decrease in empathy from early adulthood onwards (Grühn *et al.,* 2008), and to my knowledge, no study has yet attempted to compare the effects of similar proximal factors on different age groups' empathic responses or on self-reported empathy. Over the last decade, we have witnessed a burgeoning trend of industrious school-based programmes designed for 'emotion-education' (which include empathy-teaching in most cases), but only a few programmes have been objectively evaluated in how they have contributed towards prosocial behaviour, emotion recognition ability and empathy, and fewer reported the theoretical and empirical grounds upon which their interventions were built. Therefore, we are still at the brink of exploring some of the influences and strategies mentioned above.

Making a case for the common biological roots of empathy and morality

A number of cross-cultural anthropological and psychological studies have been conducted since the 1980s to examine differences and similarities in the moral sense of children from completely distinct cultures. Studies departing from Kohlberg's six-stage evaluation (e.g. Gibbs *et al.*, 2007) show that western cultures vary somewhat in the prevalent stage demonstrated by adults, but most have in common that in general, adults do not reach the highest moral stages (post-conventional). Thus, high morality does not seem to be a human universal. To darken the scenario, extensive studies conducted by Daniel Batson (e.g. 1981, 1991, 1997, 2002) indicate that human adults only marginally (less than 80 to 90 per cent) act according to the moral principles they uphold. Instead, actions seem to be driven by self-interest, for example by assigning a positive consequence task to themselves and a negative or neutral to other participants in an experiment, or by attempting to deceive the experimenter by stating that they had flipped a coin to sort tasks, when in reality they had not. This pervasive phenomenon of attempting to display moral integrity, but maximally avoiding the actual cost of behaving morally was coined *Moral Hypocrisy* (Batson *et al.*, 1997, 2002). This led Batson and colleagues to seek alternative explanations for what lies beneath the, albeit marginal, real moral conduct. For example, Batson *et al.* (2002) found a moderate and positive correlation between acting morally and the scores in the widely used empathic concern scale (Davis, 1983), which seems to capture aspects of empathy related to emotional connectedness to the suffering and emotions of others. Overall, the origins of prosocial behaviour, and particularly altruism, seem to converge to the source of moral integrity/empathy. And, I would add, emotional empathy, because it is the dimension of empathy that automatically stems from dispositional traits (Gaspar, 2014a).

Other studies have departed from different paradigms and highlighted a much brighter view of the biological roots of morality, proposing that children

spot injustice rather similarly across cultures, and act as if they are driven by an inner sense of right and wrong. In some of the earliest studies, Judith Smetana (1981) asked 3 to 4 year-old children to judge moral norm violations such as hitting another child or stealing a toy, and the breaking of conventions such as putting toys away in the wrong place or going to school without pants. Children were asked to say how bad each violation would be, and whether the violation would still be bad or acceptable at a school without rules about hitting or putting things away. Children made a clear distinction between moral norms and conventions, suggesting they already grasped the notion that conventions are arbitrary whereas moral norms are not. This finding is in line with Elliot Turiel's (2000, 2002) long line of research (starting in the 1970s) on the formulation of an innate based convention/moral distinction. Challenging Piaget's heteronomy, these results show that preschool children already evince signs of autonomy in their moral sense.

Drawing from his own experiments on children's willingness to share chocolates, Damon (1988, 1999) also stresses that the underpinnings of moral sense are present from very early on in childhood. In one experiment repeated with children of various ages (4 to 10 years old), children were assigned a sharing task as a prize for winning another task (the pretext of the experiment was a team task and the prize would be shared among members of the winning team). One (control) group had colourful cards to share whereas another had chocolates. The study showed that children divided cards with fairness but when chocolates were at stake they were much less frequently inclined towards a fair share within the team. Yet, Damon reported that moral beliefs still held somewhat, as children never abandoned concepts of fairness, justifying inequalities, for example as a function of participation in the task (those who received more chocolates had worked harder on the task). However, they were more consistent with their beliefs when they were the beneficiaries of the belief (i.e. when they for example considered themselves the one who worked the hardest). Older children behaved more congruently with their previously stated beliefs of fair distribution.

Studies comparing rural and urban Chinese children and adolescents with same age Canadians in relation to their sense of fair punishment found extraordinary cross-cultural similarities, despite relevant cultural differences, for example, on the emphasis put on obedience, deference to authority figures or on the seriousness of actions that might embarrass the family, and on the prevalence of these forms of punishment in each culture (Bower, 2009). Children and adolescents were interviewed and asked to consider three types of punishment: 1) being discouraged from the wrongful behaviour with a conversation aimed at helping with reasoning about the wrongness of the behaviour, 2) appealing to family shame and making unfavourable comparisons with other children or 3) threatening to deny love. Afterwards, they were asked to rate them according to their fairness. All children and adolescents viewed denying love as unfair and as a cause of much suffering, and favoured the reasoning conversation not only as the fairest but also as the one they thought worked the best to prevent

future misbehaviour. This result was particularly revealing of a cross-cultural predisposition because both city and village Chinese children said that their parents' most frequent form of punishment was making negative comparisons with someone whereas in Canada parents used the reasoning talk preferentially.

So, where does the shared sense of justice come from? It may, on the one hand, be generated by hardwired systems in the brain, shaped through biological evolution. On the other hand, it may be related to shared basic emotional experiences which must be situated in the parent-infant bond and in the rewarding/frustrating, comforting/causing discomfort, regulating/dysregulating axes. I will present some of the evidence supporting the notion that we may be closing in on the shared cross-cultural source of the autonomous sense of morality. The outcomes of morality are prosocial acts and the inhibition of antisocial acts. These are also the outcomes of the very powerful emotional phenomenon that is empathy.

Much like the roots of moral sense, several components of emotional empathy (e.g. emotional contagion and empathic distress) pop out gradually from quite early in life: babies cry when other babies cry; children aged approximately 1 year begin showing discomfort at the explicit pain of others, and at 18 months they try to comfort others in distress (for a review see Eisenberg and Mussen, 1989). Additionally, by 4 to 6 years old one may already be seeing the defining traits of an empathic or otherwise callous psychopathic adult (Frick *et al.*, 2003). Evidence has been mounting that there are dispositional genetic traits that affect empathy/callousness (e.g. Jabbia *et al.* 2012; Larsson *et al.*, 2006).

So, do genetic programs speak louder than culture? And if so, should we settle for the verdict that morality is imported from the outside and empathy from the inside? How do they relate? The early components of empathy do not warrant that a child will grow to become an adult with moral conduct, but they may provide underpinnings without which there will never be a truly moral conduct, only a form of prosocial behaviour that is driven by self-interest.

Several studies have shown that the affective components of empathy, namely, emotional empathy and sympathy (Blair, 2005; Eisenberg and Mussen, 1989), are higher predictors of prosocial behaviour than affective knowledge (Knafo *et al.*, 2009), attitudes or beliefs (Correia and Dalbert, 2008). Furthermore, the absence of emotional empathy is a hallmark of psychopathy (Patrick *et al.*, 2009), with sociopaths often presenting high levels of cognitive empathy (Smith, 2006).

Cognitive empathy alone fails to predict prosocial behaviour when not combined with emotional empathy (Eisenberg and Strayer, 1987). There is an interesting parallel between these distinct consequences of cognitive empathy alone with the distinction Kohlberg establishes between a highly moral person and a person with a highly developed moral judgement (Kohlberg and Candee, 1984). This implies that one could provide answers that evince a deep understanding of the other's perspective, a commitment to universal ethical values, and a reasoning bound to justice – that are rated at the highest moral stage. However, in real life, the same person would not practise them and act only motivated by self-interest. It also fits in nicely with Batson's work on moral

hypocrisy. Kohlberg's social perspective-taking is a construct much similar to Premack and Woodruff's (1978) Theory of Mind, overlapping considerably with the construct of cognitive empathy, so it is not surprising that in regard to their role in moral conduct they perform similarly.

The largest study on emotional empathy to be carried out with very young children was longitudinal and involved 409 monozygotic and dizygotic pairs of twins. Children were submitted to an empathy test (reactions to someone simulating distress), and their responses showed that prosocial behaviour was weakly predicted by both genetic and environmental factors, but more strongly predicted from empathy (Knafo *et al.*, 2008). The same study, measuring children at 14, 20, 24 and 36 months also showed that empathy is a stable dispositional trait, to which the contribution of genetic effects tends to increase with age, stabilizing between 24 and 36 months, whereas the environmental effects tend to decrease within this age range (Knafo *et al.* 2008).

Emotional empathy is largely hardwired and nested on evolutionary processes (Castro *et al.*, 2010). For example, a link was established between certain neural networks and the proneness to sympathetic prosocial behaviour and moral judgement (Decety and Michalska, 2010). Notwithstanding, the role of genetic variation in children's temperament has been downplayed. In fact, the existence of an underlying genetically coded empathy trait, consistently expressed by children, has been well supported with data (Knafo *et al*, 2008, 2009) and its heritability shown to reach h=.53 at 3 years of age. Relevant criticism to heritability studies point out the effects different individuals produce on their environments, thereby reinforcing heritable traits. Parents need to understand emotional development, especially the fact that children are not all born alike so as to adjust their own actions accordingly. Often, they are unprepared to respond to children born with irritable, hyperactive temperaments, who are at higher risk of insecure attachment and callousness (Patrick *et al.*, 2009), and they may end up inadvertently contributing to reinforcing aggressive or callous traits.

These biological predispositions to empathy/callousness develop in dynamic and flexible interaction with environmental and contextual factors (Decety and Svetlova, 2012). From birth to age 3, there is an accelerated maturation of brain structures that are crucial to empathy, reactiveness, emotion regulation and prosocial behaviour (Schore, 2001; Smith, 2006), demanding special attention from caregivers, and after which response modes tend to fixate (Essex *et al.*, 2002; Heim *et al.*, 2002). Luby and colleagues (2012) provide evidence of the effect of mother's early (ages 3 to 5 years old) nurturing behaviour on the development of the hippocampus (measured at 7 to 13 years old), a brain structure crucial to emotion processing, mood and memory. Deprived of basic contact and comfort from their mothers or caregivers, especially during the first three years of life, children tend to incur into severe brain damage and disturbed behaviour that may be irreversible in many cases (Chugani *et al.*, 2001).

Early education intervention is highly predictive of academic success and socialization years ahead (Mervis, 2011; Clemens and Sarama, 2011) while secure attachment to parents predicts positive emotional and cognitive

outcomes in the long run (Schore, 2001). Parents' personal traits and parenting styles also leave a strong trace on a child's emotional and moral development. Their responsiveness to the child's needs predicts not only the establishment of secure attachment, but also contributes to cognitive development, academic achievement and social adjustment (Baumrind, 1991; Maccoby, 1992). Even in children and adolescents who present early psychopathic traits associated with a genetic polymorphism in an oxytocin receptor gene (Beitchman *et al.*, 2012), the development of callousness and pervasive aggression is contingent with an environmental facilitator – exposure to violence; the absence of violent models inhibits the expression of the aggressive behaviour in boys with this tendency.

Hence, we can make a case of nature *via* nurture whereby the contribution of the environment to the development of empathy and prosocial behaviour is constrained by maturation windows of opportunity that require timely types of interaction. Predispositions do not dictate what a child will become. The child creates opportunities and challenges to serve his or her dispositional traits, but personality unfolds with possible interactions, in braided and inextricable processes. What seems certain is that empathy and moral models strongly participate in personality development.

Carl Rogers (1959, 1969) presented three pillars of personality development: 1) 'unconditional positive regard' (i.e. mother love and unconditional positive attention), 2) openness (i.e. being able to express one's own views and true self), and 3) empathy. Temperament, the component of personality that includes emotional reactivity and control, has been associated with empathy in developing children, particularly through a bias to experience fear (Van der Mark *et al.* 2002). At about 18 months, children who show signs of empathy also display strong signs of fearfulness whereas the most fearless children show much lower empathy. Whilst empathy can trigger prosocial behaviour, it can also generate empathic distress and aggression towards the victims' aggressor. For every child, emotion regulation is a challenge, and achieving it is a crucial developmental task. It departs from the primary mother-infant relationship with full regulation by the mother to a gradual self-regulation influenced by the growing environment. Comparative studies of cultures have provided insights on prosocial behaviour and regulatory differences in emotional empathy, as a function of culture (Cassels *et al.* 2010; Greck *et al.*, 2012). Additionally, brain studies show that areas involved in emotion regulation overlap consistently with those crucial to empathy (Schore, 2001).

To sum up, evidence seems to confirm that empathy, more than moral codes, predicts the development of prosocial behaviour, and that parental behaviour plays a major role in shaping prosociality. Morality itself seems to stem from empathy. Frans de Waal, a paladin of the intrinsic goodness of the human (and non-human) primate, stresses the need to overcome the good-evil dualism and relativize the goodness: an approach that is neither supportive of the 'good savage' nor of the 'selfish child', whereby developing children are not struggling against genetic predisposition of any kind but being 'nice enough' to accommodate genetic tendencies to become prosocial beings (de Waal, 2001).

Evolutionary perspectives on empathy and moral sense

Non-human primates have been the major targets of research looking for the 'good savage' in human origins due to their phylogenetic proximity to humans. However, basic emotional responses are shared among vertebrates, and social affective responses are seen in all mammals, who share the brain emotional networks involved in these responses (Panksepp, 1998, 2011), which suggests that we should be looking for the ancient roots of empathy much further back in our common ancestry with mammals. Researchers have focused their research efforts in seeking clues of consolation, cooperation and altruism, since these are the key manifestations of prosocial behaviour.

Consolation has been defined as reassuring behaviour by an uninvolved bystander to one of the combatants in a preceding aggressive incident (de Waal, 2006). An example, quite often observed in chimpanzees is when a third individual goes over to the loser of a fight and gently puts an arm around his or her shoulders; sometimes several chimpanzees do it in sequence or almost at once (Gaspar, 2001). Recently, both emotional contagion of distress and consolation were reported in Asian elephants (Plotnick and de Waal, 2014).

Cooperation is probably the easiest to document among non-human animals. It takes place in social animals whose lives depend on each other for major survival tasks such as hunting, nest and territory vigilance or raising offspring. 'Tit-for-tat' reciprocity, also known as 'reciprocal altruism' (a term coined by Trivers, 1971), one of the social cognition modes of morality described by Fiske in human societies (1991, 1992), is a common feature of non-human primate interactions even when it involves participating in fights and 'warfare'. Individual chimpanzees or baboons remember who aided them and repay, sometimes with great delay, which could suggest the moral emotion gratitude (de Waal, 2006). Likewise, they also seek revenge. As de Waal (2006) reports, zoos and animal exploiters are an abundant source of revenge stories, common with elephants, 'who never forget', chimpanzees and otherwise quite affectionate species, such as camels or cetaceans. Fiske's communal sharing model of morality corresponds very much to the sharing behaviour de Waal describes in chimpanzees and bonobos (de Waal, 1997; de Waal and Lanting, 1997).

Regarding altruism, Mark Bekoff and Jessica Pierce (2009) compiled an impressive account of altruistic behaviour in a variety of mammals. Some accounts are anecdotal whereas others come from lab experiments. Reports with rodents are abundant and impressive. For example, one with baby mice accidentally trapped in a sink: one, exhausted, frightened and unable to climb up the slick sides, sees another risking falling into the water and to reach the other, even more exhausted and paralysed with fear, extends him food. Another is an account of Church's (1959) paper 'Emotional Reactions of Rats to the Pain of Others' in which rats were trained to press a lever in order to get a food reward. In a neighbouring cage where the floor consisted of an electric grid that could be turned on when a rat in the first cage pressed the food lever, the pain caused by the electric shock to the second rat seemed to be evident to its neighbour,

as rats would not push the food lever if they could see that a fellow rat would receive a shock. Bekoff considers empathy the most parsimonious explanation for the rat's behaviour of withdrawing from pressing the lever to eat. These experiments are identical in apparatus and results to Milgram's experiments with rhesus monkeys (Gaspar, 2007). They stress that the altruistic behaviour in these extreme conditions is less exceptional than once thought, because high-cost true altruistic acts (different from cooperative/helping acts) have generally been deemed within the scientific community as unique to humans. Helping behaviour that does not involve the high cost of altruism is commonly seen in social animals. Bekoff compiles moving examples in rats, and de Waal (2006, 2008, 2010) in elephants, dolphins and of course, countless reports on primates.

The above examples with animals from such different taxa are by no means coincidental. Preston and de Waal (2002) proposed the perception-action model (PAM), which is currently largely supported by the neurosciences, whereby empathy is compared to a Russian doll, comprising bottom-up processes where cognitive empathy stems from emotional components of empathy, such as emotional contagion, mirror emotions and empathic distress, which are common to humans and other mammals alike (for details see for example Castro *et al.*, 2010).

A sense of fairness has been shown not only in apes but also in monkeys. When confronted with the fact that their reward for successful completion of a task is less valuable than that received by an experiment mate who does not succeed, they become aggressive, and often throw the reward to the experimenter (de Waal, 2006).

Outside empathy and in the strict realm of morality, Fiske's authority ranking is paralleled by the unambiguous deference and mutual obligations that both wild and captive chimpanzees display within their communities, big or small (de Waal, 1982; Goodall, 1986). Indeed, chimpanzee cultures include codes of conduct that make quite clear what the postures and behaviours that an individual must display when meeting another of his/her own community are. Breaking the rules leads to harsh punishments, from severe beating to complete banishment (Goodall, 1986; Nishida and Hosaka, 1996).

Among bonobos, hierarchy is not so emphasized (Kano, 1992). Coincidentally, behaviour is much more flexible and not displaying deference in encounters with higher ranking conspecifics does not bear grave consequences (Gaspar, 2001; Preuschoft and Van Hooff, 1995). Preuschoft and Van Hoof formulated the 'power asymmetry hypothesis' to explain this contrast between the stereotype of chimpanzee social behaviour and the flexibility of signals within bonobo communities. It consists on the theoretical claim that ambiguity is down-selected in societies with strong power asymmetries whereas in egalitarian societies (such as those of bonobos) many signals are under neutral selection and may overlap across contexts. In the latter, social innovation may take place much more smoothly.

Paul Bloom (2010) disagrees with a vision of morality driven by our evolutionary past and by hardwired gut reactions, arguing instead that morality

undergoes social evolution and changes throughout one's life. He illustrates this view with poignant examples such as our current loathing of slavery, child labour or animal abuse, which would not be so 200 years ago. It is so, indeed, but should we forget that power asymmetries have also diminished in the recent cultural evolution of western populations? This shift is probably what has allowed the questioning of traditional norms and the updating of values (Gaspar, 2014b), down-selecting those that are probably the least universal and the least fair, as also suggested by the contrasting examples of chimpanzee and bonobo societies.

Concluding remarks: putting it all together – where phylogeny, ontogeny and culture meet

We have seen in examples above, compiled by Bekoff and de Waal, that there is ample support for a 'wild justice' that regulates mammal societies' boundaries of good and evil. We have also seen that emotional contagion, empathic distress, and prosocial behaviour, and within it, manifestations of empathic concern, sympathy, cooperation and altruism, unfold during ontogenetic development, initially before the child develops the cognitive ability to understand the feelings of others, their behaviour and their particular perspective on events. As the child grows, values and models intertwine with this constantly unzipping maturation programme, becoming part of the core basic empathic triggers and responses.

When Haidt and colleagues set out to test Hume's theory that emotion (passions) prevails over reason (Hume (1739/1740; 1969)), they found that the cognitive perspective follows an emotional reaction and not otherwise, as we have seen in the first section (Haidt *et al*, 2000). Haidt's social intuitionist model fits a perspective where biological triggers for certain emotional experiences anticipate a 'post hoc' reasoning link to moral values (Haidt, 2001). But it is not just the hardwired biological programming that is playing a role. Moral norms and social influences also affect the appraisal of the situation and, consequently, the emotional moral response. Likewise, we can retrieve from Haidt and colleagues' work support for the view that these moral pattern recognition systems are largely hardwired in the brain, corresponding to mechanisms of basic information processing that evolved prior to language, because they serve human needs in an adaptive fashion.

Cross-cultural folktales are an amazing account of how moral norms have changed throughout the centuries, and of what has been preserved. They provide important clues to whatever hardwired mechanisms we humans use to deal with moral issues, including values such as fair punishment, obedience, generosity and sense of a 'just world'. An example of what has changed concerns the cruelty involved in punishing serious offenders, which has clearly diminished over time in contemporary western countries. 'De-humanizing' those labelled evil, by stripping them of qualities that define 'human' and by devaluing their lives in order to administer a death penalty or a cruel punishment is a feature of moral sense that has not faded out. Tales like *Little Red Riding Hood*, *Hansel and Gretel*,

Cinderella and *Snow White* teach that the evil characters have terrible endings – the wolf in *Little Red Riding Hood* is killed and cut open, the witch in *Hansel and Gretel* is shoved into the oven; in *Cinderella*, the two evil sisters are punished by having their eyes picked by doves, leaving them blind for the rest of their lives; and, in *Snow White* the witch queen is invited to Snow White's wedding where she is forced to step into a pair of burning-hot iron shoes and to dance in excruciating pain until she dies. In all, evilness is punished with cruelty and we get a sense that the lives of evil ones are less valuable. In *Hansel and Gretel*, for example, the witch is referred to as 'ungodly', reinforcing this idea. *Little Red Riding Hood* emphasizes obedience and the heroine suffers the harsh consequences of her wits and decision to overlook her mother's advice. In *Hansel and Gretel*, Gretel grabs the opportunity to save herself and her brother (who was about to be cooked first) and it is she who shoves the witch into the fire, which seems to convey the message that it is acceptable to kill to save one's life.

Like de Waal and Pinker, I take an optimistic view. We have enough information, as of now, to clearly establish that despite surface appearance, moral codes and the processes that generate them are not arbitrary. Wrapped up in cultural norms, they reveal strong foundation on human biology and humanity's shared history of cultural evolution and are strong promoters of care for one's family and community and for those in need, as well as of prevention of harm and defence of life. Empathy goes beyond this in that the prosocial conduct that springs from it is not bound to any specific norms. Its triggers are so strong that anybody can be the target of cooperation or altruism, even members of a different species. Regarding the origins of moral values, empathy seems to be, by far, the primal source, a motivational experience that nature provided to ensure maternal care, and in ever larger circles of empathy, care for one's relatives, peers, conspecifics and beyond. We have seen that moral judgement and moral conduct entail both cognitive and emotional paths, but as researchers like Haidt and Hoffman uphold and document, morality stems most strongly and coherently from emotional experiences, that are, indeed, empathic ones. While growing up, children can be exposed to empathy boosting interactions, resulting in the kinds of prosocial conduct advocated by cross-cultural moral norms. Strategies to ferment prosociality are currently under scrutiny in psychological and educational sciences, revealing an unprecedented concern with the morality of future societies.

References

Bandura, A., Ross, D. and Ross, S. A. (1961). Transmission of Aggression Through the Imitation of Aggressive Models. *Journal of Abnormal and Social Psychology*, 63 (3): 575–582. doi:10.1037/h0045925. PMID 13864605.

Batson, C. D. (1987). Prosocial Motivation: Is It ever Truly Altruistic? In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology,* vol 20 (pp. 65–122). San Diego, CA: Academic Press.

Batson, C. D. (1991). *The Altruism Question: Toward a Social-Psychological Answer*. Hillsdale, NJ: Erlbaum.

Batson D. C., Thompson, E. R. and Chen, H. (2002). Moral Hypocrisy: Addressing Some Alternatives. *Journal of Personality and Social Psychology*, 83(2): 330–339.

Batson D. C., Kobrynowicz, D., Dinnerstein, J. L., Kampf, H. C. and Wilson, A. D. (1997). In a Very Different Voice: Unmasking Moral Hypocrisy. *Journal of Personality and Social Psychology*, 72 (6): 1335–1348.

Batson, C. D., Duncan, Bruce D., Ackerman, P., Buckley, T., Birch, K. (1981). Is Empathic Emotion a Source of Altruistic Motivation? *Journal of Personality and Social Psychology*, 40(2): 290–302.

Baumrind, D. (1991). The Influence of Parenting Style on Adolescent Competence and Substance Use. *Journal of Early Adolescence*, 11(1): 56–95.

Beitchman, J. H., Zai, C. C., Muir, K., Berall, L., Nowrouzi, B., Choi, E. and Kennedy, J. L. (2012). Childhood Aggression, Callous-unemotional Traits and Oxytocin Genes. *European Child and Adolescent Psychiatry*, 2: 125–132.

Bekoff, M. and Pierce, J. (2009). *Wild Justice. The Moral Lives of Animals*. Chicago, IL: University of Chicago Press.

Berkowitz, M. W. and Gibbs, J. C. (1985). The process of moral conflict resolution and moral development. *New Directions for Child and Adolescent Development*, 1985: 71–84.

Blair, R.J. (2005). Responding to the emotions of others: dissociating forms of empathy through the study of typical and psychiatric populations. *Conscious Cognition*, 14(4): 698–718.

Blasco, P. G. and Moreto, G. (2012). Teaching Empathy through Movies: Reaching Learners' Affective Domain in Medical Education. *Journal of Education and Learning*, 1(1): 22–34.

Bloom, P. (2010). How do Morals Change? *Nature*, 446: 490.

Bower, B. (2009). Morality Play. *Science News*, 12 September: 16–19.

Brown, D. E. (1991). *Human Universals*. Boston, MA: MacGraw-Hill.

Cassels, T. J., Chan, S., Chung, W. and Birch, S. A. J. (2010). The Role of Culture in Affective Empathy: Cultural and Bicultural Differences. *Journal of Cognition and Culture*, 10: 309–326

Castro, R., Gaspar, A. and Vicente, L. (2010). The Evolving Empathy: Hardwired bases of Human and Non-human Primate Empathy, *Psicologia*, 24(2): 131–152.

Chugani, H. T., Behen, M. E., Muzik, O., Juha'sz, C., Nagy, F. and Chugani, D. C. (2001) Local Brain Functional Activity Following Early Deprivation: A Study of Post Institutionalized Romanian Orphans. *NeuroImage*, 14: 1290–1301.

Church, R. M. (1959). Emotional Reactions of Rats to the Pain of Others. *Journal of Comparative and Physiological Psychology*, 52: 132–134.

Clemens, D. H. and Sarama, J. (2011) Early Childhood Mathematics Intervention. *Scientific American Special Issue on Education*: 968–970. doi: 10.1126/science.1204537.

Correia, I. and Dalbert, C. (2008). School Bullying. Belief in a Personal Just World of Bullies, Victims, and Defenders. *European Psychologist*, 13: 248–254.

Damon, W. (1988). *The Moral Child: Nurturing Children's Natural Moral Growth*. New York: Free Press.

Damon, W. (1999). The Moral Development of Children. *Scientific American*, Aug. 19: 72–78.

Davis, M. H. (1983). Measuring Individual Differences in Empathy: Evidence for a Multidimensional Approach. *Journal of Personality and Social Psychology*, 44: 113–126.

Decety, J. and Michalska, K. J. (2010). Neurodevelopmental Changes in the Circuits Underlying Empathy and Sympathy from Childhood to Adulthood. *Developmental Science*, 13(6): 886–899.

Decety, J. and Svetlova, M. (2012). Putting Together Phylogenetic and Ontogenetic Perspectives on Empathy. *Developmental Cognitive Neuroscience*, 2: 1–24.

de Waal, F. B. M. (1982). *Chimpanzee Politics: Power and Sex among Apes.* Baltimore, MD: Johns Hopkins University Press.

de Waal, F. B. M. (1997). *Good Natured: The Origins of Right and Wrong in Humans and Other Animals*. Cambridge, MA: Harvard University Press.

de Waal, F. B. M. (2001). *The Ape and the Sushi Master*. New York: Basic Books.

de Waal, F. B. M (2006). *Primates and Philosophers: How Morality Evolved*. Princeton, NJ: Princeton University Press.

de Waal, F. B. M. (2008). Putting the Altruism back to Altruism: The Evolution of Empathy. *Annual Review of Psychology*, 59: 279–300.

de Waal, F. B. M. (2010). *The Age of Empathy*. London: Crown Publishing Group.

de Waal, F. B. M. and Lanting, F. (1997). *Bonobo: The Forgotten Ape*. Berkeley, CA: University of California Press.

De Vied, M., Van Boxtel, A., Posthumus, J. A., Goudena, P. P., Matthys, W. (2009). Facial EMG and Heart Rate Responses to Emotion-inducing Film Clips in Boys with Disruptive Behavior Disorders. *Psychophysiology*, 46, 996–1004.

Durkheim, E. (1925). *Moral Education*, Glencoe, IL: The Free Press.

Eisenberg, N. and Mussen, P. (1989). *The Roots of Prosocial Behaviour in Children*. Cambridge: Cambridge University Press.

Eisenberg, N. and Strayer, J. (1987). Critical Issues in the Study of Empathy. In N. Eisenberg & J. Strayer (Eds.), *Empathy and its Development* (pp. 3–13). Cambridge: Cambridge University Press.

Emauz, A., Gaspar, A., Esteves, F. and Carvalhosa, S. F. (2016). Adaptação da escala de empatia com animais (EEA) para a população portuguesa. *Análise Psicológica*, XXXIV,1.

Emauz, A., Gaspar, A. and Esteves, F. (in revision). Preditores de empatia com animais – um estudo trans-cultural.

Essex, M., Klein, M., Cho, E. and Kalin, N. (2002). Maternal Stress Beginning in Infancy May Sensitize Children to Later Stress Exposure: Effects on Cortisol and Behavior. *Biological Psychiatry*, 52: 776–784.

Fiske, A. P. (1991). *Structures of Social Life*. New York: Free Press.

Fiske, A. P. (1992). Four Elementary Forms of Sociality: Framework for a Unified Theory of Social Relations. *Psychological Review*, 99: 689–723.

Fowler, J. W. (1981). *Stages of Faith*. New York: Harper & Row.

Frick, P. J., Kimonis, E. R., Dandreaux, D. M. and Farell, M. S. (2003) The Four Year Stability of Psychopathic Traits in Non-Referred Youth. *Behavioral Science Law*, 21: 713–736.

Gallese V. (2001). The 'Shared Manifold' Hypothesis: From Mirror Neurons to Empathy. *Journal of Consciousness Studies*, 8: 33–50.

Gallese, V., Keysers, C. and Rizzolatti, G. (2004). A Unifying View of the Basis of Social Cognition. *Trends in Cognitive Sciences,* 8(9): 396–403.

Gaspar, A. D. (2001). *Facial Behavior in Pan and Homo. Contribution to the Evolutionary Study of Facial Expressions*. Doctoral dissertation. Lisbon, Universidade Nova de Lisboa.

Gaspar, A. D. (2007). A ética sem criacionismo – história evolutiva de uma ética natural. In A. Gaspar (Ed.), *Evolução e criacionismo: uma relação impossível*. Vila Nova de Famalicão: Quasi.

Gaspar, A. D. (2014a). Neurobiologia e psicologia da empatia. Pontos de partida para a investigação e intervenção da promoção da empatia. In: P. Henenberg and A.C. Caldas (Eds), Cérebro: o que a ciência nos diz. *Povos e Culturas* 18, 159– 174.

Gaspar, A. D. (2014b). Como a evolução elucida a ética. De anjos e demónios, à empatia entre 'nós' e os 'Outros'. *Homem: origem e evolução. Série fundamentos e desafios do* evolucionismo, volume 4: Lisboa: Glaciar.

Gaspar, A., Rocha, S. & Esteves, F. (in press, 2016) Developing Children's Ability to Recognize Animal Emotions – What Does it Take ? A Study at the Zoo. *Human Animal Interaction Bulletin*.

Gibbs, J. C., Basinger, K. S., Grime, R. L. and Snarey, J. R. (2007). Moral Judgement Development Across Cultures: Revisiting Kohlberg's Universality Claims. *Developmental Review* 27, 443–500.

Gilligan, C. (1982). *In a Different Voice: Psychological Theory and Women's Development*. Cambridge: Cambridge University Press.

Golding, W. (1954/1958). *Lord of the Flies*. Boston: Faber & Faber.

Goodall, J. (1986). *The Chimpanzees of Gombe: Patterns of Behavior*. Cambridge, MA: Belknap Press of Harvard.

Goodall, J., Lindsey, J. and Grosvenor, G. M: (1999). *Jane Goodall: 40 Years at Gombe*, The Jane Goodall Institute. New York: Harry N. Abrams, Inc.

Greck, M., Shi, Z., Wang, G., Zuo, X., Yang, X., Wang, X., Northoff, G. and Han, S. (2012). Culture Modulates Brain Activity during Empathy with Anger. *NeuroImage*, 59: 2871–2882.

Grühn, D., Rebucal, K., Diehl, M., Lumley, M. and Labouvie-Vief, G. (2008). Empathy Across the Adult Lifespan: Longitudinal and Experience-Sampling Findings. *Emotion*, 8(6): 753–765. doi:10.1037/a0014123.

Haidt, J. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review*, 108(4): 814–834.

Haidt, J. and Joseph, C. (2004). Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues. *Daedalus*, 133(4): 55–66.

Haidt, J., Koller, S. H. and Dias, M. G. (1993). Affect, Culture, and Morality, or Is It Wrong to Eat your Dog? *Journal of Personality and Social Psychology*, 65, 613–628.

Haidt, J., Björklund, F. and Murphy, S. (2000). *Moral Dumbfounding: When Intuition Finds No Reason*. Manuscript, University of Virginia.

Heim, C., Newport, D., Wagner, D., Wilcox, M., Miller, A. and Nemeroff, C. (2002). The Role of Early Adverse Experience and Adulthood Stress in the Prediction of Neuroendocrine Stress Reactivity in Women: A Multiple Regression Analysis. *Depression and Anxiety*, 15:117–12520.

Hoffman, M. L. (1975). Altruistic Behavior and the Parent–Child Relationship. *Journal of Personality and Social Psychology*, 31: 937–943.

Hoffman, M. L. (2000). *Empathy and Moral Development*. Cambridge: Cambridge University Press.

Hume, D. (1739/1740; 1969). *A Treatise of Human Nature*. London: Penguin.

Iacoboni, M., Moinar-Szakacs, I., Gallese, V., Buccino, G., Mazziota, J. C. and Rizzolatti, G. (2005). Grasping the Intentions of Others with One's Own Mirror Neuron System. *PLoS Biology*, 3(3): 0001–0007.

Jabbia, M., Kippenhana, J. S., Kohna, P., Marencob, S., Mervisc, C. B. Morris, C. A. Meyer-Lindenberga, A. and Bermana, K. F. (2012). The Williams syndrome chromosome 7q11.23 hemideletion confers hypersocial, anxious personality coupled with altered insula structure and function. *Proceedings of the National Academy of Sciences*, 109: 5164–5165.

Kano, T. (1992). *The Last Ape: Pygmy Chimpanzee Behavior and Ecology*. Stanford, CA: Stanford University Press.

Kidd, D. C. and Castano, E. (2013) Reading Literary Fiction Improves Theory of Mind. *Science*, 342(6156): 377–380. doi: 10.1126/science.1239918.

Knafo, A., Zahn-Waxler, C., Van Hulle, C. and Robinson, J. L. (2008). The Developmental Origins of a Disposition Toward Empathy: Genetic and Environmental Contributions. *Emotion*, 6, 737–752.

Knafo, A., Zahn-Waxler, C., Davidof, M., Van Hulle, C., Robinson, J. L. and Rhee, S. H. (2009). Empathy in Early Childhood. Genetic, Environmental, and Affective Contributions. Values Empathy and Fairness across Social Barriers. *Annals of the New York Academy of Sciences*, 1167: 103–114.

Kohlberg, L. (1969). Stage and Sequence: The Cognitive-developmental Approach to Socialization. In D. A. Goslin (Ed.), *Handbook of Socialization Theory and Research* (pp. 325–480). Chicago, IL: Rand McNally.

Kohlberg, L. (1981) *The Meaning and Measurement of Moral Development*. Worcester, MA: Clark University, Heinz Werner Institute.

Kohlberg, L. and Candee, D. (1984). The Relationship of Moral Judgment to Moral Action. In L. Kohlberg (Ed.), Essays in Moral development, vol. 2. The Psychology of Moral Development (pp. 498–581). New York: Harper & Row.

Larsson, H., Andershed, A. and Lichtenstein, P. (2006). A Genetic Factor Explains most of the Variation in the Psychopathic Personality. *Journal of Abnormal Psychology*, 115: 221–230.

Liebert, R. M. (1979). Moral Development: A Theoretical and Empirical Analysis. In G. J. Whitehurst and B. J. Zimmerman (Eds.). *The Functions of Language and Cognition* (pp. 199–264). New York: Academic Press.

Luby, J. L., Barch, D. M., Belden, A. Gaffrey, M. S., Tillmwn, R., Babb, C., Nishino, T., Suzuki, H. and Botteron, K. N. (2012). Maternal Support in Early Childhood Predicts Larger Hippocampal Volumes at School Age. *Proceedings of the National Academy of Sciences,* 109, 8: 2854–2859.

Maccoby, E. E. (1992). The Role of Parents in the Socialization of Children: An Historical Overview. *Developmental Psychology*, 28: 1006–1017.

Mervis,J (2011) Past Success Shapes Effort to Expand Early Intervention. *Sci. Am. Special Issue on Education*: 952–956. doi: 10.1126/science.333.6045.952

Nishida, T., and Hosaka, K. (1996). Coalition Strategies Among Adult Male Chimpanzees of the Mahale Mountains, Tanzania. In W. C. McGrew, L. F. Marchant and T. Nishida (Eds.), *Great Ape Societies* (pp. 114–134). Cambridge: Cambridge University Press.

Panksepp, J. (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. New York: Oxford University Press.

Panksepp, J. (2011). Cross-species Affective Neuroscience Decoding of the Primal Affective Experiences of Humans and Related Animals. *PLOS ONE*, 6(9): e21236.

Patrick, C. J., Fowles, D. C. and Krueger, R. F. (2009). Triarchic Conceptualization of Psychopathy: Developmental Origins of Disinhibition, Boldness, and Meanness. *Developmend &Psychopathology,* 21: 913–938.

Paul, E. (2000). Empathy with Animals and with Humans: Are they Linked? *Anthrozoos,* 13(4): 194–202.

Paul, E. S. and Serpell, J. A. (1993). Childhood Pet Keeping and Humane Attitudes in Young Adulthood. *Animal Welfare,* 2: 321–337.

Piaget, J. (1932/1965). *The Moral Judgment of the Child.* New York: Free Press.

Pinker, S. (2011). *The Better Angels of our Nature. Why Violence has Declined*. New York: Penguin Books.

Plotnick, J. M. and de Waal, F. B. M. (2014). Asian Elephants (*Elephas maximus*) Reassure Others in Distress. *PeerJ*, 2: e278.

Premack, D. and Woodruff, G. (1978). Does the Chimpanzee Have a Theory of Mind? *The Behavioral and Brain Sciences*, 4(5): 5–526.

Preston, S. D. and de Waal, F. B. M., (2002). Empathy: It's the Ultimate and Proximate Bases. *Behavioral and Brain Sciences*, 25: 1–20.

Preuschoft, S. and van Hooff, J. A. R. A. M. (1995). Homologizing Primate Facial Displays: A Critical Review of Methods. *Folia Primatologica*, 65: 121–137.

Rogers, Carl. (1959). A Theory of Therapy, Personality and Interpersonal Relationships as Developed in the Client-centered Framework. In S. Koch (ed.), *Psychology: A Study of a Science. Vol. 3: Formulations of the Person and the Social Context*. New York: McGraw Hill.

Rogers, C. (1969). *Freedom to Learn: A View of What Education Might Become* (1st ed.) Columbus, OH: Charles Merrill.

Schore, A. N (2001). Effects of Secure Attachment Relationship on Right Brain Development, Affect Regulation, and Infant Mental Health. *Infant Mental Health Journal*, 22(1–2): 7–66.

Shweder, R. A. and Haidt, J. (1993). The Future of Moral Psychology: Truth, Intuition, and the Pluralist Way. *Psychological Science*, 4: 360–365.

Shweder, R. A., Much, N. C., Mahapatra, M. and Park, L. (1997). The 'Big Three' of Morality (Autonomy, Community, and Divinity) and the 'Big Three' Explanations of Suffering. In A. Brandt and P. Rozin (Eds.), *Morality and Health*. New York: Routledge, pp.119– 169.

Smetana, J. (1981). Preschool Children's Conceptions of Moral and Social Rules. *Child Development*, 52: 1333–1336.

Smith, A. (2006). Cognitive Empathy and Emotional Empathy in Human Behavior and Evolution. *The Psychological Record*, 56: 3–21.

Staub, E. (2003). *The Psychology of Good and Evil: Why Children, Adults, and Groups Help and Harm Others*. New York: Cambridge University Press.

Turiel, E. (2000). The Development of Morality. In W. Damon, R. M. Lerner and N. Eisenberg (Eds.), *Handbook of Child Psychology: Vol. 3: Social, Emotional, and Personality Development* (5th ed) (pp. 863–932). Hoboken, NJ: Wiley.

Turiel, E. (2002). *The Culture of Morality: Social Development, Context, and Conflict*. Cambridge: Cambridge University Press.

Trivers, R. L. (1971). The Evolution of Reciprocal Altruism. *The Quarterly Review of Biology*, 46 (1): 35–57. doi:10.1086/406755.

Vail, R. V. (2011). *Children and Their Development* (6th ed.). Boston, MA: Pearson.

van der Mark, I. L. van Ijzendoorn, M. H. and Bakermans-Kranenburg, M. J. (2002). Development of Empathy in Girls During the Second Year of Life: Associations with Parenting, Attachment, and Temperament. *Social Development*, 11(4): 451–468.

Webster-Stratton, C. (1998). Preventing Conduct Problems in Head Start Children: Strengthening Parenting Competencies. *Journal of Consulting and Clinical Psychology*, 66(5): 715–730.

Zahn-Waxler, C. and Radke-Yarrow, M. (1990). The Origins of Empathic Concern. *Motivation and Emotion*, 2: 107–130.

5 Software sans emotions but with ethical discernment

Luís Moniz Pereira

Introduction

Some of our previous research (Pereira & Saptawijaya, 2011; Han *et al.*, 2012; Pereira & Saptawijaya, 2015a, 2015b; Saptawijaya & Pereira, 2015a, 2015b) has focused on using logic programming techniques to computational modelling of morality sans emotions. In the realm of the individual, we have addressed questions of permissibility and the dual-process of moral judgments by framing together ingredients that are essential to moral agency: abduction, integrity constraints, preferences, argumentation, counterfactuals, and updates. Computation over these ingredients has become our vehicle for modelling the dynamics of moral cognition within a single agent, without addressing the cultural dimension (Prinz, this volume), because this is still absent in machines. In the collective realm, we have reported on computational moral emergence (Han *et al.*, 2015a), again sans emotions, using techniques from Evolutionary Game Theory (EGT). We have shown that the introduction of cognitive abilities, like intention recognition, commitment, revenge, apology, and forgiveness, reinforce the emergence of cooperation in diverse populations, comparatively to their absence, by way of EGT models.

In studies of human morality, these distinct but interconnected realms – one stressing above all individual cognition, deliberation, and behaviour; the other stressing collective morals and how they have emerged with evolution – seem separate but are synchronously evinced (Pereira & Saptawijaya, 2015b). There are issues concerned with how to bridge the two realms also addressed in this volume (see, for example, Gaspar, this volume). Our account affords plenty of room for an evolutionary phylogenetic emergence of morality, as illustrated below, thereby supplementing the limitations of focusing just on ontogeny. The bridging issues concern individual cognitive abilities and their deployment in the population: namely the one of recognizing the intention of another, even taking into account how others recognize our intention; the abilities of requesting commitment, and of accepting or declining to commit; those of cooperating or defecting; plus those of apologizing, be they fostered by guilt, and of taking revenge or forgiving.

This chapter relies mainly on our collective realm research, and considers the modelling of distinct co-present strategies of cooperative and uncooperative behaviour. Such driving strategies are associated with moral 'emotions' that motivate moral discernment and substantiate ethical norms, leading to improved general conviviality on occasion, or not. To wit, we can model moral agency without explicitly representing embodied emotions, as we know them. Rather, such software-instantiated 'emotions' are modelled as (un)conscious heuristics empowered in complex evolutionary games.

In the next two sections, starting with the ground breaking work of Alan Turing, functionalism is employed to scaffold a philosophical perspective on emotions and morality. The further five sections after those review materials from our EGT-based research in support of this perspective. This work has substantiated the philosophical viewpoint through an admixture of intention recognition, commitment, revenge, apology, and forgiveness. The final section conjectures on guilt, and its relationship with counterfactual reasoning, as a next natural step in our research programme.

Turing is among us

Turing's relevance arises from the timelessness of the issues he tackled, and the innovative light he shed upon them (Pereira, 2012). He first defined the algorithmic limits of computability, via an *effective* well-specified mechanism, and showed the generality of his definition by proving its equivalence to other general, but less algorithmic and non-mechanical, more abstract formulations of computability. His originality lies on the essential simplicity of the mechanism invoked – the now dubbed Turing Machines (or programs), which he called A-Machines – and the proof of existence of a Universal A-Machine (i.e. the digital computer, known in academia as the Universal Turing Machine), which can simulate any other A-Machine, that is, execute any program.

Interestingly, he raised the issue of whether human beings are a measure for his 'machines', and, in mechanizing human cognition, Turing implicitly introduced the modern perspective since known as 'functionalism'. According to this paradigm, what counts is the realization of function, independently of the hardware embodying it. Such 'multiple realization' is afforded by the very simplicity of his devised mechanism, relying solely on the manipulation of discrete information, where data and instructions are both represented just with symbols. The twain are stored in memory, instructions doubling as data and as rules for acting – the stored program. To this day, no one has invented a computational mechanical process with such general properties, which cannot be theoretically approximated with arbitrary precision by some A-Machine, where any interactions with the world outside are captured by Turing's innovative concept and definition of 'oracle' – the very word employed by him for the purpose – as a means to interrogate that world by posing queries to one or more outside oracles. This concept of oracle is regularly taught in computer science today, namely in the essential study of computation complexity, though

not every student knows it came from Turing. In the midst of a computation a query may be posed to an outside oracle about the satisfaction of some truth, and the computation continued once an answer obtained, rather than the computer testing for an answer in a possibly infinite set of them.

Turing further claimed that his machines could simulate the effect of *any* activity of the mind, not just a mind engaged upon a 'definite method of proceeding' or algorithm. He was clear that discrete state machines included those with learning or self-organizing abilities, and stressed that these still fall within the scope of the computable. Turing drew attention to the apparent conflict between self-organization and the definition of A-Machines as having fixed tables of behaviour, but sketched a proof that self-modifying machines are still definable by an unchanged instruction set (Hodges, 1997; McDermott, 2001).

The promise of this approach in studies of morality is that it represents a universal functionalism, the terms of which enable the bringing together of the ghosts in the several embodied machines (silicon-based, biological, extra-terrestrial or otherwise), to promote their symbiotic epistemic co-evolution, as they undertake moral action within a common moral theatre.

Functionalism and emergence

The principle of the distinction between software and hardware appears clear-cut with the advent of the digital computer and its conceptual precursor, the Universal Turing Machine. The diversity of technologies employed to achieve the same function, confirms it ever since the first computers. One program is executable in physically different machines, precisely because the details of its execution below an ascertainable level of analysis are irrelevant, as long as an identical result at the level of discourse is produced. That said, however, the distinction between hardware and software is not so clear as it might seem. Hardware is not necessarily represented by physical things but rather by what, at some level of analysis, is considered fixed, given, and whose analysis or non-analysability is irrelevant for the purpose at hand. Historically, in the first computers, that level coincided with that of the physical parts of the machine. Subsequently, especially due to rapidly increasing computing power, 'hardware' has become increasingly 'soft', with the physical basis for the hardware/software distinction finally blurred by the concept of the 'abstract machine': a fixed collection of mathematically defined instructions supporting a set of software functions, independently of the particular physical processes underlying the implementation of the abstract machine, that is, realizing it.

Hence, 'multiple realization' stands for the thesis that a mental state can be 'realized' or 'implemented' by different physical states. Beings with different physical constitutions can thus be in the same mental state, and from these common grounds can cooperate, acting in mutual support (or not). According to classical functionalism, multiple realization implies that psychology is autonomous: in other words, biological facts about the brain are irrelevant

(Boden, 2008). Whether physical descriptions of the events subsumed by psychological generalizations have anything in common is irrelevant to the truth of the generalizations, to their interestingness, to their degree of confirmation, or, indeed, to any of their epistemological important properties (Fodor, 1974).

Functionalism has continued to flourish, being developed into numerous versions by thinkers as diverse as David Marr, Daniel Dennett, Jerry Fodor, and David Lewis (Fodor, 1974; Dennett, 2005). It helped lay the foundations for modern cognitive science, being the dominant theory of mind in philosophy today. In the latter part of the twentieth and early twenty-first centuries, functionalism stood as the dominant theory of mental states. It takes mental states out of the realm of the 'private' or subjective, and gives them status as entities open to scientific investigation. Functionalism's characterization of mental states in terms of their roles in the production of behaviour grants them the causal efficacy that common sense takes them to have. In permitting mental states to be multiply realized, functionalism offers an account of mental states compatible with materialism, without limiting the class of minds to creatures with brains like ours (Levin, 2010).

Biological evolution is characterized by a set of highly braided processes, which produce a kind of extraordinarily complex combinatorial innovation. A generic term frequently used to describe this vast category of spontaneous, and weakly predictable, order-generating processes, is 'emergence'. This term became a sort of signal to refer to the paradigms of research sensitive to systemic factors. Complex dynamic systems can spontaneously assume patterns of ordered behaviours not previously imaginable from the properties of their constitutive elements or from their interaction patterns. There is unpredictability in self-organizing phenomena – preferably called 'evolutionary' (Turing 1950) – with considerable variable levels of complexity, where 'complexity' refers to the emergence of collective properties in systems with many interdependent components. These components can be atoms or macromolecules in physical or biological contexts, and people, machines or organizations in socioeconomic contexts.

What does emerge? The answer is not something defined physically but rather something like a shape, pattern, or function. The concept of emergence is applicable to phenomena in which the relational properties predominate over the properties of the compositional elements in the determination of the ensemble's characteristics. Emergence processes are due to starting configurations and interaction topologies, not intrinsic to the components themselves (Deacon, 2003). This functionalism is, almost by definition, anti-substance-essence, anti-vital-principle, anti-monopoly of *qualia*.

Building intelligent machines may seek a partial understanding of the emergence of higher-level properties, like morality. Here, functionalism affirms the salience of the results of this work in assessing, for example, human morality. Again, functionalism holds that the material substrate is not of the essence, and that it suffices to realize equivalent functionality albeit by way of a different material vehicle. Moreover, distinct roads to the same behaviour may be had, thereby adding to our understanding of what, say, 'general intelligence'

or 'mind' means. Thus, on our estimation, the most fruitful inquires into the nature of 'mind' or 'general intelligence' will certainly include the use of Artificial Intelligence aided in time by the embryonic field of artificial emotions, qua strategies, to simulate complex mental operations, as already foreseen (Turing, 1950).

Learning to recognize intentions and committing resolve cooperation dilemmas

Few problems have motivated the amalgamation of so many seemingly unrelated research fields as has the evolution of cooperation (Nowak, 2006; Sigmund, 2010). Several mechanisms have been identified as catalysers of cooperative behaviour (see survey in Nowak (2006) and Sigmund (2010)). Yet these studies, mostly grounded on evolutionary dynamics and game theory, have neglected the important role, which is played by intention recognition (Han & Pereira, 2013) in behavioural evolution. In our work (Han *et al.*, 2011, 2012a), we explicitly studied the role of intention recognition in the evolution of cooperative behaviour. The results indicate that intention recognizers prevail against the most successful strategies in the context of the iterated Prisoner's Dilemma (e.g. win-stay-lose-shift, and tit-for-tat like strategies), and promote a significantly high level of cooperation, even in the presence of noise plus the reduction of fitness associated with the cognitive costs of performing intention recognition. Thus, our approach offers new insights into the complexity of – as well as enhanced appreciation for the elegance of – behavioural evolution when driven by elementary forms of cognition and learning ability.

Moreover, our recent research (Han, *et al.*, 2015a and b) into the synergy between intention recognition and cooperative commitment sheds new light on promoting cooperative behaviour. This work employs EGT methods in agent-based computer simulations to investigate mechanisms that underpin cooperation in differently composed societies. High levels of cooperation can be achieved if reliable agreements can be arranged. Formal commitments, such as contracts, promote cooperative social behaviour if they can be sufficiently enforced, and the costs and time to arrange them provide mutual benefit. On the other hand, an ability to assess intention in others has been demonstrated to play a role in promoting the emergence of cooperation.

An ability to assess the intentions of others based on experience and observations facilitates cooperative behaviour without resort to formal commitments like contracts. Our research found that the synergy between intention recognition and commitment strongly depends on the confidence and accuracy of the intention recognition. To reach high levels of cooperation, commitments may be unavoidable if intentions cannot be assessed with sufficient confidence and accuracy. Otherwise, it is advantageous to wield intention recognition to avoid arranging costly commitments.

Now, conventional wisdom suggests that clear agreements need to be made prior to any collaborative effort in order to avoid potential frustrations for the

participants. We have shown (Han *et al.*, 2013a) that this behaviour may actually have been shaped by natural selection. This research demonstrates that reaching prior explicit agreement about the consequences of not honouring a deal provide a more effective road to facilitating cooperation than simply punishing bad behaviour after the fact, even when there is a cost associated to setting up the agreement. Typically, when starting a new project in collaboration with someone else, it pays to establish up-front how strongly your partner is prepared to commit to it. To ascertain the commitment level one can ask for a pledge and stipulate precisely what will happen if the deal is not honoured.

In our study, EGT is used to show that when the cost of arranging commitments (for example, to hire a lawyer to make a contract) is justified with respect to the benefit of the joint endeavour (for instance buying a house), and when the compensation is set sufficiently high, commitment proposers become prevalent, leading to a significant level of cooperation. Commitment proposers can get rid of fake co-operators that agree to cooperate with them yet act differently, also avoiding interaction with the bad guys that only aim to exploit the efforts of the cooperative ones.

But what happens if the cost of arranging the commitments is too high compared to the benefit of cooperation? Would you make a legal contract for sharing a cake? Our results show that in that case those that free ride on the investment of others will 'immorally' and inevitably benefit. Establishing costly agreements only makes sense for specific kinds of projects. Our study shows that insisting that your partner share in the cost of setting up a deal leads to even higher levels of cooperation, suggesting the evolution of cooperation for a larger range of arrangement costs and compensations. This makes sense, as equal investment will ensure the credibility of the pledge by both partners. Agreements based on shared costs result in better friends.

We also compared this behaviour with costly punishment, a strategy that does not make any prior agreements and simply punishes afterwards. Previous studies show that by punishing strongly enough bad behaviour cooperation can be promoted in a population of self-interested individuals (Fehr & Gachter, 2002). Yet these studies also show that the punishment must sometimes be quite excessive in order to obtain significant levels of cooperation. Our study shows that arranging prior agreements can significantly reduce the impact-to-cost ratio of punishment. Higher levels of cooperation can be attained through lower levels of punishment. Good agreements make good friends indeed.

Emergence of cooperation in groups: avoidance vs. restriction

Public goods, like food sharing and social health systems, may prosper when prior agreements to contribute are feasible and all participants commit to do so. Yet, free-riders may exploit such agreements (Han *et al.*, 2013a), thus requiring committers to decide not to enact the public good when others are not attracted to committing. This decision removes all benefits from free-riders (non-contributors), but also from those who are wishing to establish the beneficial

resource. In (Han *et al.*, 2014) we show, in the framework of the one-shot Public Goods Game (PGG) and EGT, that implementing measures to delimit benefits to 'immoral' free-riders, often leads to more favourable societal outcomes, especially in larger groups and in highly beneficial public goods situations, even if doing so incurs in new costs.

PGG is the standard framework for studying emergence of cooperation within group interaction settings (Sigmund, 2010). In a PGG, players meet in groups of a fixed size, and all players can choose whether to cooperate and contribute to the public good or to defect without contributing to it. The total contribution is multiplied by a constant factor and is then equally distributed among all. Hence, contributors always gain less than free-riders, disincentivizing cooperation. In this scenario, arranging a prior commitment or agreement is an essential ingredient in motivating cooperative behaviour, as abundantly observed both in the natural world (Nesse, 2001) and lab experiments (Cherry and McEvoy, 2013). Prior agreements help clarify the intentions and preferences of other players (Han *et al.*, 2012a). Refusing agreements may be conceived as intending or preferring not to cooperate (the non-committers).

In (Han *et al.*, 2014), we extend the PGG to examine commitment-based strategies within group interactions. Prior to playing the PGG, commitment-proposing players ask their co-players to commit to contribute to the PGG, paying a personal proposer's cost to establish that agreement. If all of the requested co-players accept the commitment, the proposers assume everyone will contribute. Those who commit yet later do not contribute must compensate the proposers (Han *et al.*, 2013a). As commitment proposers may encounter non-committers, they require strategies to deal with these individuals. Simplest is to not participate in the creation of the common good. Yet, this avoidance strategy, AVOID, also removes benefits for those wishing to establish the public good, creating a moral dilemma. Alternatively, one can establish boundaries on the common good, so that only those who have truly committed have (better) access, or so that the benefit of non-contributors becomes reduced. This is the RESTRICT strategy.

Our results lead to two main conclusions: (i) Both strategies can promote the emergence of cooperation in the one-shot PGG whenever the cost of arranging commitment is justified with respect to the benefit of cooperation, thus generalizing results from pairwise interactions (Han *et al.*, 2013a); (ii) RESTRICT, rather than AVOID, leads to more favourable societal outcomes in terms of contribution level, especially when group size and/or the benefit of the PGG increase, even if the cost of restricting is quite large.

Why is it so hard to say sorry?

When making a mistake, individuals are willing to apologize to secure further cooperation, even if the apology is costly. Similarly, individuals arrange commitments to guarantee that an action such as a cooperative one is in the others' best interest, and thus will be carried out to avoid eventual penalties for

commitment failure. Hence, both apology and commitment should go side by side in behavioural evolution. In Han *et al.* (2013b), we studied the relevance of a combination of these two strategies in the context of the iterated Prisoner's Dilemma (IPD). We show that apologizing acts are rare in non-committed interactions, especially whenever cooperation is very costly, and that arranging prior commitments can considerably increase the frequency of apologizing behaviour. In addition we show that, with or without commitments, apology resolves conflicts only if it is sincere, i.e. costly enough. Most interestingly, our model predicts that individuals tend to use a much costlier apology in committed relationships than otherwise, because it helps better identify free-riders, such as fake committers.

Apology is perhaps the most powerful and ubiquitous mechanism for conflict resolution (Abeler *et al.*, 2010; Ohtsubo & Watanabe, 2009), especially among individuals involving in long-term repeated interactions (such as a marriage). An apology can resolve a conflict without having to involve external parties (e.g. teachers, parents, courts), which may cost all sides of the conflict significantly more. Evidence supporting the usefulness of apology abounds, ranging from medical error situations to seller-customer relationships (Abeler *et al.*, 2010). Apology has been implemented in several computerized systems, such as human-computer interaction and online markets, to facilitate users' positive emotions and cooperation (Tzeng, 2004; Utz *et al.*, 2009).

The Iterated Prisoner's Dilemma (IPD) has been the standard model to investigate conflict resolution and the problem of the evolution of cooperation in repeated interaction settings (Axelrod, 1984; Sigmund, 2010). The IPD game is usually known as a story of tit-for-tat (TFT), which won both Axelrod's tournaments (Axelrod, 1984). TFT cooperates if the opponent cooperated in the previous round, and defects if the opponent defected. But if there can be erroneous moves due to noise (i.e. an intended move is wrongly performed), the performance of TFT declines, because an erroneous defection by one player leads to a sequence of unilateral cooperation and defection. A generous version of TFT, which sometimes cooperates even if the opponent defected (Nowak & Sigmund, 1992), can deal with noise better, yet not thoroughly. For these TFT-like strategies, apology is modelled implicitly as one or more cooperative acts after a wrongful defection.

In Han *et al.* (2013b), we describe a model containing strategies that explicitly apologize when making an error between rounds. An apologizing act consists in compensating the co-player an appropriate amount (the higher the more sincere), in order to ensure that this other player cooperates in the next actual round. As such, a population consisting of only apologizers can maintain perfect cooperation. However, other behaviours that exploit this apologetic behaviour could emerge, such as those that accept apology compensation from others but do not apologize when making mistakes (fake apologizers), destroying any benefit of the apology behaviour. Employing EGT (Sigmund, 2010), we show that when the apology occurs in a system where the players first ask for a commitment before engaging in the interaction (Han *et al.*, 2012b and c; Han, 2013), this exploitation can be

avoided. Our results lead to these conclusions: (i) Apology alone is insufficient to achieve high levels of cooperation; (ii) Apology supported by prior commitment leads to significantly higher levels of cooperation; (iii) Apology needs to be sincere to function properly, whether in committed relationships or commitment-free ones (which is in accordance with existing experimental studies, e.g. in Ohtsubo and Watanabe (2009)); (iv) A much costlier apology tends to be used in committed relationships than in commitment-free ones, as it can help better identify free-riders such as fake apologizers: '*commitments bring about sincerity*'.

In Artificial Intelligence and Computer Science, apology (Tzeng, 2004; Utz *et al.*, 2009) and commitment (Winikoff, 2007; Wooldridge & Jennings, 1999) have been widely studied, namely how their mechanisms can be formalized, implemented, and used to enhance cooperation in human-computer interactions and online market systems (Tzeng, 2004; Utz *et al.*, 2009), as well as general multi-agent systems (Winikoff, 2007; Wooldridge & Jennings, 1999). Our study provides important insights for the design and deployment of such mechanisms. For instance, what kind of apology should be provided to customers when mistakes are made, and whether apology can be enhanced if complemented with commitments to ensure cooperation, e.g. compensation for customers who suffer wrongdoing.

Apology and forgiveness evolve to resolve failures in cooperative agreements

Making agreements on how to behave has been shown to be an evolutionarily viable strategy in one-shot social dilemmas. However, in many situations agreements aim to establish long-term mutually beneficial interactions. Our analytical and numerical results (Martínez-Vaquero *et al.*, 2015) reveal for the first time under which conditions revenge, apology and forgiveness can evolve, and deal with mistakes within ongoing agreements in the context of the Iterated Prisoners Dilemma. We showed that, when agreement fails, participants prefer to take revenge by defecting in the subsisting encounters. Incorporating costly apology and forgiveness reveals that, even when mistakes are frequent, there exists a sincerity threshold for which mistakes will not lead to the destruction of the agreement, inducing even higher levels of cooperation. In short, even when to err is human, revenge, apology and forgiveness are evolutionarily viable strategies, playing an important role in inducing cooperation in repeated dilemmas.

Using methods from EGT (Hofbauer & Sigmund, 1998; Sigmund, 2010), we provide analytical and numerical insight into the viability of commitment strategies in repeated social interactions, modelled through the Iterated Prisoners Dilemma (IPD) (Axelrod & Hamilton, 1981). In order to study commitment strategies in the IPD, a number of behavioural complexities need to be addressed. First, agreements may end before the recurring interactions are finished. As such, strategies need to take into account how to behave when the agreement is present and when it is absent, on top of proposing, accepting or

rejecting such agreements in the first place. Second, as shown within the context of direct reciprocity (Trivers, 1971), individuals need to deal with mistakes made by an opponent or by themselves, caused for instance by 'trembling hands' or 'fuzzy minds' (Sigmund, 2010; Nowak, 2006). A decision needs to be made on whether to continue the agreement, or end it collecting the compensation owed from the other's defection.

As errors might lead to misunderstandings or even breaking of commitments, individuals may have acquired sophisticated strategies to ensure that mistakes are not repeated or that profitable relationships may continue. Revenge and forgiveness may have evolved exactly to cope with those situations (McCullough, 2008; McCullough *et al.*, 2011). The threat of revenge, through some punishment or withholding of a benefit, may discourage interpersonal harm. Yet, often one cannot distinguish with enough certainty if the other's behaviour is intentional or just accidental (Han *et al.*, 2011; Fischbacher & Utikal, 2013). In the latter case, forgiveness provides a restorative mechanism that ensures that beneficial relationships can still continue, notwithstanding the initial harm. An essential ingredient for forgiveness, analysed in our work, seems to be (costly) apology (McCullough, 2008), a point emphasized in Smith (2008).

The importance of apology and forgiveness for sustaining long-term relationships has been brought out in different experiments (Abeler *et al.*, 2010; Takaku *et al.*, 2001; Okamoto & Matsumura, 2000; Ohtsubo & Watanabe, 2009). Apology and forgiveness is of interest as they remove the interference of external institutions (which can be quite costly to all parties involved), in order to ensure cooperation.

Creating agreements and asking others to commit to them provides a basic behavioural mechanism present at all the levels of society, playing a key role in social interactions (Nesse, 2001; Sterelny, 2012; Cherry & McEvoy, 2013). Our work reveals how, when moving to repeated games, the detrimental effect of having a large arrangement cost is moderated, for a subsisting commitment can play its role for several interactions. In these scenarios, the most successful individuals are those who propose commitments (and are willing to pay their cost) and, following the agreement, cooperate unless a mistake occurs. But if the commitment is broken then these individuals take revenge and defect in the remaining interactions, confirming analytically what has been argued in McCullough (2008), and in McCullough *et al.* (2011). This result is intriguing as revenge by withholding the benefit from the transgressor may lead to a more favourable outcome for cooperative behaviour in the IPD, as opposed to the well-known reciprocal behaviour such as TFT-like strategies. Forgivers only do better when the benefit-to-cost ratio is high enough.

Yet, as mistakes during any (long-term) relationship are practically inevitable, individuals need to decide whether it is worthwhile to end the agreement and collect the compensation when a mistake is made or whether it is better to forgive the co-player and continue the mutually beneficial agreement. To study this question, the commitment model was extended with an apology-forgiveness mechanism, where apology was defined either as an external or individual parameter in the

model. In both cases, we have shown that forgiveness is effective if it takes place after receiving an apology from the co-players. However, to play a promoting role for cooperation, apology needs to be sincere, in other words, the amount offered in the apology has to be high enough (yet not too high), which is also corroborated by recent experimental psychology (McCullough *et al.*, 2014). This extension to the commitment model produces even higher cooperation levels than in the revenge-based outcome. In the opposite case, fake committers that propose or accept a commitment with the intention taking advantage of the system (defecting and apologizing continuously) will dominate the population. In this situation, the introduction of the apology-forgiveness mechanism destroys the increase of the cooperation level that commitments by themselves produce. Thus, there is a lower-limit on how sincere apology needs to be, as below this limit apology and forgiveness even reduce the level of cooperation one could expect from simply taking revenge. It has been shown in previous works that mistakes can induce the outbreak of cheating or intolerant behaviour in society (Martínez-Vaquero & Cuesta, 2013, 2014), and only a strict ethics can prevent them (Martínez-Vaquero & Cuesta, 2014), which in our case would be understood as forgiving only when apology is sincere.

Commitments in repeated interaction settings may take the form of loyalty (Schneider & Weber, 2013; Back & Flache, 2008), which is different from our commitments regarding posterior compensations, for we do not assume a partner choice mechanism. Loyalty commitment is based on the idea that individuals tend to stay with or select partners based on the length of their prior interactions. We go beyond these works by showing that, even without partner choice, commitment can foster cooperation and long-term relationships, especially when accompanied by sincere apology and forgiveness whenever mistakes are made.

Ohtsubo's experiment (Ohtsubo & Watanabe, 2009) shows that a costlier apology is better at communicating sincerity, and as a consequence will be more often forgiven. This observation is shown to be valid across cultures (Takaku *et al.*, 2001). In another laboratory experiment (Fischbacher & Utikal, 2013), the authors showed apologies work because they can help reveal the intention behind a wrongdoer's preceding offence. In compliance with this observation, in our model, apology best serves those who intended to cooperate but defect by mistake.

Despite the fact that 'to err is human' (Pope, 1711), our research results demonstrate that behaviours like revenge and forgiveness can evolve to cope with mistakes, even when they occur at high rates. Complicating matters is that mistakes are not necessarily intentional, and that even if they are then it might still be worthwhile to continue a mutually beneficial agreement. Here, a sincerity threshold exists whereby the cost of apologizing should exceed the cost of cooperating if the encouragement of cooperation is the goal.

Future work: emotional and counterfactual guilt

A natural extension of our work on intention recognition, commitment, revenge, apology, and forgiveness involves adding guilt, shame, and confession

with surplus apology. We leave shame alone for now as it involves reputation, which we did not address above so as to concentrate on the more basic model of pairwise interactions, without the intrusion of reputational hearsay. Though both have ostensibly evolved to promote cooperation, we believe that guilt and shame can be treated separately. Guilt is an inward phenomenon that can foster apology, and even spontaneous public confession. Shame is inherently public, and it too may lead to apology and request for forgiveness. Shame, however, hinges on being caught, on failing to deceive, and on a mechanism being in place that lets one fall into disrepute.

From an evolutionary viewpoint, guilt is envisaged as an in-built mechanism that tends to prevent wrong doing because of internal suffering that pressures an agent to confess when wrongs are enacted, alongside a costlier apology and penance, plus an expectation of forgiveness to alleviate or dispel the guilt-induced suffering.

The hypothesis, consequently, is that the emergence of guilt within a population is evolutionarily advantageous as it represents an extra-costly apology compared to a non-guilty one, enacted as it is in order to decrease the added suffering. We can test this hypothesis by adapting our existing model comprising commitment, revenge, apology, and forgiveness, via piggybacking guilt onto it. To do so, one introduces a present/absent guilt parameter such that, on defection by a guilt-ridden player, not only is thereby increased the probability of apology (confession), but also the player spontaneously pays a costlier apology, as a means to atone internal guilt. On the other hand, the co-player will more readily accept a guilty extra-valued apology, and forgive. In addition, this co-player's attitude, if copied, will contribute to favour his own forgiveness by others in the population, in case his own super-apologetic confession of guilt replaces of the standard one in the absence of acknowledged guilt. The prediction is that guilt will facilitate and speed-up the emergence of cooperation, in spite of its heavier cost. One reason behind this prediction is that costs of cooperation are compensated for by the costlier guilt apology paid by others. Another reason is that it is in general more conducive to forgiveness, especially in the border cases where the standard apology is outright insufficient.

We know that guilt is alleviated by private confession, e.g. to a priest or psychotherapist, with cost in prayers or fees, plus the renunciation of past failings. In the context of our research, such ersatz confessions and atonements, precisely by exacting a cost, should render temptation to defect less probable – a preference reversal (Correia, this volume) – with the proceeds appertained to some common good (e.g. in a Public Goods Game, or like through charity).

In summary, future research will attempt to show, by simulation if not analytically, that guilt naturally connects with apology and forgiveness mechanisms because of its emergent evolutionary advantage. It seems not too difficult to incorporate into the present framework, by splitting each strategy into one variant experiencing guilt in case of defection, plus a guiltless one. The population at the start would now contain, instead, an admixture of all of both types, for a given fixed cost and extra cost of guilty apology, plus the usual

other parameters, namely a forgiveness threshold. The prediction again is that guilt is evolutionarily advantageous, within a range of the overall parameters defining a starting population composition, via EGT evolution with the usual social imitation of strategies with high payoff success.

This further opens the way to treatment of emotions modelled as strategies, guilt being a widely acknowledged one. It should show that one does not need a specific kind of body (namely an anthropomorphic one) for guilt to serve the role of a moral emotion, useful as it is in population settings where moral cooperation attains good value for all regardless of means of embodiment.

Finally, counterfactual reasoning (Byrne, 2007; Collins *et al.*, 2004; Pereira & Saptawijaya, 2015a) could be wielded to prime and tune guilt. Presupposing that the agent can reason counterfactually, e.g. given the by-now-known sequence of plays by its co-players it might reason: 'Had I before felt guilty instead, and played according to such guilt, then I would have fared better.' As a consequence, the player would then meta-reflectively (Mendonça, this volume) modify its 'feeling level' of guilt for the future.

One could envisage the whole of our above approach as purveying a form of fiction, though recognizably a rather abstract one, yet still adumbrated as per the 'Moral Feelings from rocky fictional ground' (John, this volume), the next chapter in this volume. Indeed, our abstract mathematical and computational fictional simulations might be construed and stretched to fit a bill whereby such fiction would not necessarily offer theorists of emotion or morality immediate embodied evidence, as in novels, say. In contradistinction, it can possibly offer interesting, challenging and conjectural ideas that might benefit the theorizing in these domains. A computer scientist friend bemusedly jokes about my 'soap opera' research, what with intention recognition, commitment proposal, defection, guilt, apology, forgiveness, revenge…

Acknowledgements

Profound thanks are due to my co-authors of joint published work herein cited, without which the personal summing up and specific philosophical viewpoint above would not have been possible at all. Alphabetically: Ari Saptawijaya, Francisco C. Santos, Luis Martínez-vaquero, The Anh Han, and Tom Lenaerts. Moreover, the author thanks Ari Saptawijaya, Sara Graça da Silva and Jeffrey White for comments on prior drafts, and the support from grant FCT/MEC NOVA LINCS PEst UID/CEC/04516/2013.

References

Abeler, J., Calaki, J., Andree, K., & Basek, C. (2010). 'The power of apology'. *Economics Letters,* 107(2): 233–235.

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

Axelrod, R. & Hamilton, W. D. (1981). The evolution of cooperation. *Science* 211, 1390–1396.

Back, I. & Flache, A. (2008). 'The adaptive rationality of interpersonal commitment'. *Rationality and Society*, 20: 65–83.

Boden, M. A. (2008). 'Information and cognitive science'. In P. Adriaans & J. van Bentham (eds.),*Philosophy of Information* (pp. 41–761). Amsterdam: North-Holland, Elsevier.

Byrne, R. M. J (2007). *The Rational Imagination: How People Create Alternatives to Reality*. Cambridge, MA: MIT Press.

Cherry, T. L. & McEvoy, D. M. (2013). 'Enforcing compliance with environmental agreements in the absence of strong institutions: An experimental analysis'. *Environmental and Resource Economics*, 54(1): 63–77.

Collins, J., Hall, N. & L. A. Paul, L. A. (eds.). (2004). *Causation and Counterfactuals*. Cambridge, MA: MIT Press.

Deacon, T. W. (2003). 'The hierarchic logic of emergence: Untangling the interdependence of evolution and self-organization'. In: H. W. Weber, D. J. Depew (eds.), *Evolution and Learning: The Baldwin Effect Reconsidered*. Cambridge, MA: MIT Press, 273–308.

Dennett, D. C. (2005). *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness*. Cambridge, MA: MIT Press.

Fehr, E. & Gachter, S. (2002). 'Altruistic punishment in humans'. *Nature*, 415: 137–140.

Fischbacher, U. and Utikal, V. (2013). 'On the acceptance of apologies'. *Games and Economic Behavior,* 82: 592–608.

Fodor, J. A. (1974). 'Special sciences, or the disunity of science as a working hypothesis'. *Synthèse*, 28: 77–115.

Han, T. A. (2013). 'Intention Recognition, Commitments and Their Roles in the Evolution of Cooperation: From Artificial Intelligence Techniques to Evolutionary Game Theory Models'. *SAPERE series*, 9. Berlin: Springer-Verlag.

Han, T. A. & Pereira, L. M. (2013). 'State-of-the-art of intention recognition and its use in decision making'. *AI Communication*, 26, 237–246.

Han, T. A., Pereira, L. M., & Santos, F. C. (2011). 'Intention recognition promotes the emergence of co-operation'. *Adaptive Behavior*, 19: 264–279.

Han, T. A., Pereira, L. M., & Santos, F. C. (2012a). 'Corpus-based intention recognition in cooperation dilemmas'. *Artificial Life Journal*, 18(4): 365–383.

Han, T. A., Pereira, L. M., & Santos, F. C. (2012b). 'Intention recognition, commitment, and the evolution of cooperation'. In *Proceedings of IEEE Congress on Evolutionary Computation* (pp. 1–8). Brisbane: IEEE Press.

Han, T. A., Pereira, L. M., & Santos, F. C. (2012c). 'The emergence of commitments and cooperation'. In *Proceedings of the Eleventh International Conference on Autonomous Agents and Multiagent Systems* (pp. 559–566). International Foundation for Autonomous Agents and Multiagent Systems.

Han, T. A., Pereira, L. M., and Lenaerts, T. (2014). 'Avoiding or restricting defectors in public goods games?' *Journal of the Royal Society Interface*, 12(103). doi: 10.1098/rsif.2014.1203.

Han, T. A., Pereira, L. M., Santos, F. C., and Lenaerts, T. (2013a). 'Good agreements make good friends'. *Scientific Reports,* 3 (2695). doi:10.1038/srep02695.

Han, T. A., Pereira, L. M., Santos, F. C., and Lenearts, T. (2013b). 'Why is it so hard to say sorry: The evolution of apology with commitments in the iterated prisoner's dilemma'. In *Proceedings of the 23nd International Joint Conference on Artificial Intelligence (IJCAI 2013)*. Palo Alto, CA: AAAI Press.

Han, T. A., Pereira, L. M., Santos, F. C., and Lenearts, T. (2015a). 'Emergence of cooperation via intention recognition, commitment, and apology: A research summary', *AI Communications*, doi:10.3233/AIC-150672, vol. 28.

Han, T. A., Santos, F. C., Lenaerts, T., and Pereira, L. M. (2015b). 'Synergy between intention recognition and commitments in cooperation dilemmas', *Nature Scientific Reports*, *Sci. Rep.* 5: 9312.

Han, T. A., Saptawijaya, A., & Pereira, L. M. (2012). 'Moral reasoning under uncertainty'. In N. Bjørner, & A. Voronkov (Eds.), *Proceedings of the Eighteenth International Conference on Logic for Programming Artificial Intelligence and Reasoning (LNCS)* (Vol. 7180, pp. 212–227). Berlin: Springer-Verlag.

Hodges, A. (1997). *Alan Turing: One of The Great Philosophers*. London: Phoenix.

Hofbauer, J. & Sigmund, K. (1998). *Evolutionary Games and Population Dynamics*. New York: Cambridge University Press.

Levin, J. (2010). Functionalism. In E. N. Zalta (Ed.), *The Stanford Encyclopaedia of Philosophy*. Retrieved from http://plato.stanford.edu/archives/sum2010/entries/functionalism/

Martínez-Vaquero, L. A. & Cuesta, J. A. (2013). Evolutionary stability and resistance to cheating in an indirect reciprocity model based on reputation. *Physical Review E*, 87: 052810.

Martínez-Vaquero, L. A. & Cuesta, J. A. (2014). 'Spreading of intolerance under economic stress: Results from a reputation-based model'. *Phys. Rev. E* 90, 022805.

Martínez-Vaquero, L. A., Han, T. A., Pereira, L. M., & Lenaerts, T. (2015). 'Apology and forgiveness evolve to resolve failures in cooperative agreements', *Nature Scientific Reports, Sci. Rep.* 5, 10639. doi:10.1038/srep10639

McCullough, M. E. (2008). *Beyond Revenge, The evolution of the Forgiveness Instinct*. San Francisco, CA: Jossey-Bass.

McCullough, M. E., Kurzban, R., & Tabak, B. A. (2011). 'Evolved mechanisms for revenge and forgiveness'. In P. R. Shaver & M. Mikulincer (Eds.), *Human Aggression and Violence: Causes, Manifestations, and Consequences.* Herzilya series on personality and social psychology (pp. 221–239). Washington, DC: American Psychological Association.

McCullough, M. E., Pedersen, E. J., Tabak, B. A., & Carter, E. C. (2014). 'Conciliatory gestures promote forgiveness and reduce anger in humans'. *Proceedings of the National Academy of Sciences of the United States of America*: 111, 11211–11216.

McDermott, D. (2001). *Mind and Mechanism*. Cambridge, MA: MIT Press.

Nesse, R. M. (2001). *Evolution and the Capacity for Commitment*. Russell Sage Foundation Series on Trust. New York: Russell Sage.

Nowak, M. A. (2006). 'Five rules for the evolution of cooperation'. *Science,* 314(5805): 1560–1563. doi: 10.1126/science.1133755.

Nowak, M. A. & Sigmund, K. (1992). 'Tit for tat in heterogeneous populations'. *Nature,* 355: 250–253.

Ohtsubo, Y. & Watanabe, E. (2009). 'Do sincere apologies need to be costly? Test of a costly signaling model of apology'. *Evolution and Human Behavior*, 30(2): 114–123.

Okamoto, K. & Matsumura, S. (2000). 'The evolution of punishment and apology: An iterated prisoner's dilemma model'. *Evolutionary Ecology* 14: 703–720.

Pereira, L. M. (2012). 'Turing is among us'. *Journal of Logic and Computation*, 22(6): 1257–1277.

Pereira, L. M. & Saptawijaya, A. (2011). 'Modelling morality with prospective logic'. In M. Anderson and S. L. Anderson (Eds.), *Machine Ethics* (pp. 398–421). New York: Cambridge University Press.

Pereira, L. M. & Saptawijaya, A. (2015a). 'Abduction and beyond in logic programming with application to morality'. In *IfCoLog Journal of Logics and Their Applications (Special Issue on 'Frontiers of Abduction'*. London: College Publications.

Pereira, L. M. & Saptawijaya, A. (2015b). 'Bridging two realms of machine ethics'. In J. White & R. Searl (Eds.), *Rethinking Machine Ethics in the Age of Ubiquitous Technology* (pp. 197–224), Hershey, PA: IGI Global.

Pope, A. (1711). *An Essay on Criticism, Part II*. London: W. Lewis.

Saptawijaya, A. & Pereira, L. M. (2015a). 'Logic programming applied to machine ethics'. In F. Pereira, P. Machado, E. Costa, and Cardoso, A. (Eds.), *Proceedings of the Seventeenth Portuguese International Conference on Artificial Intelligence.* LNCS vol. 9273. Berlin: Springer-Verlag.

Saptawijaya, A. & Pereira, L. M. (2015b). 'The potential of logic programming as a computational tool to model morality'. In R. Trappl (Ed.), *A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations*. Berlin: Springer-Verlag.

⋆Schneider, F. & Weber, R. A. (2013). 'Long-term commitment and cooperation'. Tech. Rep., Working Paper Series, University of Zurich, Department of Economics.

Sigmund, K. (2010). *The Calculus of Selfishness*. Princeton, NJ: Princeton University Press.

Smith, N. (2008). *I was Wrong: The Meanings of Apologies,* vol. 8. New York: Cambridge University Press.

Sterelny, K. (2012). *The Evolved Apprentice*. Cambridge, MA: MIT Press.

Takaku, S., Weiner, B. & Ohbuchi, K. (2001). A cross-cultural examination of the effects of apology and perspective taking on forgiveness. *Journal of Language and Social Psychology* 20: 144–166.

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

Turing, A. M. (1950), 'Computing machinery and intelligence'. *Mind*, 59: 433–460.

Tzeng, J.-Y. (2004). 'Toward a more civilized design: studying the effects of computers that apologize'. *International Journal of Human-Computer Studies,* 61(3): 319 – 345.

Utz, S., Matzat, U., & Snijders, C. (2009). 'On-line reputation systems: The effects of feedback comments and reactions on building and rebuilding trust in on-line auctions'. *International Journal of Electronic Commerce,* 13(3): 95–118.

Winikoff, M. (2007). 'Implementing commitment-based interactions'. In *Proceedings of the Sixth International Joint Conference on Autonomous Agents and Multiagent Systems* (pp. 868–875). New York: Association for Computing Machinery.

Wooldridge, M. & Jennings, N. R. (1999). 'The cooperative problem-solving process'. *Journal of Logic and Computation,* 9(4): 563–592.

6 Moral feelings from rocky fictional ground

Eileen John

Introduction

From a certain philosophical angle, all works of fiction can look like rocky ground for eliciting emotion. Fiction is the stuff of 'mere imagination' and as such can seem too metaphysically remote and practically inconsequential to get an emotional grip on us. From a casual survey of how people engage with fiction, however, those philosophical considerations do not seem to do a good job of predicting or explaining how we respond to it (Carroll and Gibson 2011, Currie and Ravenscroft 2003, Djikic *et al.* 2009, Feagin 1996, Matravers 1998, Oatley 2011, Plantinga and Smith 1999, Robinson 2005). We can be moved by works of fiction and often seem to count a work's emotional power as central to its value.

Additionally, many philosophers have argued that fiction has morally transformative potential, both positively, taking fiction to have potential to *improve* moral understanding (Carroll 2002, Gaut 2007, Gibson 2007, Nussbaum 1990, Schellekens 2007, Zamir 2007), and negatively, most famously in Plato's *Republic* Book X (see also Hamilton 2003, O'Neill 1986). Regardless, the power of fiction seems importantly tied to the way readers engage with it as emotional and moral beings.

I want to follow a certain path of questions in bringing literary fiction, emotion and morality together. Supposing that emotions involve factors we take to be relevant to our well-being, I will ask how emotional responses to fiction can do that. What concerns or interests can we have at stake as readers? My partial answer is that the way representational resources are used is central to the emotional force of fiction. I will raise several other questions, conscious that they might not be resolved. Does experience with fiction challenge the idea that a person's emotions track his or her concerns? How closely bound together are emotions and moral judgements? What emotional responses are morally significant when engaging with fiction?

Readers' emotions do not simply affirm, embody or coordinate with moral judgements. They can involve a critical interplay between emotion and moral understanding. I will illustrate some of these claims at the end of the chapter by discussing two novels that offer particularly 'rocky ground' for emotional

response: Saramago's *Blindness* (2005) and *Seeing* (2007) do not set out to absorb us in the experiences of people living plausible lives. To understand the power of such novels, we need to focus on their representational tactics and on how the emotional response in relation to those tactics can be a source for moral reflection, rather than constituting moral judgement.

Emotional response to fiction

Setting aside initially the question of moral significance, let me give a very broad sketch of issues surrounding emotion and fiction (Dadlez 1997, Davies 2007, Matravers 1998, and Neill 1993). The sensible starting point *seems* to be that fiction offers us a depiction of people (or beings crucially similar to people) to whom we respond emotionally. The people portrayed might be grieving, in danger or in the throes of unrequited love. The work puts us in a position to know about or to be witness to their situation, and we respond to them to some degree as we would in real life.

Typically there are no real people undergoing whatever the work of fiction depicts, and audiences are well aware of this. What accounts for emotional response under these conditions? One major explanatory approach pivots on the idea that we *make believe* that there are such people living in these circumstances (Currie 1990, Walton 1990). The richness and power of fiction is a matter of the reader's imaginative activity. Within my 'game of make-believe', I come to know about, and respond emotionally to life. In Kendall Walton's view, these affective states do not function as fully genuine emotions. This is in part because the beliefs that emotions seem to require are missing (such as a belief that someone is in danger when feeling fear). Walton points further to the failure of these states to have the motivational force that is characteristic of given emotions, arguing that they should thus be classified as 'quasi-emotions' (Walton 1990: 201–2).

A number of theorists argue that the 'quasi-emotion' classification does not do justice to affective responses to fiction and have resisted the claims that belief and motivating force are conditions for genuine emotion (Lamarque 1981, Carroll 1990, Dadlez 1997, Matravers 1998). Such views, sometimes called 'thought theories', tie emotional response to fiction to the content and experiential qualities of the thoughts we entertain in engaging with fiction: 'the thought theorist maintains that it is not only *beliefs* about the intentional object of an affective state that can cause genuine emotions, but also the very act of entertaining the thought of that object' (Davies 2007: 127). Hence, we can respond emotionally to fictional characters because they are in some sense constituted by emotionally moving content of thought.

In different ways, both of these approaches see the reader's psychological contact with a person as the basis for emotional response to fiction. Either we make believe that we are coming to know about a person, or we have thoughts whose content is structured and united, as in thinking about a person. The emotional or quasi-emotional responses are supposed to have the kind of significance that emotional responses to real people have, including specifically

moral significance (showing, for instance, that we have proper concern when others are harmed unjustly). However, though I agree that how we think and feel about real people is at work in some way, I do not think that the priority in understanding what we do with fiction is to show how it enacts or parallels the experience of psychological contact with people. The 'real people' model is unlikely to do a good job of showing why these responses are interesting and important. Either the responses will be relatively safely contained within a game of make-believe, or they will emerge from the sheer thought contents associated with people, where it is not clear enough what we have at stake in entertaining such thoughts.

Consider how emotion is commonly understood to function in relation to a person's interests. In Jesse Prinz's terms, 'concerns are organism-environment relations that bear on well-being', and 'emotions are defined by the concerns they have the function of detecting' (Prinz 2007: 85). Jenefer Robinson states as a generally shared view that

> what we pay attention to in emotion are precisely those aspects of the world that we see as important to our own interests, wants, and goals. […] emotions seem to be ways in which an organism appraises the environment as satisfying or failing to satisfy its wants and interests.
>
> (Robinson 2005: 26)

Martha Nussbaum, though working with a different overall theory of emotion, makes this point as well: 'emotions are forms of evaluative judgment that ascribe to certain things and persons outside a person's own control great importance for the person's own flourishing' (Nussbaum 2001: 22). What interests are at stake for the reader who responds emotionally to fiction? It does not seem that my well-being could be affected by how well or badly things go for fictional characters, conceiving of them as people living their lives. The fact that life works out well for Elizabeth Bennet every time one reads *Pride and Prejudice*, or badly for Emma Bovary in Flaubert's novel, seems to have no bearing on the reader's well-being.

While I will suggest that we have interests at stake in fiction, I also want to note that the assumption that people track their own interests as readers is too simple. Emotional responsiveness is extremely flexible and does not seem driven solely by the demand to track and protect the emoter's own interests. Given the range and fluidity of what people feel when engaging with fiction, we might say that emotions are indeed acute detectors of interests and concerns, just not only our own. Perhaps we are built to respond emotionally when we become aware that *someone's* well-being is or even could be at stake.

Emotions and values

The 'thought theorist' can point out that I have so far construed the thoughts relevant to fictional characters too narrowly. Fiction involves us in thoughts not

only about what characters do and undergo, but about how to evaluate their lives, in moral and other terms, and these aspects of thought content can coincide with or diverge from a reader's actual evaluative commitments. My emotional responses can track whether or not such thought contents affirm the value commitments I have an interest in upholding. Of Shakespeare's Macbeth and the monstrousness of his behaviour, Jenefer Robinson says, 'we feel ourselves personally involved in this view of Macbeth, because of our own concerns, interests, and values: we are human; we fear and dislike a man whose actions show such contempt for our fundamental human values' (Robinson 2005: 112). It seems right that Shakespeare's play engages us with fundamental human values, but I would give a different account of the nature of such engagement and the feelings it involves. Do we really fear Macbeth the man? I find it more apt to say that the play offers a representation of a man, and we react to the implicit claim of such a representation, that it reflects an important understanding of humanity. This is perhaps not a deep disagreement with the thought theorist, but I will emphasize that we respond to the real representational 'labour' embodied in the play and the kind of danger or problem the representation poses for us, rather than sheerly to the abhorrent content of the 'man-thoughts' we entertain. Tzachi Zamir's discussion of *Macbeth*, for instance, presents the play as posing a live question about nihilism (Zamir 2007: 92–111). It seems that the grip of the play involves us having moral judgements in view (e.g., thou shalt not kill) and powerful emotional responses driven by a need to grasp the basis for those moral judgements. For Zamir, an emotional and morally dissolving abyss opens up because a real philosophical challenge comes alive for us in responding to the play. My emotions are not directly a rejection of Macbeth but rather show my own insecurity as to the basis on which I reject the nihilism that he is designed to embody.

Broadly, then, I do not think that the thought contents in themselves will account well for our emotional responses (more on this below). At this point, I want to make two brief further comments about the relation between fictional thought contents, emotions and values. It seems crucial to change of values that we be able to experiment with and contemplate different forms of value (Prinz 2007: 100). Fiction could, it seems, provide a safe space for engaging with evaluative thought contents we do not already embrace, and it would be important for such experimental thinking that our emotions not always serve to affirm our value commitments. If fiction is to help us examine and change our values, it would be helpful if emotions did not work only in a defensive mode, detecting 'threats' to value commitments and bolstering our resistance to them.

The second point is that, in any case, it is not at all clear that emotional responses to fiction are reliable detectors of thought-borne threats to a reader's interests. If a novel puts a protagonist at the centre of attention, it comes pretty easily to us to have emotional or quasi-emotional responses that make sense in terms of that protagonist's perspective, independently of how we would ourselves evaluate such events and actions. For instance, my real feminist commitments stand in an awkward relation to many emotional responses I have

to novels and films. I can be gleeful at a brutal revenge plot, caught up in the worries of a repressive social system I reject, or bored by the morally decent worldview driving a character, et cetera.

I appreciate that the phenomena of emotional response to fiction may seem too peripheral, too derivative from the core functions of emotion, to be able to put much pressure on how we conceive of emotion and values. We might say, 'Yes, we seem to go emotionally and evaluatively "off-piste" with fiction, but that is not deeply illuminating about the real relations of emotion and value.' This way of side-lining the phenomena would, I think, ignore an important source of philosophical questioning about the relations between emotion, concern and value judgement. I will suggest some questions that seem worth pursuing. One might see our emotional susceptibility, so evident in relation to fiction, as a sign of an influential, contentious and dynamic relation between emotion and value judgement. Emotions are sophisticated ways of registering and prioritizing information that is needed for value judgement (information about who cares about what, from what vantage-point, why they care, how different vantage-points relate). Value judgements, and moral judgements quite specifically, are about what should be cared about, while the emotions will primarily tell us whether anyone (real or fictional) does care. The work of fiction seems able to give us a reflective space in which we can realize that moral judgement and emotion can be in tension with, or differently focused from, each other. Emotional response can reveal things ignored by moral understanding, and moral understanding can reveal the limitations of emotional response. That seems to me a relatively ordinary possibility of awareness in response to fiction. For example, I can feel repelled by the character Casaubon in *Middlemarch*, and find that my response to him is morally ungenerous. The complexity of works like Sophocles' *Antigone* or *Oedipus Rex* seems due, at least in part, to the interplay and tension between emotional and moral response, as our compassionate feelings in response to characters' predicaments need not coincide with our moral judgement of their actions. The emotions we feel in responding to a work of fiction seem able to pose questions about how and whether concerns and value judgements align.

This picture leads me further to suggest, in a very schematic way, that experience with fiction is challenging with respect to sentimentalist accounts of moral meaning and judgement.[1] Morally significant emotional response and moral judgement are crucially intertwined elements in engagement with the vast majority of works of fiction. But the reason those elements are so central to engrossing us does not obviously seem to be because the emotions are constitutive of moral judgement. Rather, a novel, play or film seems to engross us by taking advantage of the possibilities for tension and mutual pressure between emotion and moral understanding. How this kind of experience with fiction bears on philosophical theorizing about emotion and morality is not straightforward. It is possible that we do things with fiction that are in fact philosophically 'fantastic' in some way (such as separating moral judgement from emotion in a way that is not deeply defensible). At any rate, while fiction might seem to be a domain in

which emotion plainly serves as the key to being morally 'anchored', I think the phenomena are not immediately easy to understand in those terms.

It is also worth noting that the emotions referred to as *the* moral emotions, especially those carrying moral disapproval such as guilt, anger, resentment and contempt, are not so prominent when responding to fiction. Even when we think a character, such as Macbeth, is a concentrated site of moral wrongness, it just does not seem that we spend much, if any, emotional energy being angry at or contemptuous of him. Moral understanding seems to capture what readers are often working toward in fiction better than emotionally approving or disapproving judgement. A view such as Jonathan Haidt's, for instance, that takes moral reasoning typically to follow after and to reinforce emotional reactions for socially strategic purposes (see, e.g., Haidt 2012: 25, 55), seems unsuited to doing justice to the kind of interplay between emotion and moral judgement, and to the freedom from some socially strategic demands, that seem available to us in fiction.[2]

Responding to representation

I have surveyed above what I consider to be some difficulties in grounding our emotional responses to fiction either in concern for the well-being of the people we imagine or in the thought of value commitments that are evoked by imagining those people. My alternative account of what we have at stake, and what is capable of directly engaging our emotions, rests on the point that in fiction we encounter the products of representational activity. Let me grant that this answer can sound unhelpful, as if it just says that we respond emotionally to what is represented. But I want to dwell on this idea that we respond to the productive activity of representation, taking a novel, for instance, to be a record of labour, choice and achievement. Readers encounter the results of someone selecting means for conceiving of and articulating aspects of life. With the help of the guiding conventions of literature, we can further assume that a value judgement is being communicated. This selective practice is being offered as good or adequate for some purpose. We respond to fictional characters at least in part as the products of representational activity, as things that manifest choices and possibilities for identifying and presenting what is worth noticing and understanding about human life. Our ability to summon them up as 'real people' in our engagement with a work then needs to be framed as something that the work enables or discourages through its strategies of representation.

This idea was well stated by Michael Weston early on in contemporary discussion of these issues. On Weston's view,

> talk about seeing characters as "real" people […] is directed at the quality of the realization of the fictional character, and points, therefore, towards the mode of representation employed in the work – that is, towards the *kind* of representation it is.
>
> (Weston 1975: 87)

For Weston, 'to be moved by Mercutio's death is to respond in the light of one's interpretation […] and hence is part of one's response to the sense we see in the play as a whole' (ibid., 86). Similarly, Flint Schier argues that 'our reaction to fictional characters is […] a reaction to them as represented in the text', and in responding or not responding to characters, we interact 'with the controlling intelligence of the artist' (Schier 1983: 85).[3] The general point is that we encounter fictional characters as aspects of a sophisticated artefact, a work that manifests an artist's intelligence and calls out for us to interpret characters as functioning parts of that artistic whole.

The readers' concerns are embodied in fiction because of our own goals and needs in relation to representation. We carry out and are the subjects of representational activity on a daily basis, and it matters a great deal to us whether we represent our lives and world adequately, and whether we and the world are represented adequately by others. The resources that are available to us as representers, and that are taken seriously within our communicative social exchanges, make a huge difference to whether we can identify things of importance to our lives and can make them intelligible to others. When I ask a teenager what he did at school that day and he replies, 'Stuff', he is taking control of the representational resources and refusing to make intelligible to me what I hope to understand about his day. Meanwhile, literary authors seize the available resources very differently, with ambitions for articulating and changing what can be noticed and found important. As ordinary representing agents, we regularly fail to be ambitious and innovative, and even when we try, we usually will not feel that we have done justice to what was there to be experienced, understood, and expressed. There is a great deal more to be said about how people can suffer, or have their needs and interests furthered, through human failures, achievements, and changes in representational activity. In the wealth of representational activity on display in literary fiction, we encounter something of concern to us, in a relatively straightforward way. Nussbaum says, 'To "put" things is to do an assessible action' and, citing Henry James, 'our whole conduct is some form of artistic "putting"' (Nussbaum 1990: 163).[4]

A story is a record of someone having taken responsibility for summoning up a living context, and the question of whether that responsibility is well used is in question for the reader. How do the choices made limit, focus, and open up possibilities for noticing, thinking, and feeling? The reader may address these questions more or less attentively, but I think that our implicit sense of how well a representational project is being carried out, to what purpose, will show up in our emotional responses. Our emotions detect and appraise how a work meets serious representational needs and goals. The fact that the context is only imagined means that the purpose of presenting facts and describing actual states of affairs accurately is not the guiding criterion. Rather, we appraise whether the work lets information, perception, emphasis, relation, mood, pace, breadth, depth and various gestalt properties have a satisfying presence and role. The fact that we bypass the goal of telling the truth about

the actual world, and there is no single 'replacement purpose', enhances the cognitive potential of fiction, in the sense that it makes us consider a fuller range of representational goals. Instead of taking for granted a single criterion for success, the reader can try out different criteria. One might be whether a work gets the individual lives of people to 'come to life' in thought and imagination for the reader, but, as I shall now discuss when analysing Saramago, a novel can help us to consider how and whether we benefit from such a 'realist' criterion for representation.

Responding to Saramago's *Blindness* and *Seeing*

Blindness (*Ensaio sobre a cegueira*) and *Seeing* (*Ensaio sobre a lucidez*) use quite dramatic representational techniques that do not let 'contact with real people' be the central imaginative project for the reader. The characters are for the most part nameless, some picked out with repeated descriptive tags ('the old man with the black eyepatch') or job titles ('the inspector'), and the setting is a city, also unnamed. Each novel has as its premise that something odd and never explained happens: in *Blindness*, all the citizens of the city except one ('the doctor's wife') go blind, and in *Seeing*, set in the same city four years later, the vast majority of the citizens leave their ballots blank at an election. These implausible occurrences lead to extreme consequences, partly due to the panic and brutality of the governmental response to these situations, but also, in *Blindness*, due to the terrible degradation of the people who suddenly cannot find their way around in the world.

If readers wanted to sink into the portrayal of these events, to 'lose themselves in the fictional world', this immersion would be difficult to do. The structure and punctuation of the narration and dialogue make it challenging to follow the ordinary details of who is doing and saying what. Paragraphs can go on for pages, and the speech of different characters is run together so that one easily loses track of who is speaking or if it is speech at all. The narration is also relatively intrusive and shifts abruptly, so that how the story is being told is itself a matter of explicit discussion. For example, in *Blindness*:

> From this point onwards, apart from a few inevitable comments, the story of the old man with the black eyepatch will no longer be followed to the letter, being replaced by a reorganised version of his discourse, re-evaluated in the light of a correct and more appropriate vocabulary.
>
> (Saramago 2005: 115)

This tone of prickly officiousness and the confusing promise of a correction that does *not* follow a character's story to the letter, of course carry no conviction for the reader. These awkward narratorial interventions are themselves part of what signals to us that this is a work that does not offer us a comfortable way of 'working'. In another intrusive passage in *Seeing*, the narrator dwells on a rather banal question uttered by one of the nameless officials:

> The question, as well as being superfluous, was, how can we put it, just the teensiest bit dishonest […] because it was obvious that the person asking the question was taking advantage of the authority inherent in his position to shirk his duty, since it was up to him, in voice and person, to initiate any exchange of information. If we bear in mind the sigh he uttered and the rather querulous tone we thought we detected at one point […].
>
> (Saramago 2007: 9)

This passage mixes apparently confident moral criticism and analysis of implicit expectations with some coyness about whether the narratorial 'we' really is authoritative, e.g. about such details as querulousness of tone. Despite the somewhat laborious attention to the official's behaviour, this character turns out to have no further role in the novel. It is not easy to develop a sense of what is important in what is offered to the reader's attention. There are very few passages in which one is tempted to think, 'Ah, this is a perspective with some wisdom to offer on these strange circumstances.'

These novels are dense with conversation and commentary, but the cumulative effect is not that readers are able to build up a densely imagined and emotionally attentive experience of the lives of the characters. Compare some of the ways in which the case is made for the moral significance of specifically 'realist' works of literary fiction. Philosopher Berys Gaut's argument for the morally illuminating powers of art highlights the 'test of imaginative acquaintance': by asking the reader 'imaginatively to adopt the target's position (asking her to imagine what it is like to be the target), [an artwork] enhances the power and precision of her feelings towards the target'. Our imaginative access to the character 'focuses the power of moral judgement' (Gaut 2007: 160). Nussbaum emphasizes the achievement of a kind of epistemic and felt intimacy with the characters: 'We actively care for their particularity, and we strain to be people on whom none of their subtleties are lost, in intellect and feeling' (Nussbaum 1990: 162).

Novels such as *Blindness* and *Seeing* persistently fail to support such experiences of imaginative acquaintance and fine-grained, caring attention. They nonetheless offer possibilities for emotional and moral engagement. My claim about our real concern for how people use representational resources is intended to help make this case. Saramago uses the representation of fictional people and events to give us a disconcerting, disorienting, not very richly imagined experience. The tactics of the novels are, in a certain way, very aggressive: readers are not allowed to feel they are making comfortable progress in the understanding of a fictional world, but are rather forced to halt over the tactics. Why is this story being told in this way? Why is it not allowing the perspectives of people undergoing the depicted events to occupy the centre of attention? Even the visual appearance of the pages will raise questions: why are the paragraphs (if they can be called paragraphs) so overwhelmingly long and unaccommodating to desires for structure, focus, and intelligible order?

Now, it might seem that these sorts of questions will leave us only able to think of such novels as intellectual puzzles. If we halt over such tactical

questions about the representational practice on offer, where is the scope for emotion or for emotionally charged moral engagement? I do not think there is a single or simple answer as to why these novels are *not* (or not only) intellectual puzzles. Let me suggest a few reasons why this 'halting engagement' seems to be emotionally and morally fruitful. First, the extremity of the novels' tactics can produce a sense that there is something urgent behind this practice. It seems obvious that something is 'not OK', and the novels show drastic steps being taken to awaken us to whatever that is. Although the novels are often also quite funny (perhaps slightly on display in the high-handed, somehow ridiculous passages quoted above), the tactics themselves convey what I would call a mood of desperation or dread. 'Do not be complacent!' is somehow implicit in the novels' failure to give us a substantial foothold for imagining and following a story. The novel's qualities as a representation seem able to evoke a felt awareness that we may fail to identify what is wrong with a social and political world. The mood of desperation or dread can lead us to feel a concern for, and the absence of, confident moral judgement.

The implicit evaluative claim seems to be that there are conditions of life for which these are the adequate or needed representational practices, and the reader has to figure out how that claim might be defended. We probably rarely do this for a novel in more than a provisional, exploratory way, but the idea is that there is always a basic question for readers as to why one would portray life in this way. The reader can then develop a sense, or again, more rarely, an articulated judgement, as to whether and how the answer to this 'why question' bears on the reader's concerns.

To do this for either of Saramago's novels would be a long story, but here is a quick sketch of what might be involved. In the case of *Seeing*, one might say that the relevant conditions are ones in which, while there may be institutions, such as voting, that officially acknowledge people's abilities to reason, critique and change the conditions under which they live, those institutions do not actually expect people to exercise those abilities and do not endow them with power to change their society. The novel proceeds to take those conditions as needing to be represented in a way that foregrounds a kind of impoverished and stilted awareness of human interaction and autonomy, and a frustrating loss of ability to take anyone in particular to be of much interest. Whether this is a 'need' is of course difficult to assess. But the claim of the novel seems to be that the urgency of the problem, of conditions supposedly realizing individual autonomy and collective responsibility and failing to do either, means that our attention needs to be directed negatively to what would be counted as worth knowing and saying about people in such conditions. To some degree, the novel's perspective enforces the loss and impoverishment within those conditions.

These are not the only things the representational practice of *Seeing* offers, and the variations are, I would say, quite important to the emotional impact of the novel. Let me give one example (of course I cannot assume that my example will chime with other readers' experience). So far I have emphasized that *Seeing* saddles readers with a kind of distanced relation to, or almost a

nostalgic awareness of, what it would be to care about and have a clear moral relation to a person. A passage toward the end of the novel explicitly concerns this aspiration. It is a conversation between three government officials who have gradually become willing to betray their covert mission (which moved from utter pointlessness to paranoid viciousness). Amongst the three, the leader has reached this willingness first:

> Because I was afraid, Afraid of what, we're not monsters, Afraid that the need to find a guilty party at all costs would stop you seeing the person who was there before you, Did you trust us so little, sir, It wasn't a question of trust, of whether I did or didn't trust you, it was more as if I had found a treasure and wanted to keep it all to myself, no, that's not it, it wasn't a question of feelings, that wasn't what I was thinking, I simply feared for that woman's safety, I thought that the fewer people who questioned her, the safer she would be, So put in plain and simple language, and forgive my boldness, sir, said the sergeant, you didn't trust us, No, you're right, I admit it, I didn't, Well, don't bother asking our forgiveness, said the inspector, you're forgiven already […].
>
> (Saramago 2007: 263–4)

In lifting this passage out of its context towards the end of the novel, it is hard to convey how exciting it is. The dialogue between these three characters has to that point been almost unbearably empty and cautious, only vaguely hinting at what one would think they desperately need to talk about. The passage here shows them suddenly trying to understand and report on their own motives, and noting a simple, morally relevant emotion, fear for another's safety. They admit their own failures of trust and trustworthiness, which are met with immediate forgiveness. Although this scene is not a deep opening up of human experience and personality, it can be a tremendous relief for the reader who has been deprived of dialogue depicting people with capacities to think independently and to make judgements about human needs and mutual responsibilities. The experience of relief, perhaps even gratitude, and more broadly the rush of pleasure in the way the novel briefly hints at how people might reflect on and express themselves to each other can, I think, be taken by the reader as a genuine emotion that signals something about his or her real concerns. I have something at stake in whether people are conceived and represented as having these self-critical and mutually responsive capacities.

What I have suggested above about the complex relation between concerns and values means that I would not immediately take this kind of emotional affirmation to constitute a moral judgement. I think the relief, gratitude and basic pleasure in the passage are informative to me. They are crucial elements to incorporate into my sense of how the novel addresses my concerns. Integrating them into a moral framework depends on developing moral understanding. In general, this understanding requires moving from what I have found I care

about (e.g., in this small example, portraying people in this way rather than as executors of official roles) to having a sense of why this is worth caring about.

In closing, let me acknowledge that these remarks do not amount to an adequate account of this complex territory. This discussion has primarily raised questions about what we can learn about emotion and morality from engagement with fiction. I hope also to have made a schematic case for understanding at least some important emotional responses to fiction as responses to the work of representation that we directly encounter through this process. Fiction elicits emotional responses that engage with deep and expansive human concerns, and these responses can contribute to reflective questioning and understanding of moral value.

Notes

1 Statements from differing sentimentalist positions include: 'Sentimentalists claim that the emotions do not just detect values but partly serve to constitute them – as the funny is not just detected by our amusement but shaped by the human sense of humour' (D'Arms and Jacobson 2014: 266); 'An action has the property of being morally wrong (right) just in case there is an observer who has a sentiment of disapprobation (approbation) toward it' and 'moral judgments simultaneously express how we feel and represent things' (Prinz 2007: 92, 100).

2 Haidt notes that moral intuitions 'can be shaped by reasoning, especially when reasons are embedded in a friendly conversation or an emotionally compelling novel' (Haidt 2012: 71). I am trying to suggest here that the emotional power of a novel can also be examined and critiqued by moral understanding, so that emotions are not inevitably 'driving' the reasoning.

3 See Matravers (1998), especially pp. 85–88, for another discussion that emphasizes engagement with a representation, but with the 'real people' model playing more of a role. Robinson (2005) is one of the most helpful discussions, in highlighting the role of a reader's interests in supporting emotional response.

4 Nussbaum puts this even more strongly: 'novels do not function […] as pieces of 'raw' life: they are a close and careful interpretative description. All living is interpreting; all action requires seeing the world *as* something. So in this sense no life is 'raw' and […] throughout our living we are, in a sense, makers of fictions' (Nussbaum 1990: 47).

References

Carroll, Noël. (1990). *The Philosophy of Horror*. New York: Routledge, Chapman and Hall.

Carroll, Noël. (2002). 'The Wheel of Virtue: Art, Literature, and Moral Knowledge'. *Journal of Aesthetics and Art Criticism* 60 (1), 3–26.

Carroll, Noël and John Gibson (eds.) (2011). *Narrative, Emotion, and Insight*. University Park, Pennsylvania State UP.

Currie, Gregory (1990). *The Nature of Fiction*. Cambridge: Cambridge UP.

Currie, Gregory and Ian Ravenscroft (2003). *Recreative Minds: Imagination in Philosophy and Psychology*. Oxford: Oxford UP.

Dadlez, E. M. (1997). *What's Hecuba to Him?* University Park: Pennsylvania State UP.

D'Arms, Justin and Daniel Jacobson (2014). 'Sentimentalism and Scientism'. In Justin D'Arms and Daniel Jacobson (eds.), *Moral Psychology and Human Agency*. Oxford: Oxford UP, 253–78.

Davies, David (2007) *Aesthetics and Literature*. London: Continuum.
Djikic, Maja, Keith Oatley, Sarah Zoeterman and Jordan Peterson (2009). 'On Being Moved by Art: How Reading Fiction Transforms the Self'. *Creativity Research Journal* 21, 24–29.
Feagin, Susan (1996). *Reading with Feeling*. Ithaca: Cornell UP.
Gaut, Berys (2007). *Art, Emotion and Ethics*. Oxford: Oxford UP.
Gibson, John (2007). *Fiction and the Weave of Life*, Oxford, Oxford UP.
Haidt, Jonathan (2012). *The Righteous Mind*. London: Allen Lane.
Hamilton, Christopher (2003). 'Art and Moral Education'. In José Luis Bermúdez and Sebastian Gardner (eds.), *Art and Morality*. New York: Routledge, 37–55.
Lamarque, Peter (1981). 'How Can We Fear and Pity Fictions?' *British Journal of Aesthetics* 21, 291–304.
Matravers, Derek (1998). *Art and Emotion*. Oxford: Oxford UP.
Neill, Alex (1993). 'Fiction and the Emotions'. *American Philosophical Quarterly* 30, 1–11.
Nussbaum, Martha (1990). *Love's Knowledge*. New York: Oxford UP.
Nussbaum, Martha (2001), *Upheavals of Thought*. Cambridge: Cambridge UP.
Oatley, Keith (2011). *Such Stuff as Dreams: The Psychology of Fiction*, Oxford: Wiley–Blackwell.
O'Neill, Onora (1986). 'The Power of Example'. *Philosophy* 61, 5–29.
Plantinga, Carl and Murray Smith, eds. (1999). *Passionate Views: Film, Cognition, and Emotion.* Baltimore: Johns Hopkins UP.
Prinz, Jesse (2007). *The Emotional Construction of Morals*. Oxford: Oxford UP.
Robinson, Jenefer (2005). *Deeper than Reason*. Oxford: Oxford UP.
Saramago, José (2005). *Blindness*, trans. Giovanni Pontiero. London: Vintage.
Saramago, José (2007). *Seeing*, trans. Margaret Jull Costa. London: Vintage.
Schellekens, Elisabeth (2007). *Aesthetics and Morality*. London: Continuum.
Schier, Flint (1983). 'Tragedy and the Community of Sentiment'. In Peter Lamarque (ed.), *Philosophy and Fiction*. Aberdeen: Aberdeen UP, 73–92.
Walton, Kendall (1990). *Mimesis as Make-Believe*. Cambridge, MA: Harvard UP.
Weston, Michael (1975). 'How Can We Be Moved by the Fate of Anna Karenina? (II)'. *Proceedings of the Aristotelian Society* 49, Suppl. vol. 6, 81–93.
Zamir, Tzachi (2007). *Double Vision*. Princeton, NJ: Princeton UP.

7 Emotional rescue and, *au ralenti*, some stories about images

Carlos Augusto Ribeiro

> Nothing is as seductive and as cursed as a secret.
>
> (Søren Kierkegaard)

Introduction

Plastic arts and image have their origin in love, according to Pliny's myth: the shadow gives the similarity needed to the affective and mnesic role of representation. The lover draws the outline of the shadow of her lover's face as he is about to leave, so that the image is the proof and the symbol of their loving bond and, while he is away, the remedy against separation. In 'The Oval Portrait' (1842), by Edgar Allan Poe, a painter strives to achieve maximum similarity in the portrait of his wife. The woman dies when the portrait is finished, and the disastrous consequences affect all those involved.

'The Oval Portrait' and *The Invention of Morel* (1940), by Adolfo Bioy Casares, attest to the continuity of the relationship between image and love. A sort of distorted relationship is present in some of the characters, either by a tendency towards excess, or by a sadist touch. In both stories, a similar principle prevails: the life received is not only paid with the inexorable death but the life in the image also demands a previous yielding of life itself. The compensation from such obligation is the immortality given by a substitute.

The chapter is divided into three parts. In the first, part I propose a close reading of 'The Oval Portrait', and will do the same with *The Invention of Morel* in the second part. As a reader, immersed in the fictional reality, I assumed an imaginative attitude towards the characters' situations, perceiving and feeling them as if from the inside, as if facing them in reality. At the same time, I am aware, as any reader is, that I will come upon the scene too late, with no power to intervene. The third and final part emphasizes the common threads between the two texts regarding emotions and morality. Both stories attest to the power of artistic and technological images over the characters' emotions and behaviours, showing the inseparability between the cognitive and emotional processing. In the portrayed fictional worlds, people surrender themselves entirely to an arbitrary power of a brilliant genius (male artist or inventor), seemingly united by the same desire of immortality.

'The Oval Portrait'

The narrator of 'The Oval Portrait' spends the night in a chateau which had been recently abandoned.[1] Wounded and weakened, he does not dissuade his valet from his intention to force entry into this strange, grand and melancholic chateau. They establish themselves in one 'of the smallest and least sumptuously furnished apartments'[2] located in the tower which is full of various antiques (tapestries and armorial trophies), and paintings (OP: 127). Before falling asleep, the narrator contemplates the pictures and reads a book he had found on the pillow. Many hours later, as he shifts the candelabrum to throw more light on the book, an oval picture that had been hidden in the dark corner of the room is suddenly revealed. Shaken from a 'melancholic stupor' or 'semi-dreamy' state, the narrator is deeply touched by this portrait, which frightens, confounds and subdues him (OP: 127–128). He tries to understand the reason behind the 'spell of the picture', and to reach the calmness of spirit necessary for a 'more sober and more certain gaze' (OP: 128–129). Although he is not certain that the execution and beauty of the countenance are the most plausible reasons for taking the painted head for that of a living person, he firmly rejects that it is due to his fancy (obscured by stupor). He claims that the presence of the painting and the observation of its qualities are enough to dissipate the possibility of a temporary hallucination. After careful observation of the painting, he finds that its spell resides in the 'absolute life-likeliness of expression' and lets himself fall in bed (OP: 129). With 'deep and reverent awe', he puts the candelabrum back in its former position, takes the small book (a catalogue which discussed the story of each painting), and looks for the number which designated the oval portrait (ibid.). In the catalogue, he finds out the fatidic story of the painting, and the description of its execution, which confirms the feelings that haunt him. The nocturnal experience of the narrator echoes the experience of the painter. Twice, the painting provokes both spell and horror at the same time: first to the painter and then to the intrusive observer. They both reproduce the same exclusive gaze that is intrinsic to the dynamics of fascination. The production of this peculiar image – a portrait that mortally touches its model – is determined by unusual conditions. Passionate, 'wild and moody', lost in reveries, the painter in Poe's tale works on the portrait, night and day, as if possessed (ibid.). Subtly, his deep love and attention are entirely transferred from his lover to the production of the painting – the creation of a miraculously perfect resemblance. He rarely takes his eyes from the canvas. He does not want to see that the tints spread on the canvas were taken from the cheeks of his wife.

Jealous of the time her lover devotes to Art (the rival that disputes his attention), the young model 'of rarest beauty' learnt how to tame her body to pose for unlimited time, in spite of her initial horror to the painter's proposition (ibid.). Patient and kind, 'humble and obedient', she smiles through pain until the portrait is finished (ibid.). This may be due to her love for the painter, as she notices in his face 'a fervid and burning pleasure in his task' (ibid.). Or it may be due to vanity, as the painter 'had high renown' (ibid.). Whatever the reason,

she does not dare to interrupt him and accepts the isolation from the world, immersing herself between two spaces (the atelier and the canvas), and the two objects of her affection (the painter and the portrait). Weakened but not guessing her fate, she gives her life to inhabit the image (a new body) which takes her as a model. She gains another flesh and is transfigured into an immortal gaze.

The witnesses who visit the place choose not to disturb the course of the event. Not even the evident growing transfiguration of the model into a live spectrum and finally into a corpse incites them to empathetic emotion or compassion. On the contrary, all those who contemplate the execution of the painting speak in a low voice about the perfect resemblance, acknowledging and religiously appreciating its 'miraculous' pictorial qualities (ibid.). They reveal great fascination for the power of the painter (painting miraculously) and less for the deep love uniting the two lovers (painter and wife), even if it means the sacrificial death of the model. When the portrait is about to be finished – 'save one brush upon the mouth and one tint upon the eye' – the witnesses decide to prohibit more visitors to the tower as the painter had grown wild due to the ardour of the work (OP: 129–130).

The last brush on the canvas, which took hours, days and consecutive weeks, coincides with the last breath of his beloved. Almost simultaneously, the painter steps back to admire the painting and so, to admire himself: 'the painter stood entranced before the work which he had wrought' (OP: 130). Pale and aghast, he cries 'with a loud voice' that his painting is 'indeed Life itself!' (ibid.). When he suddenly turns to check on his lover, he finds her dead. The coincidence generates illusion (that of a single cause for what happens), hallucination and perplexity. There seems to be a disturbing contagious law, acting over our heads, which makes us believe in a system of transfusion between referents and representations, objects and images; and suspect of a perverse relationship between visible and invisible, material and immaterial, vision and blindness.

In this scene, the painting seems to be the cause which leads to the distribution of roles and destiny of each protagonist. The influence of the image closes them all in an emotional space where strong agitation alternates with immobility and amazement. The image, as an instrument of metamorphosis, turns human agents into objects and spectra, ultimately dehumanizing them.

Involuntarily, the painter is converted into the creator of a deadly image which, by means of doubling and transfusion, absorbs and vampirizes life. Overtaken by an idolatrous passion, he is unable to distinguish and balance two types of love: the love for a real person and the passion for the image of that person. The division between an image of life (open) and an image of death (encirclement), between life and its simulacrum, is implicit between the indifference or convenient ignorance of the painter and the submissive and resigned behaviour of the model. Both lovers ignore the fact that if on the one hand an image can unite them, on the other hand, mutual love cannot be restricted to a single and final image. Just as it happens with an artistic installation when the exhibition comes to the end, when the castle is ruined, it is time for the dispersal of the collections. The scenery of a catastrophe is kept in the portrait and in the catalogue.

The Invention of Morel

Compared with 'The Oval Portrait', *The Invention of Morel*, by Adolfo Bioy Casares, depicts the intensification and extension of the domain of the image by establishing a technological environment which replaces the natural environment of an island. Fulfilling the 'myth of total cinema' (Bazin, 1992: 23), Morel's invention aims at guaranteeing immortality by means of surgical extraction of doubles (simulacra) and, concomitantly, by discarding the bodies as mere carcasses or hollow shells.

The narrator is a fugitive of the law sentenced to death who hides in a phantasmagoric island, deserted except for a group of visual apparitions (white summer visitors in *démodé* costumes). The island is the focus of a strange disease which kills from the inside out in fifteen days, leaving the body without nails and hair, and drying the skin and the eyes. To make matters worse, the ecosystem of this insular private paradise is collapsing: sick trees, loss of regularity of high tides, and vegetation (plants, meadows, flowers of all seasons) which, although abundant and diverse, occur out of time and at unusual places.

In an attempt to organize his thoughts and memories, and evaluate what has already been written or thought, the narrator describes the days of his struggle for survival. Hunted by the law and harassed by natural threats such as heat, tides or floods, dispossessed and confined to the narrowest place in the island (marshes), he sets to describe the morphology of the island and its constructions which date from 1924 and include a museum, a sanatorium, a library, a pool and a chapel. He details the materials used, the countless rooms and their decoration and distribution, providing us with his aesthetic appreciation and literary and philosophical references.

Affected by various deprivations and intoxications, the narrator becomes almost paranoid and suffers infirmities and hallucinations, oscillating between cycles of fear and anxiety, followed by relative calm, marked by flashes of consciousness and feelings of helplessness. His diary, turned into a will, shows the fluctuations of his convictions and his instability and duplicity regarding what is real/unreal; what he says and does not say; the island and its strange people; the senses and consciousness. The world seems to have lost the coherence and boundaries between vigil and dream, between real and fictional world. Exiled from human conviviality, he feels a mere shadow among shadows, persecuted and terrified by fascinating apparitions, 'fleeting giants', which appear indifferent and calm, as alleged representatives of a civilized and policed world. Other times, he is dominated by a feeling of nostalgia and desire of reciprocity, approaching the apparitions. Eventually, he falls in love with one, Faustine, whose intimacy he desperately yearns. He engages with her as if she were not a ghost, and makes repeated and vain attempts to declare and prove his love. These always end with him begging and screaming at her indifference and silence, accusing her of a persistent feigned and simulated attitude. In one of his efforts, he builds her a garden, which in turn is part of a strategy for self-representation. Through it, the narrator declares himself as an 'invisible

photographer' and painter who carefully organized and staged a scene from his everyday life. A picture is taken showing two figures (in *silhouette*): the narrator himself as a short man kneeling before an enormous woman, Faustine, seated and looking at the sea and the sunset. Later on, when the narrator sees himself surrounded by the images recorded and projected by the sophisticated Morel's machine, he sees two distinct times overlapped: the past images of two apparitions, Faustine and Morel, walking side by side, juxtaposed with the images of what seems to be now his ruined garden. The narrator is constantly projecting his own sentiments onto the moving images he perceives, fancying (seeing-in) the apparitions as living subjects who still have a complex inner life. He accuses Faustine and Morel of being lovers in collusion against him, beyond the limits of evidence. The destruction of his garden reinforces the narrator's paranoid suspicions.

He identifies himself with the inventor, Morel, and yet regards him as a rival: a rival in love and a rival concerning the control over Morel's machine. In the end, it seems that the narrator succeeded not only in usurping the machine, its plans and operation, but also in mirroring the same death as his inventor, dying at the hands of the reproduction machine. This is the way he found to belong to their world.

The narrator and Morel seem to be reciprocal and rival doubles. At a certain moment, the narrator informs the reader that he has himself carried out a process of merging and appropriation of an alien text. So, and according to one editor's footnote, the narrator's diary joins a quoted text (Morel's typed document about his own invention) and a text without quotation marks referring to marginal notes, handwritten in pencil with the same handwriting of the rest of the diary. However, the reader only sees a unified text (even if truncated in some chapters, according to the narrator's claims), a text with and without quotation marks, reproduced in typographic characters. To complicate things, the narrator knows (and the reader does too) that some facts he himself provides are contradicted by his own annotations as editor. Whilst the narrator still denounces suppressions on the author's text made by the editor (the narrator himself), his true identity is confirmed by an editor's footnote to the reader claiming he had removed some parts of the author's text due to a lack of space. There is no evidence whatsoever to confirm that the text without quotation marks belongs entirely to either the narrator or Morel. From this perspective, narrator and Morel represent metaphors for editor and author, and the Morel's machine a metaphor for the machinery of a book.

Morel's invention consists of a machine to remedy spatial and temporal absence by means of sensory prosthetics (visual, auditory, olfactory, thermal and tactile). Powered by tidal energy, the machine works in three phases: 1) collecting life, 2) capturing it, and 3) archiving it. It is an invention that creates simulacra of people, identical to living ones, but lacking self-consciousness similar to characters from a film or novel. Each living being (both human and non-human) becomes the generator of partial or entire identical duplicates. The resulting images (emanations) surround the individuals and involve the whole

space, regardless of whether it is day or night, each time reconstituting them in different simulacra. Each reproduction of an original is another object; each reproduction of life, a living reproduction. Reproducing is (almost) equivalent to creating a soul for each image. Hence, there is a parallelism between the destiny of men and that of images. Life is what is latent in the recording or reproduction. However, when life is produced, its original source is also killed.

The process is constantly repeated, despite variations in times and circumstances, including the images, the situations, the events, the people, the behaviours, the conversations, the tides, and so on. Everything comes in duplicate and juxtaposed in time: the heavenly bodies, the sun, the seasons, the objects, the living beings, the everyday acts, the individual death, the fact and its memory, the destruction and reconstruction, the repetition in loop and its reassembling.

Ironically, in *The Invention of Morel*, white[3] people constitute a foreign minority and are regarded as objects of museological curiosity (as if they were less developed), exotic and hybrid (speaking 'perfect French as South Americans'). However, due to their technological power, they are also seen as emissaries of death wherever they arrive, annihilators of places and of any possibility of a happy and healthy life, either in the present or future.[4]

Common aspects between the two stories

There are several common points between the narrators of both stories, who share a similar physical and psychological condition: they are both wounded, impaired and terrorized by images. Both consider themselves intruders (castle, island), and face complex and disturbing experiences associated with events that took place in the past or connected to realities from the past but phantasmally incorporated in the present. Likewise, they are both affected by contradicting emotions. Besides surprise and satisfaction, fear and panic are the most dominant emotions.

Both stories attest to the power of images (static or moving, traditional or animated) in controlling the characters' emotions. Images are presented as instruments (technologies) of immortality, life-suction forces, enchantments of terror. They monopolize the attention and reasoning of the narrators, unleashing, in each of them, a continuous process of thoughts, emotions and feelings that are mutually interlinked and involved. Ruined or abandoned spaces are more than mere pretexts for extraordinary occurrences. They provide conditions for the production of special images whilst offering protection and preservation of those very same figures. These spaces shape and reveal the life inside characters, their thoughts, emotions and temperaments. The spatial and temporal disposition of elements mentioned by the narrators determines the focus of each situation and is crucial for the reported emotional experience, attesting to the inseparability between the cognitive and the emotive, in mental life.[5] Although belonging to different genres, both stories confront the reader with suffering, bad luck and sacrifice on behalf of a love not moderated by

images, with the adversities and growing degradation of human life (and, also, the degradation of an ecosystem) in the name of art or technology.

The narrator of 'The Oval Portrait' appears emotional towards the quality of life of the oval portrait – and never with the beauty of the figure represented in it. Like the witnesses who beheld the production of the oval portrait, he is not worried with the horrible process of transferring life, with the quality of the painting, or its making. What moves him the most is the extraordinary event of life being retained and conserved in the painting and not so much the collateral damage (martyrdom and death) inflicted to the woman who posed for it. Even less moving to him is the transformation of the model's patience into apathy.

The narrator of *The Invention of Morel* makes the reader oscillate between identification (compassion and sympathy for the evil that befell him) and a certain detachment. He wins the reader's compassion and sympathy in the instants in which, for example, he claims to be manipulated by the author and the editor during the writing of his diary. Even when he shifts the way he presents himself: as a spectator of the cinematographic characters (into which the people of the island were converted), as an actor (playing the role of one or more characters), or, even, as a cinematographic and literary montage operator (crossing the ontological fracture between the spectator/reader and the fictional characters). There is a certain degree of distancing deriving from a feeling of dismay towards the oppression he exerted (or appears to have exerted – regardless if he did it as an author or an accomplice) over the victims of the invention. Immortality is reached through a process of transforming people into spectra, subjecting their bodies to the painful experience of a slow death without having the opportunity to express their choice, and unaware of the consequences of the ongoing reproduction process. They end up encased in a closed world, parallel, of circular temporality. The island has remained (apparently) intact until the arrival of the fugitive/narrator. Once he reached this 'paradise', the intruder (considering himself to be the victim of an oppressive legal and prison system) is won over by a desire of being part of that immortalized past from which, at first, he felt excluded. An uncontrollable envy causes him to act against himself (pathetically subjecting himself to the transformation allowed by Morel's invention) and against the others (destroying the authenticity and coherence of an archived world by causing its temporal homogeneity to explode). Such a desire to embody the represented (film) reality is also the desire to challenge and compensate the technical possibility of on-site distancing towards scenarios or situations, by means of a new montage.

The view adopted by the narrators of both stories seems to favour the oppressive attitude demonstrated by the painter and the inventor, and the indifference of the witnesses who display a somewhat sadistic delight. In *The Invention of Morel*, the witnesses are neither totally imaginary nor totally real. According to Edmund Burke, their indifference translates into an absence of sympathy, seen as the ability of the subject to put him or herself in the place of another (1993: 52). The painter's dehumanizing oppression is the result of an absolute attachment to Art, which leads him to ignore the value of the

woman he loves, as if she was not present. The absolute attachment to Art – also shared by the witnesses – may hide an attraction for the macabre, an appetite for visions of decay and pain. In the case of *The Invention of Morel*, the attachment to technology leads to a similar attitude of complacency towards the injustices committed by so-called geniuses. When faced with the violation of the dignity and individuality of the victims, they show the same lack of tolerance that seems to approve their transformation into mere instruments of others' desires.

Evil is strongly present in both stories. According to Paul Ricoeur, evil is the ability to refuse the value of life (1998: 280). Jean-Paul Changeux describes it as the destructive impulse of the community, which opposes the individual's life instinct, the survival of his/herself and of others, the harmonious life in society, the common good or the joy of living, in short, against the 'ability to understand another, to represent one's mental states, to witness sympathy, friendship and perhaps love' (ibid.). The inhuman exploitation exerted by the painter over the model represents a show entirely consented and managed by the witnesses, the same witnesses who decided to close the tower (a private club of sorts) to curious eyes, allowing the show to go on in secret, up to the woman's last living moment. The desensitization process reached all participants (including the model). All engaged in an exercise of indomitable desire to explore the limits between life and death, as if what was happening before their eyes could not affect them emotionally, or as if they thought that what happened was not really happening. Such observation is morally disturbing.

To the reader, who imaginarily puts him or herself in the place of the fictional characters, there is not an insurmountable distance between feeling oppressed or moved by the suffering and fate of a character, and feeling oppressed by the fate and suffering of a real person. When the reader displays an attitude of detachment towards a piece (book or film) or renounces an 'imaginative attitude' because he or she does not identify with that reality,[6] he or she cannot really feel moved (Neill, 1993: 354). Likewise, we can all feel emotionally detached from reality, and art is a privileged means to accomplish this. Andy Warhol advocated the need for an attitude of alienation, consisting in perceiving reality as a film, and film as if it was real. Such philosophy is based on a double assumption: first, that we can perceive the characters of a film as real people; and second, that we can see real people as characters of a film. Naturally, this approach entails some ethical problems. Someone might not respond emotionally to reality, showing a lack of concern for real consequences on real people (treating them as fictional characters).

What causes the aversion to reality may, according to E. Burke, become a source of intense pleasure according to the imitation provided by fiction and by art. Likewise, the delight in the real or imaginary suffering of one's equals happens when one's life is safe from any impending disaster – although that same condition does not impede that sympathy for another's maladies might occur. Burke suggests two possible explanations for this: on the one hand, the relief felt by realizing that one is before a work of fiction, as darker and terrible as it may be, and, on the other hand, the presumption of being safe from the

woes that one is witnessing (1993: 53). We are also moved by catastrophe or misfortune, be it real or imaginary. Overall, we are constantly dominated by an unconscious impulse:

> there is no other show that we so avidly seek like the one of some atrocious and uncommon disgrace; therefore, either if it happens before our eyes or if it takes place in the story, it always brings us pleasure. It is not pure, but mixed with reasonable unease. The pleasure we take from those scenes of great suffering stops us from avoiding them, and the pain felt induces us to comfort ourselves by doing it to those who suffer.
>
> (Burke, 1993: 54)

Burke admits 'a certain delight – and probably not a small one – in the misfortunes and real pains of others' (1993: 53). However, some religious perspectives opposed to modern sensibility assume that the contemplation of extreme suffering can prompt a transfiguration in which pain is associated with sacrifice, and sacrifice to exaltation. This tradition of empathetic meditation can be found, for instance, in the images of the suffering of Christ (the Passion of Christ) and that of his saints (Freedberg, 1991: 174–175). As auxiliaries for concentration and meditation, they were given the power to generate, in their recipients, a greater affective approach to Christ; and, in such a way, a greater ease in the modelling of their lives by the imitation of their examples. More recently, Susan Sontag argued in favour of this when describing a photograph owned by George Bataille, where a prisoner is shown being subjected to death by a hundred strikes, in China. She explains the need for one to take such an image of unbearable atrocity as an object of contemplation, as Bataille did every day, to 'elevate oneself above weakness', to 'become insensitive' or 'admit the eradicable' (Sontag, 2003: 104).

In either Poe's story or Bioy Casares' novel, human activities involve both those whose value comes from exercising a certain degree of control to dominate and administrate the natural or human environment (art, technology, handicraft), and those in which some human beings are – or actively choose to be – at the mercy of forces transcending their control. In the eyes of the developments of contemporary art, the model in 'The Oval Portrait' endures a long-term performance.[7] Just like a spectator of a play, witnesses suspend their disbelief before the said performance, and surrender to an illusion of creation and transference of life through the genius of a painter. However, contrarily to what would most likely happen to a spectator of a play, that illusion is not broken when they realize that the moment of the real (and not fictional) death of the model (performer) is approaching. Actually, it appears to be their most desired moment. The extreme resemblance produced by the wonderful skills of the painter is what least matters to them – rather, it is the reality promised by illusion, an illusion which, by means of an evasion, becomes a rival, exclusive reality. What matters to them is allowing the absent to become present, following the gradual withdrawal of the model (performer) from the world of the living.

It is as if they were indifferent to the distinction between theatrical staging and live-reality, refusing to acknowledge that the theatrical staging stops being a faithful representation to become a live-reality. If so, the life of the painting, the life captured in image, is literal (and not metaphorical).

Both stories provide feminine muses, configuring a type of contemporary art where the performance highlights the use of the body as a material, a means and an art object, whilst involving the audience in the artistic process. Resembling the abnegation and alienation of the feminine character from 'The Oval Portrait', Marina Abramovic presented herself to an audience, in Naples, in 1974, putting herself in a position of great psychological and emotional vulnerability. In this performance, *Rhythm 0*, the audience was instigated to use the instruments displayed on a table as they well pleased for six hours – these included a pistol and several instruments of pain and pleasure. For the time provided, the artist remained, as promised, passive to the turmoil of collective pulses and impassive to a crescendo of cruelty. In the first hours, the participants surrounded and grabbed her. In the following hours, they ripped her clothes, hurt her skin with razors and sucked her blood. The reaction of the participants to the challenge of rendition posed by the artist moved, later on, towards a dangerous and out-of-control situation. When someone placed a pistol in her hand, with one finger on the trigger, the audience divided itself into two factions: one, protective, and another, instigator.

This performance has some parallels with the blind obedience experiment (1963) conducted by Stanley Milgram,[8] and also with Philip G. Zimbardo's Stanford Prison Experiment (1971). In Milgram's experiment, a subject is ordered to administer increasingly severe punishments to victims in a context of learning experience where the victim must act obediently. The Stanford Prison Experiment is a well-known study of psychological and behavioural effects of prison life on volunteers who were randomly divided into two equally numbered groups, and assigned the role of either prisoners or guards.

In Abramovic's performance, the audience behaves obediently, accepting to be complicit in a depersonalization process as proposed by the artist. She is committed to being just a mere object, playing the role of a masochistic victim encouraging sadistic behaviour. However, the audience seems to enact the challenge too easily, tacitly assuming the teacher role or the prison guard. In doing this, each member of the audience gives up his own moral sense. Apparently, no one feels responsible for that woman, who stubbornly behaves like a mute object, a powerless prisoner. Before the performance began, she was still the authoritative figure, urging others to obey her rules. Afterwards, she annuls herself. An opportunity for diffusion of responsibility is created: everyone has projected responsibility outside the self and onto others. Behaving virtually as psychopaths or torturers, the audience has taken pleasure in exercising power over her body, intimacy, dignity, and humanity. Presumably, the audience's obedient behaviour was reinforced by observing similar obedient behaviour in peers (Zimbardo, 2004: 27), without thinking through the meaning or consequences of those actions. The level of aggression was increased in gradual steps. Audience and artist have proven to

be committed to each other's cooperation or compliance. She devoted herself to the 'cause' of depersonalization, becoming indifferent to others and less sensitive through division between self and body. At the same time, the audience is implicated by having accepted this situation. Just before she could become a voluntary martyr for the cause of art, the performance ends with the audience divided in two opposing groups. The blurring line between role-playing and reality enacted by the performance stops. We can speculate about the reasons for ending this situation. It may be because a subject (or a group) has urged the instigator group to exercise restraint and has found the idea of bringing death to a helpless woman, simulating her suicide, morally wrong. Given that emotions co-occur with moral judgments and some drive us to act, the subject may have been motivated by a sentiment of disapprobation, encompassing emotions of anger (even shame or guilt, in case of having also been a transgressor), when he/she saw the gun (Prinz, 2006: 30). Maybe someone realized the risks involved if the instigator's group had succeeded. Even in event of the artist's accidental death, this type of art would have lost any credibility. For sure, this moment brought the acknowledgment of a profound truth: we and our body are not two separate things, but only one. What is damaging to the body is also damaging for the self. When we cause harm to the body, we harden our soul (Scruton, 2009: 146–147).

As we have seen in 'The Oval Portrait' and *The Invention of Morel*, images are instigators of emotions even when a process of emotional erasure and silencing has prevailed in their production. In both stories, emotion is either the engine of artistic creativity, or artistic creativity the engine of emotion. In both, a curious inversion: the anthropomorphizing of the image (target of devotion and reverence) and the objectification of the living person (as if devoid of feelings, volitions and desires). But can art (the love for images) be a legitimate justification for the imposing of any type of suffering and sacrifice? Are all the adversities and sacrifices inflicted by men to other men in the name of art, all the decaying of human life in the name of a collective need for art, justified as long as they are considered a tribute to art? Most of all, is such an art worthy of such sacrifices and adversities? Although art does not have to moralize or to be at the service of propaganda, thereby impairing the artist's freedom and his sense of human affinity, art is never morally or politically neutral.

A true art appeals for and nourishes our higher nature (Scruton, 2009: 169). Exercising metaphorical imagination (or fancy) of things that do not really exist, helps us, readers and spectators, to nourish a 'generous construal of the world' (Nussbaum, 1995: 38). Whenever art or literature aid us in understanding the world, they inspire compassion and the passion for justice, giving voice to our questions, offering a new order or a glimpse of a better world, a new interpretation and evaluation of the past (ibid., 31).

The spectre of death without memory is present in both stories, and is reserved for those who are socially invisible. The power of art can fight this simple disappearing by providing intense sensorial experiences suitable for involving and committing a spectator/reader towards a meditation over the often unpleasant aspects of our human condition.

I conclude by quoting a timeless poem by Fernando Pessoa which acknowledges the need to distinguish between sentimental effusion and artistic activity:

> The poet is a pretender. / He so completely pretends / He even pretends it is pain / The pain that he truly feels. // And those that read what he writes, / In the pain they read they feel, /Not those pains that he felt, / But only that which they can't feel. // And so in the wheel tracks / It winds, entertaining reason, / That clockwork train / Which is called the heart.[9]

The feelings and emotions expressed in sentimental effusion by the poet who dislodges them from a comfortable system of habits are not the content, but, rather, a filter. Feelings transfigured by art, bearing universal meaning, are a demand for any artistic activity.

Notes

1 My analysis of an Edgar Allan Poe's 'Oval Portrait' is a partial and slightly modified version of the chapter 'Oval Office' from my doctoral thesis: Carlos Augusto Ribeiro, *We Are Not Alone Under the Skin – For A Possible Exposure About Doubles*, FSCH / Universidade Nova de Lisboa, 2007, pp. 84–92.

2 All references to 'Oval Portrait' report to a Portuguese edition of Poe's complete novels from 1972. All subsequent references to the novel will be identified in the text by the abbreviation OP.

3 For industrialized countries, distant territories represent the unknown and the trace of a lost past. Throughout the colonial and postcolonial periods, the camera did more than merely familiarize the Western spectator with foreign views, often capturing a disappearing past or preserving and packaging a world that appeared to be uncorrupted by industry and urbanization. Despite claims for its accuracy and trustworthiness, however, the camera was used as an instrument of symbolic control. It was used to record and define those who were colonized according to the interests of the West. The non-European world is portrayed as underdeveloped, stressing the indigenous nature of people, their settled lives, picturesque or exotic appearance and timeless existence. The enjoyment of advantages is widely represented by cinema, photography and television – by which the beneficiaries of material progress are convinced by the naturalness of their rights.

4 The narrator mentions the dream in which Morel appears as his rival double. In that dream, Morel is the warden of a madhouse. Morel's island is, simultaneously, a madhouse, purgatory or heaven. The whole life of the island was transformed into images, by which the persistence of the dead among the living becomes effective. Even the narrator presents himself as a ghost, a dead man for the image-people. It happens all the time an integral representation of reality by constant animation of images from an automated archive, as well as a restitution of a perfect illusion by means of sound, colour, size, smell and temperature. Morel's island is surrounded by a continuous flux of images. Multiple devices allow the narrator to review the island's past but also to programme the future through constant assembly of previous sequences, combined with seemingly random sequences. Morel's invention brings a new possibility, by which all that could be recreated, instead of being literally copied, could also be altered, possibly for the better, through recreation and montage.

5 The distinction between thought and emotion does not make sense in the eyes of recent advances in the neurosciences. There is a connection between electric circuits for emotion and cognition in a brain (Goleman, 2005: 225–254) or (Damásio, 2012: 195–235).
6 It is not necessary for the reader to believe in the reality of what happens to fictional characters to enjoy fiction. On the other hand, in real life, the emotional response presumes the belief in the reality of what happens, or what we see happening (either live or deferred), to real people.
7 Performance relates to the idea of a pure event, immediate, non-mediated (without representation), here-and-now.
8 The American psychologist Stanley Milgram carried out research on the role of destructive obedience in social and political life. He considered obedience to be the psychological bridge that links individual action to political purpose, men to systems of authority. A particular form of obedience studied by Milgram is blind obedience, which entails acts of aggression against others, such as those infamous episodes from contemporary history. But Milgram also recognizes that obedience may be ennobling and educative and refers to acts of charity and kindness, and not exclusively to destruction. He described a procedure for this study in an article published in *The Journal of Abnormal and Social Psychology*, Vol. 67, No. 4, 1963. See http://www.columbia.edu/cu/psychology/terrace/w1001/readings/milgram.pdf (accessed 25 July 2015).
9 Fernando Pessoa, 'Psicografia' in *Poesias. Fernando Pessoa.* Lisboa: Ática, 1995 (1942): 235.

References

Alain, Émile-Auguste (1999). *Système des Beaux-Arts*, Paris: Gallimard [1926].
Bazin, Andre (1992). *O que é o cinema?* Lisboa: Livros Horizonte [1958].
Beloff, Zoe (2015). 'The Days of the Commune – Notes and Reflections', in Pascal Gielen and Niels Van Tomme (eds). *Aesthetic Justice and Moral Perspectives*. Amsterdam: Valiz.
Burke, Edmund (1993). *Uma investigação filosófica sobre a origem de nossas ideias do sublime e do belo*. São Paulo: Papirus [1757].
Casares, Adolfo Bioy (1984). *A invenção de Morel*. Lisboa: Antígona [1940].
Castro, Ilda Teresa de (2013). 'Empatia e consciência moral', in João Mário Grilo and Maria Irene Aparício (ed)., *Cinema e filosofia – compêndio*, Lisboa: Colibri.
Costello, Diarmuid and Willsdon, Dominic (2008). *The Life and Death of Images – Ethics and Aesthetics*. London: Tate.
Damásio, António (2012). *Ao encontro de espinosa – as emoções sociais e a neurologia do sentir*. Lisboa: Círculo de Leitores [2003].
Fischer, Ernst (2010). *The Necessity of Art*. London andNew York: Verso [1971].
Freedberg, David (1991). *The Power of Images – Studies in the History and Theory of Response*. Chicago and London: The University of Chicago [1989].
Eco, Umberto (1989). *The Open Work*. Cambridge, MA: Harvard University Press.
Goleman, Daniel (2005). *Emoções destrutivas e como dominá-las – um diálogo científico com Dalai Lama*, Lisboa, Círculo de Leitores [2003].
Muller, Nat (2015). 'No Show – Refusal as Critique'. In Pascal Gielen and Niels Van Tomme (eds). *Aesthetic Justice and Moral Perspectives*. Amsterdam: Valiz.
Neill, Alex (1993). 'Fiction et émotions'. In J.-P. Cometti, J. Morizot and R. Pouivet (eds), *Esthétique contemporaine – art, représentation et fiction*. Paris: Vrin.

Nussbaum, Martha Craven (1995). *Poetic Justice – The Literary Imagination and Public Life.* Boston, MA: Beacon Press.

Poe, Edgar Allan (1972). 'O retrato oval'. In *Histórias completas de Edgar Poe*, trans. João Costa. Lisboa: Arcádia, pp. 127–130.

Prinz, Jesse (2006). 'The Emotional Basis Of Moral Judgements'. *Philosophical Explorations*, 9(1): 29–43.

Ribeiro, Carlos Augusto (2007). *Não estamos sós sob a pele – para uma exposição possível acerca de duplos*, PhD thesis, Universidade Nova de Lisboa.

Ricoeur, Paul and Changeux, Jean-Pierre (1998). *O que nos faz pensar?* Lisboa: Edições 70.

Scruton, Roger (2009), *Beleza.* Lisboa: Guerra & Paz.

Schimmel, Paul (ed). (1998). *Out of Actions – Between Performance and the Object 1949–1979*, Museum of Contemporary Art, Los Angeles. New York and London: Thames and Hudson.

Sontag, Susan (2003). *Olhando o sofrimento dos outros*. Lisboa: Gótica.

Singer, Peter (2005). *Como havemos de viver? A ética numa época de individualismo*. Lisboa: Sociedade Portuguesa de Filosofia/Dinalivro [1993].

Tolstöi, Léon (2006). *Qu'est-ce que l'art?* Paris: Puf [1931].

Tomme, Niels Van (2015). 'Six Acts – Or an Experimental Approach to Justice' (Conversation with Carlos Motta), in Pascal Gielen and Niels Van Tomme (eds), *Aesthetic Justice and Moral Perspectives*. Amsterdam: Valiz.

Watzlawick, Paul (1991). *A realidade é real?* Lisboa: Relógio D'Água [1977].

Zimbardo, Philip G. (2004). 'A Situationist Perspective on the Psychology of Evil: Understanding How Good People Are Transformed into Perpetrators'. In A. G. Miller (ed.), *The Social Psychology of Good and Evil*. New York: Guilford Press.

Index