Breaking Shared Norms, the Moral Commitment Problem, and Moral Emotions

Breaking Shared Norms, the Moral Commitment Problem, and Moral Emotions:

An Evolutionary Analysis

By

Lieven J. R. Pauwels and Ann De Buck

Cambridge
Scholars
Publishing

Breaking Shared Norms, the Moral Commitment Problem, and Moral Emotions: An Evolutionary Analysis

By Lieven J. R. Pauwels and Ann De Buck

This book first published 2026

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

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

ISBN: 978-1-0364-6476-9
ISBN (Ebook): 978-1-0364-6477-6

TABLE OF CONTENTS

FOREWORD

This book, *Breaking Shared Norms: The Moral Commitment Problem and Moral Emotions: An Evolutionary Analysis*, holds both academic and personal significance. It is the second book resulting from the *Human Cooperation Survey,* a research project launched in 2019 to deepen our understanding of cooperation and its various breakdowns. The project was conceived to explore individual differences in cooperative behavior, not solely through game-theoretical approaches that measure generosity beyond kinship, but also through survey-based methods capturing trait-like characteristics associated with both cooperative and psychologically selfish tendencies.

The first book, *Explaining Judgments on Rule Violations: On Empathy, Moral Intuitions, and Emotions* (2022), dealt with a partial test of the Moral Foundations Theory, developed by Jonathan Haidt, Jesse Graham, and colleagues. This study used both the Moral Foundations Questionnaire (MFQ) and the Moral Foundations Sacredness Scale (MFSS). We examined how strongly empathy, moral intuitions, and emotions were related to moral judgments of various everyday rule violations, ranging from minor theft and breaking promises to the willingness to punish free riders. Building on this foundation, the present book expands the inquiry by situating the concept of rule violations within the well-established framework of the moral commitment problem—an issue that is ubiquitous in everyday life. It further explores the evolutionary roots of moral emotions and social preferences as evolved mechanisms for navigating situations that involve moral commitment problems.

We are both deeply passionate about understanding the evolutionary origins of human behavior, particularly the biological and social underpinnings of cooperation, morality, and social norms. At the same time, as criminologists and social scientists, we are driven by a desire to better understand the causes of antisocial behavior. Contemporary criminological theory, in our view, often falls short in this respect. Many dominant theories do not sufficiently distinguish between proximate and ultimate explanations, nor do they always acknowledge the double-edged nature of key criminological constructs.

There are many reasons why the evolutionary perspective remains underutilized in criminology and the social sciences more broadly. Without claiming to be exhaustive, we highlight several common pitfalls: (1) Some stem from misunderstandings, most often invoking some form of genetic or biological determinism; (2) Others result from flawed studies that have fueled stereotypes and contributed to the poor reputation of evolutionary social science, particularly due to misguided interpretations in the field of evolutionary psychology; (3) Ideological concerns also play a role, such as the genuine fear that evolutionary science might be misused for political purposes, as has occurred in the past—though it should be noted that social sciences, especially learning theories, have likewise been appropriated by harsh authoritarian or totalitarian regimes; (4) And finally, a widespread misconception is that evolution refers solely to genes, brain morphology, or innate cognitive architecture, while in reality, culture—broadly defined—is also an evolving system.

As we write this foreword, we are already well into the 21st century. Indeed, a quarter of it has already passed. Remarkably, evolutionary science has made significant progress during this time. Recent studies have shed new light on our deep evolutionary past, with important implications for how we, as social scientists, view human nature and social order. It even affects the way we look at religion. These developments should not give rise to skepticism or moral outrage toward evolutionary approaches in the social sciences. Rather, they invite us to adopt a more humble and nuanced perspective on our place in the universe.

That said, let us return to the core of this foreword: introducing what this evolutionary-inspired book is all about. Commitment problems are the explanandum in this book, and specific norms (self-serving justifications), emotions, empathy, and self-control are independent variables studied in relation to these different commitment problems. It is rare for different commitment problems with overlapping evolutionary roots to be studied together. Concepts such as shame, guilt, empathy, and self-control are often, particularly in outdated versions of criminological theories, treated as unidirectional forces that either promote or inhibit antisocial behavior. We argue that these concepts function as context-sensitive tools evolved to regulate social behavior. As the renowned neurobiologist and primatologist Robert Sapolsky emphasizes, these traits are double-edged swords reflecting the deep complexity of social behavior: the very same mechanisms can lead to opposed outcomes depending on the context in which they are activated. For example, a hormone, a small brain region, or a personality trait might promote both pro-social and antisocial behavior in different settings. This

contingent nature of sociality underlines the fact that humans are capable of both violent destructiveness and profound altruism, sometimes within the same individual. Cooperation, too, is a double-edged sword—it can foster social cohesion but also lead to exclusion (ostracism) or harm, depending on the circumstances.

Viewed through an evolutionary lens, the target of moral and social norms, as well as behavior, becomes crucial. Mechanisms such as kin selection, direct and indirect reciprocity, social (including sexual) selection, and gene-culture co-evolution form evolutionary pillars underlying more proximate determinants of whom individuals choose to help, protect, punish, or harm. While some criminological research has begun to explore these distinctions, we believe that such evolutionary perspectives remain underutilized and hold significant promise for advancing the field. Social scientists should not fear adopting evolutionary insights to help organize knowledge.

An evolutionary perspective that underestimates the role of the environment in favor of evolutionary biology is no more helpful than a perspective that neglects our evolutionary heritage altogether. Metaphors, though undeniably powerful, have at times given rise to simplistic one-liners about "human nature." Yet it is precisely this concept of human nature (along with the notion of social order) that has been at the center of long-standing theoretical battles within criminology. These battles, in retrospect, have not proven especially fruitful. Still, perhaps we needed to go through them to realize that their practical value was limited.

Among the many thinkers who have shaped our approach, we would like to explicitly acknowledge the influence of Richard D. Alexander. His seminal works, such as *Darwinism and Human Affairs* (1979) and *The Biology of Moral Systems* (1987), laid the groundwork for a field at the intersection of evolutionary biology and behavioral sciences. Alexander was part of what Robert Cliquet—professor emeritus of anthropology and social biology at Ghent University—described as the “third Darwinian revolution”: a core group of scientists who helped resolve the apparent paradox of cooperation within a Darwinian framework of competition and natural selection. This intellectual tradition forms the backdrop against which we examine the emotional, intuitive, and moral mechanisms that regulate human social behavior, both prosocial and antisocial.

One of the authors of this book was a former student of Robert Cliquet and now teaches his course on biological anthropology for criminology students. Robert Cliquet’s key message was that to truly understand phenomena such

as racism, sexism, ageism, and egoism, one cannot meaningfully discuss these topics without a solid understanding of recent advances in evolutionary biology. While evolutionary biology may have had a troubled start, according to some uninformed or poorly informed scholars, this is certainly no longer the case today.

Earlier, economist Robert Frank introduced this evolutionary perspective into economics, challenging the overly simplistic view of the rational, self-interested human animal that prevailed at the time. Building on his foundation, we examine everyday commitment problems, such as theft, breaking promises, and the punishment of free riders, from the perspective that selective evolutionary forces have shaped empathy and emotions to help address these challenges, at least in part. Consequently, individual differences are to be expected, and these differences should be related to how individuals deal with commitment problems. In the empirical chapters of this book, we pay considerable attention to individual and sex differences, aiming to explain them in light of contemporary evolutionary science. We demonstrate that small sex differences—rather than large ones—are to be expected, and that it is important to consider not only average differences but also variation within groups. Of course, Richard Alexander and Robert Frank are not the only thinkers whose work we draw upon. Contemporary scholars such as Michael Tomasello, the late Frans de Waal, and Robert Sapolsky have also deeply informed our thinking. Alongside many others, too numerous to list here, their contributions have been invaluable to the development of this book. All are duly acknowledged in our bibliography.

This book brings together the complementary expertise of both authors in evolutionary science and criminological theorizing in general, and in the study of moral emotions in criminal decision-making in particular. We hope that this volume contributes to the growing dialogue between criminology, evolutionary theory, and the broader behavioral sciences. At the very least, we hope it sparks curiosity about the fundamental question that drives our work: not just why people ultimately break norms, but why, so often, they choose not. We hope to reinvigorate criminologists' attention to integrating both ultimate and proximate research questions and mechanisms, an approach commonly used in standard studies of animal behavior.

Some scholars deserve to be mentioned in this foreword, though any list would inevitably omit worthy names. We thank everyone who believed in the value of the evolutionary approach. Special gratitude goes to the anonymous reviewers for their supportive and constructive comments,

which helped refine this book. We also gratefully acknowledge Karin Van Peteghem for her assistance in preparing and properly formatting the tables and figures. Finally, we warmly thank Dr. Johan Braeckman, former professor of moral philosophy at Ghent University and a true expert on the life of Charles Darwin, for his invaluable support in integrating evolutionary thinking into the social sciences and humanities within our research.

Ghent, August 2025
Prof. dr. Lieven Pauwels & dr. Ann De Buck

INTRODUCTION AND AIM OF THE STUDY

1. Cooperation and the moral commitment problem

Cooperation is undeniably a hallmark of human societies. Although not exclusive to humans, the **breadth and complexity** of our cooperative behavior far surpass those seen in other mammals (Apicella & Silk, 2019; Henrich & Muthukrishna, 2021; Raihani, 2021). Across cultures, people routinely help one another: parents care for their children (Barash, 2001, 2007), friends offer mutual support (Dunbar, 2021), and even strangers often show kindness without expecting anything in return (McCullough, 2020). Scholars have described human societies as **ultra-social** (Richerson & Boyd, 1998; Turchin, 2015), **super-cooperative** (Nowak & Highfield, 2012), and **socially interdependent** (Tomasello, 2016, 2019). Cooperation takes many forms, including helping, avoiding harm, sharing, teaching, reciprocating, caregiving, collective childcare, and the creation and enforcement of shared norms and institutions.

Although definitions of cooperation vary, at its core, it refers to behavior that benefits another individual (the recipient). Depending on the outcomes, cooperation may be altruistic, mutually beneficial, or reciprocal (Axelrod & Hamilton, 1981; Sober & Wilson, 1998; Trivers, 1971, 2006). On a larger scale, cooperative individuals can achieve shared goals that would be difficult or impossible to reach alone. For our ancestors, group living likely became a key survival strategy. For most of human history—until about 15,000 years ago—humans lived in small hunter-gatherer bands and faced volatile ecological threats such as resource scarcity and predation (Høgh-Olesen, 2010). These pressures created strong incentives for cooperation. The more an individual's survival depended on the group, the more valuable cooperative behavior became for the individual.

A fundamental feature of human cooperation is that individuals must either suppress their self-interest in favor of others (e.g., helping, sharing, avoiding harm) or align their self-interest with that of others (e.g., reciprocity) (Tomasello & Vaish, 2013). In this context, self-interest is viewed as a psychological mechanism by which individuals are motivated to prioritize themselves over others, such as by keeping resources for themselves (Tomasello, 2016), distinguishing it from behavior driven by evolutionary

fitness (Sober & Wilson, 1998). Evolutionary biologist Richard Alexander (1979, 1987) emphasized the evolutionary significance of conflicts of interest. These arise from the fact that each human being is genetically unique, with a distinct set of personal interests shaped by individualized genetic variation. As a result, people often compete over limited resources—pursuing goals that not everyone can attain equally. These insights are central to Alexander's work (1979, 1987).

To understand cooperation fully, it is important to distinguish between its **ultimate function** (what it is good for) and its **proximate internal mechanisms** (how it arises) (Mayr, 1961; Tinbergen, 1963). Ultimate considerations of cooperation refer to the benefits that an organism or its close relatives gain from cooperating, which helps explain why such behavior has been favored by natural selection in response to complex social interactions. From an evolutionary perspective, cooperation in humans was advantageous in contexts where working together was more beneficial than acting alone, such as cooperative hunting, food sharing, aiding in times of danger (e.g., defense against predators or rival groups), sharing knowledge, and building shelter (Trivers, 2002). Strong selection pressures likely favored behavioral tendencies that promoted cooperation when it contributed to an individual's reproductive success (Duntley & Shackelford, 2008; Ridley, 1997).

The proximate perspective, by contrast, focuses on the immediate triggers of cooperative behavior and the psychological mechanisms that support it (de Waal, 2008). Evolutionary theory suggests that the structure and variability of human social cooperation stem from traits, and preferences that evolved through natural and sexual selection (Cronin, 1994), along with other evolutionary processes. Over generations, these traits became widespread in populations (Dawkins, 2016 [originally published 1976]). Research in evolutionary cognitive neuroscience (Shepherd, 2017) suggests that human social behaviors evolved in response to selection pressures, making our hominin ancestors more social, tolerant, and capable of forming stronger group bonds (de Waal, 2006; Turner, 2021; Wrangham, 2019). Evolution ultimately favored a large and flexible human brain capable of adapting to environmental conditions (Camilleri, Rockey & Dunbar, 2023; Mitchell, 2018), along with specialized cognitive mechanisms for decision-making based on conditional "*if-then*" rules (Platek, Keenan & Shackelford, 2007).

Moreover, the human brain became "soft-wired" for flexible learning, capable of acquiring social norms, experiencing emotions, exercising self-

control, and adjusting to interdependent group life within the local social and ecological environment (e.g., Buss, 2016; Chudek & Henrich, 2010; Henrich, 2016; Tomasello, 2022; Tomasello et al., 2012). Crucially, humans are not merely learners but also rule users: they actively represent and apply conditional "if-then" rules to guide behavior. Individuals with the physiological capacity to apply conditional "if-then" rules in cooperative contexts—such as helping, not hurting, reciprocating, forgiving, or using strategies like tit-for-tat and win-stay, lose-shift (Nowak & Sigmund, 1993)- had a higher chance of survival and reproduction. These capacities are supported by neurobiologically grounded systems of cognitive control that enable the flexible formulation and application of conditional rules (Bortfeld & Bunge, 2024; Bunge, 2004; Zelazo, 2008; Zelazo & Frye, 1997).

Whether we are driven primarily by self-interest or by the desire to prioritize the greatest good for the greatest number (Sapolsky, 2017) remains a longstanding puzzle that has sparked century-old debates and discussions about coordination, cooperation, and reciprocity. This deeply embedded tension between individuality and community, cooperation and cheating, is captured by the term "the commitment problem" (although other terms are also used). Commitment problems (Frank, 1988, 2001, 2011) stem from conflicts between short-term self-interest and longer-term collective interest (Dawes, 1980). If all humans shared the same interests, their goals would align, and cooperation would be straightforward. However, while cooperation offers benefits beyond what individuals could achieve alone, each person may still be tempted to give less and take more, undermining the collective effort. If everyone acts selfishly, cooperative systems can break down to the detriment of all involved (Greene, 2013).

One of the best-known examples of the commitment problem is probably the Prisoner's Dilemma, a theoretical model from game theory that mathematically explores strategic social interactions, including cooperation and cheating (Camerer & Fehr, 2004; Rand & Nowak, 2013). In essence, commitment problems are cooperation problems rooted in conflicts of interest. They are significant because they are pervasive in social life.

2. Solutions to moral commitment problems

A commitment is a pledge (or threat) to take action that an individual might not otherwise be inclined to take (Hirshleifer, 1987: 309), such as promising to help or threatening to harm someone. These commitments serve as signals that limit behavioral options, favoring a specific course of action in

certain anticipated future circumstances (Nesse, 2001). The commitment problems discussed here involve moral conflicts of interest – the tension individuals experience when acting against their immediate (psychological) self-interest to uphold societal or personal moral norms. We refer to these as *moral commitment problems*.

Resolving these conflicts requires mechanisms that shift individual incentives to align with the desired action, ensuring that a person follows through on their promise or threat (Schelling, 1960). One compelling solution lies in the role of *moral emotions*, which serve as psychological mechanisms that promote cooperation in social situations involving conflicts of interest.

Robert Frank (1988), along with others such as Hirshleifer (1987) and Schelling (1960), was among the first to propose that natural selection might have shaped moral emotions to secure commitments and encourage cooperation. He argued that emotions like (moral) indignation (a subtype of anger, Haidt, 2003; but see Turner, 2007), and guilt may have evolved as emotional commitment devices. The term "*emotional commitment devices*" describes commitments facilitated by emotions (Frank, 2001). In this view, emotions serve to influence behavior by motivating individuals to follow through on their committed course of action when the situation arises. Put differently, commitment devices promote cooperative behavior by limiting an individual's behavioral options, steering them toward the behavioral options available to the individual (Kelly, 2011).

For our purposes, emotions such as (moral) indignation and guilt function as proximate emotional mechanisms that may have evolved to shift individuals from selfish to cooperative behavior. While the temptation to steal might offer immediate rewards to an individual, if all group members acted similarly, it would disrupt the social harmony of the group in the long run. Although emotional commitments may limit access to potentially valuable short-term opportunities, they also offer significant long-term benefits (Frank, 2001). Frank's commitment model suggests that emotional predispositions help drive cooperative behavior. There is considerable evidence suggesting that evolutionary forces favored such emotions, at least in part, for this very reason. However, it is important to emphasize that emotions were unlikely to be selected solely for solving commitment problems. Rather, they must promote cooperation in ways that, over time, enhance the organism's overall fitness or adaptive success (Frank, 2001).

The argument that emotions, like (moral) indignation and guilt, may have evolved as proximate psychological mechanisms motivating individuals to commit to actions that may not serve their immediate interests (though they could benefit them in the long run) represents only one aspect of the explanation. The other involves understanding how emotions themselves were shaped as emotional commitment devices. This leads us to consider the precursors of moral emotions. Frans de Waal (2006, 2009, 2022) has long argued that moral commitment devices may have evolutionary precursors in tendencies such as empathy and altruism, with foundational building blocks already present in nonhuman primates. In his biosocial criminological theory, Anthony Walsh (2024) identifies empathy as an important psychological factor, an emotional deterrent, that motivates people to refrain from harming others. Empathic tendencies were likely shaped by kin selection (Hamilton, 1964) within the context of committed relationships with friends or mates, who represent potential partners for reciprocal exchange (Trivers, 1971). These tendencies may also be reinforced through reputation management (Alexander, 1987) or the avoidance of social punishment (Boehm, 2012).

3. Aim of the study

The primary aim of this study is to deepen our understanding of how moral emotions influence decision-making in scenarios involving commitment problems, also referred to as cooperation failures. These terms will be used interchangeably throughout this book. Specifically, we explore ***the direct and indirect relationships among three key moral commitment problems: (1) the tendency to refrain from stealing each other's property, (2) the tendency to honor promises, and (3) the tendency to punish free riding. We examine these in relation to the emotions of moral indignation and anticipated guilt, as well as several key explanatory variables identified in existing literature and prior studies.***

We focus specifically on these three types of commitment problems, refraining from stealing, honoring promises, and punishing free riders, because they reflect core dimensions of cooperation that are important to maintain social cohesion. Hoffman (2014) argues that human beings evolved to cooperate—often reluctantly—to survive as a highly social species. This cooperation centered primarily on two essential norms: **not stealing from others and not breaking promises.** These behaviors are foundational to trust and exchange. Punishment, though costly, served as a mechanism to deter such violations by making antisocial behavior more

expensive. Curry's (2016) **Morality-as-Cooperation** (MAC) theory further supports the relevance of these specific behaviors. Based on game-theoretical reasoning, MAC identifies seven forms of cooperation that address key social challenges. Among them are **recognition of property rights, reciprocity,** and **coordination for mutual benefit,** which align directly with the selected norm violations: **theft, broken promises**, and **free riding**. These behaviors threaten stable cooperation by undermining trust and fairness.

Our study is founded on an evolutionary-inspired integrated model of moral antecedents related to commitment problems. This research builds on our previous work (De Buck & Pauwels, 2022a), in which we introduced the concept of the moral commitment problem and explored the interrelationships between anticipated guilt and personal moral norms. Additionally, we expand upon our earlier findings regarding the moral antecedents of moral judgments (De Buck & Pauwels, 2022b), where we presented an evolutionary-inspired conceptual model that encompasses both distal and proximal explanatory variables of moral judgments in the context of four distinct moral violations.

In the present study, we further develop this framework by considering *intentions to steal, intentions to break promises, and intentions to punish a free rider* as outcome variables. We elaborate on the model by conceptualizing moral emotions such as *moral indignation and anticipated guilt,* along *with personal moral norms*, as proximate factors or mechanisms hypothesized to reduce the likelihood of intentions to steal, break a promise, and punish a free rider. Within this context, moral emotions are understood as *emotional commitment* devices, whereas personal moral norms are viewed as *cognitive commitment* devices. This conceptualization situates moral emotions and personal moral norms within a broader framework that also encompasses precursors or building blocks.

The precursors of moral emotions are deliberately chosen, and their roles, significance, and interrelationships are theoretically framed within evolutionary theories: **Gene-based selection** emphasizes how genetic diversity—variations of alleles within a human population—is encoded in our genomes and wired into our brains (Lindenfors, 2017). There is extensive genetic variation across individuals in every species, including humans, which affects most psychological traits indirectly and non-specifically through the wiring and development of the human brain (Mitchell, 2018). **Kin selection and reciprocity** focus on cooperative tendencies directed at kin (Hamilton, 1964), friends, and specific individuals

likely to reciprocate the benefits of cooperation (Axelrod & Hamilton, 1981; Trivers, 1971). **Gene-culture coevolution or dual inheritance** underscores cooperative tendencies based on a form of social or collective commitment directed at all members of one's moral community (Henrich & Henrich, 2007). Modern humans do not create their social commitments; they are born into them. However, individuals must self-regulate their behavior according to shared social and moral norms (Tomasello, 2016). A central aspect is the concept of evolved norm psychology, shaped by the co-evolution of cultural social norms in human societies over tens or even hundreds of thousands of years, alongside the genetic evolution of the human brain (Chudek & Henrich, 2010; Henrich & Muthukrishna, 2021).

Our conceptual model partially draws upon the arguments made by evolutionary economist Robert Frank (1988, 2001) and the logic of the evolutionary-based multiple moralities approach proposed by evolutionary developmental psychologist Michael Tomasello (2016, 2018). Frank argues that moral emotions (and shared moral norms) are powerful drivers of human decision-making, explaining why individuals often make choices that seem irrational from a narrow self-interest perspective. Tomasello's multiple moralities framework offers a useful framework for the conceptual organization of moral emotions and their precursors or determinants to explain individual variability in solving moral commitment problems.

This study is unique as it combines insights from evolutionary theory, sociobiology, evolutionary psychology, social sciences, and cultural evolution theory. We argue that modern evolutionary theorizing can serve as an explanatory principle for a wide range of human social behaviors, including crime and our responses to it (see e.g., Braeckman, 2001; Durrant & Ward, 2015; Raine, 1997; Walsh & Jorgensen, 2018). In criminological studies, the dependent variable is called crime, delinquency, or rule-breaking; however, biosocial criminologists often use the term antisocial behavior. Antisocial behavior is broader and includes socially harmful behaviors that may not always be illegal (Durrant & Ward, 2016; Fishbein, 2010). We have opted for the concept of moral commitment problems or cooperation failures because it facilitates a balanced interpretation of (psychological) self-interest and social behavior. This concept resonates with discussions of human nature. Since the time of the Ancient Greek philosophers, scholars have grappled with explaining human nature, often unsuccessfully. Since Darwin, it has become clear that human nature is shaped by evolution. Through natural selection (and other evolutionary processes such as social/sexual selection, kin selection, and direct and indirect reciprocity, cultural multi-level selection (Cliquet, 2010)), living

creatures, including humans, possess a genetically based behavioral repertoire that evolved to maximize their biological fitness in response to the challenges and opportunities faced as members of small hunter-gatherer bands, rather than under contemporary conditions (Turner, 2021; van Schaik, 2016; van Schaik & Michel, 2016; Wilson, 1978/2004).

The capacity for moral commitment is part of this behavioral repertoire. While living in social groups offers valuable benefits for survival and reproduction, it also introduces problems and conflicts over scarce resources. The difficulty of achieving cooperation among individuals pursuing their self-interest (Alexander, 1987) essentially encapsulates the commitment problem. However, we should keep in mind that "selfish" genes do not necessarily imply that self-interest—defined as the tendency to optimize outcomes for oneself over others (Ricard, 2015)—is central to human nature (Dawkins, 1976/2016; Williams, 1966). From an evolutionary perspective, emotions are adaptations[1] that partly motivate humans to make emotional commitments, leading them to forgo potentially valuable short-term opportunities for important long-term benefits, such as cooperation that benefits the group. Although the evolutionary mechanisms underlying the moral emotions characteristic of *Homo sapiens* are not yet fully understood, several plausible explanatory models exist (e.g., Alexander, 1987; Frank, 1988; Gintis, 2000; Henrich, 2016; Trivers, 1971).

Additionally, this study is distinctive in its focus on a sub-type of anger: moral indignation. Moral indignation belongs to the broader family of "other condemning" emotions, which includes contempt and disgust, and it serves as an emotional response to perceived violations of moral rules. This emotion can motivate individuals to punish those who fail to cooperate (Haidt, 2003). Moral indignation is a facet of human psychology that evolved, at least in part, to facilitate cooperation (Trivers, 1971). Therefore, our study aims to gain deeper insights into the determinants of morality, a crucial concept for understanding commitment problems.

[1] Central to an evolutionary perspective is the notion of adaptations (Williams, 1966). Adaptations are the evolutionary processes that consistently shape the features and characteristics of organisms, enabling them to effectively address recurrent challenges related to survival and reproduction, either directly or indirectly (Hamilton, 1964; Tooby & Cosmides, 1990). The primary function of adaptations is to solve an adaptive problem in a manner that enhances the organism's reproductive success (Buss, Haselton, Shackelford, Bleske & Wakefield, 1998).

4. Outline of the chapters

This book is organized into nine chapters. **Chapter One: Defining Moral Commitment Problems** introduces the concept of moral commitment problems and lays the groundwork for understanding them as cooperation failures. It provides definitions of cooperation and moral commitment problems.

Chapter Two: The Evolutionary Logic Behind Moral Commitment Problems explores the evolutionary foundations of moral commitment problems, explaining how evolved psychological mechanisms like moral emotions and personal moral norms help mitigate failures in cooperation. The chapter outlines evolutionary processes such as kin selection, reciprocal altruism, and gene-culture coevolution and situates moral commitment problems and moral commitment devices within this broader evolutionary framework.

Chapter Three: Towards a Testable Model introduces a conceptual model to explain individual differences in addressing moral commitment problems. Shifting from evolutionary to proximate explanations, the chapter examines how various psychological factors—referred to as commitment devices—like empathic concern, altruistic preferences, moral indignation, anticipated guilt, and personal moral norms influence moral decision-making. Tomasello's multiple moralities framework is used to explore these factors, providing the basis for a testable model that aims to explain individual differences in moral choices in situations entailing moral commitment problems.

Chapter Four: Study Sample and Design outlines the study's sample, methodology, and materials. It describes the operational definitions of key constructs, including moral commitment problems, and discusses the statistical methods used to analyze the data. This chapter offers a concise overview of the study's design.

Chapter Five: Three Moral Commitment Problems: Individual and Sex Differences investigates three moral commitment problems—intentions to steal, break promises and punish free riders—focusing on individual and sex differences. The first part of the chapter provides descriptive statistics that illustrate the variation in participants' moral intentions and the correlations between these variables. The second part examines sex differences, exploring whether males and females show

significant differences in their intentions related to stealing, promise-breaking, and punishing free riders.

Chapter Six: Evolved Social Preferences: Individual and Sex Differences explores evolved social preferences that shape cooperation, such as empathic concern, altruistic preferences, and self-enhancing/group-inclusive motives. The chapter analyzes how these preferences vary across individuals, with a particular focus on sex differences.

Chapter Seven: Commitment Devices: Individual and Sex Differences focuses on moral commitment devices—moral indignation, anticipated guilt, and personal moral norms—that influence moral decision-making. The chapter presents descriptive statistics on how these devices vary among individuals and examines potential sex differences. It explores how these moral emotions and judgments shape responses to situations entailing moral commitment problems.

Chapter Eight: Multivariate Analyses of the three Commitment Problems applies multiple regression models to investigate how evolved social preferences and moral commitment devices account for individual differences in intentions to steal, break a promise, and punish a free rider. The chapter begins with sequential logistic regression analyses to explore how different predictor variables contribute to explaining moral intentions. It then uses path analysis to examine direct and indirect effects, exploring how social preferences influence moral intentions through commitment devices.

Chapter Nine: Discussion, Limitations of the study, and Future Directions addresses the limitations of the current study. It highlights the theoretical and methodological challenges associated with the research. The chapter identifies potentially fruitful avenues for future exploration. Additionally, it offers recommendations for crime prevention policy.

CHAPTER ONE

DEFINING MORAL COMMITMENT PROBLEMS

1. Introduction

In this chapter, we focus primarily on defining the construct we seek to explain. Since we use the concept of *commitment problems* as failures of cooperation, it is important to begin by clarifying what *cooperation* entails. A significant source of confusion in scientific research, particularly in interdisciplinary contexts, stems from variations in definitions. Different scientific communities, and even individual scholars within the same field, often use inconsistent terminology. They may refer to the same phenomenon using different terms or use the same term to describe different phenomena (West, Griffin & Gardner, 2007).

This lack of clarity and poor communication can hinder scientific progress. In psychological disciplines, such semantic and conceptual confusion is known as the *jingle-jangle fallacy* (Block, 1995). The *jingle fallacy* (Thorndike, 1904) refers to using the same term for distinct concepts, while the *jangle fallacy* (Kelley, 1927) involves using different terms for equivalent or highly similar concepts.[2]

We begin by defining *cooperation* as a foundation for understanding *commitment problems*. We then turn to the definition of the *moral commitment problem*, illustrating how it can be formalized to more effectively capture its social and strategic dimensions.

2. What is cooperation?

There is a vast body of literature about cooperation, resulting in a wide range of definitions (e.g., Axelrod & Hamilton, 1981; Barclay & van Vugt, 2014; Durrant & Ward, 2015; Melis & Semmann, 2010; Nowak & Coakley, 2013; Tomasello, 2009; West et al., 2007) (see Table 1.1). In everyday language,

[2] For instance, definitional assumptions regarding the concept of self-control illustrate this issue (Milyavskaya, Berkman & De Ridder, 2019).

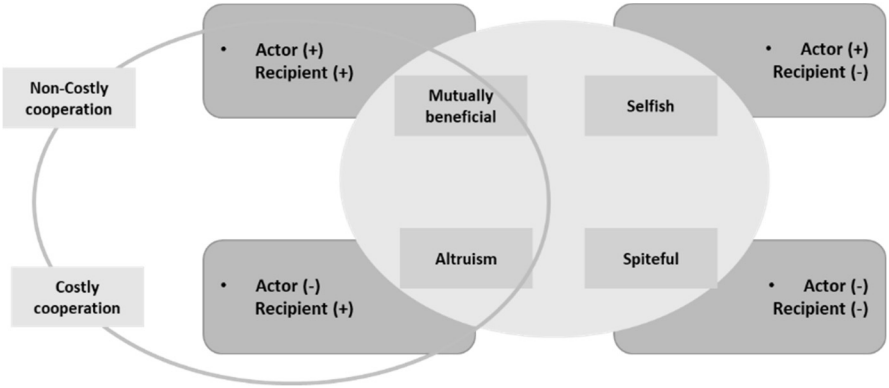
cooperation is often understood as teamwork, individuals working together toward a shared goal. According to Coakley and Nowak (2014), scientists sometimes blur the distinction between *cooperation* and a related phenomenon, *altruism*, which incorporates specific motivational and intentional elements. They emphasize the importance of distinguishing between a form of cooperation that merely "happens"—and can occur across the evolutionary spectrum—and altruism, which is driven by intentional goals or emotional commitments.

This distinction is critical, as it speaks to one of the most intriguing and debated questions in evolutionary theory: at what point does cooperation become altruism? Nevertheless, Coakley and Nowak also point out that cooperation can be precisely defined when framed mathematically, for example, in the context of the Prisoner's Dilemma (PD) or other strategic games (e.g., Maynard Smith, 1982).

The distinction between cooperation and altruism is further clarified by examining the consequences of social behavior (West et al., 2007a, 2011). Building on Hamilton's (1964) classification of social behaviors, these actions can be categorized according to whether they confer benefits or costs to both the actor and the recipient. For example, a behavior that benefits the actor but harms the recipient (+/−) is considered *selfish*, while one that benefits both parties (+/+) is *mutually beneficial*. *Altruism* occurs when a behavior is costly to the actor but beneficial to the recipient (−/+), whereas *spiteful* behavior imposes costs on both (−/−) (Hamilton, 1964, 1970; West et al., 2007b) (see Figure 1.1).

West et al. (2007) define cooperation as a behavior that provides a benefit to another individual (the recipient) and is selected because of its beneficial impact on the recipient, even though it may be beneficial or costly to the actor. Cooperation can involve temporary costs to the actor, unlike other forms of behavior that either impose no costs or yield immediate benefits to the actor (non-costly cooperation). This definition encompasses both altruistic (−/+) behaviors, which are a form of costly cooperation, and mutually beneficial (+/+) behaviors, which involve no cost to the actor (non-costly cooperation). Tomasello (2009) supports this view by defining altruism as *one individual sacrificing in some way for another,* and cooperation as *multiple individuals working together for mutual benefit* (p. XVII).

Figure 1.1. *The evolutionary outcomes of social behavior (Adapted from Durrant & Ward, 2015: 102)*



Note. Social interactions between members of the same species can result in a range of outcomes. Selfish behavior (+/-) benefits the actor (+) at the cost of the recipient (-). Spiteful behavior (-/-) is costly to both the actor (-) and recipient (-). Altruistic behavior (-/+) benefits the recipient (+) at a cost to the actor (-); Mutually beneficial behavior (+/+) benefits both the actor (+) and the recipient (+). Cooperative behavior (oval) includes altruistic behaviors (costly cooperation) and some mutually beneficial behaviors (noncostly cooperation) (Barclay & van Vugt, 2014; Durrant & Ward, 2015; West et al., 2007).

For this study, we define cooperation as *any action in which one individual incurs a cost, at least temporarily, while another gains a benefit as a result.* In terms of the convergence and conflict of interests, cooperation requires individuals to either suppress their self-interest or, at least, equate it with the interests of others (Tomasello, 2016; Tomasello & Vaish, 2012). Here, *self-interest* refers to the tendency to prioritize oneself over others (see also Tomasello, 2016).

This definition emphasizes the *behavioral* dimension of cooperation and frames it as a form of social behavior. *Altruism*, by contrast, refers to a subset of cooperative behaviors: specifically, actions in which an individual—motivated by goodwill, concern for others, or sympathy—willingly incurs a cost so that another individual may benefit. This definition incorporates both behavioral and *psychological* components. Thus, altruism can be viewed as a *costly form of cooperation*, where the benefit to the recipient comes at the expense of the actor (see also Jensen, 2012, p. 566, in Durrant & Ward, 2015). Altruism inherently involves some degree of *self-sacrifice* (Atkins, Wilson, & Hayes, 2019).

From an evolutionary perspective, *costly cooperation* presents a significant puzzle: How can other-regarding behavior that imposes costs on individuals, while benefiting others, be sustained through natural selection? Why would an individual engage in cooperative behavior that appears costly to perform yet advantageous to someone else (Hamilton, 1963, 1964)? Natural selection generally favors the evolution of behaviors that enhance an individual's fitness, that is, the likelihood of passing on copies of one's genes to future generations. Individuals gain *inclusive fitness* either directly, through their own reproductive success, or indirectly, through the reproductive success of genetic relatives (Hamilton, 1964).

Actions that increase the fitness of a recipient at the expense of the actor seem to contradict this evolutionary logic, as they do not directly enhance the actor's reproductive success (Apicella & Silk, 2019). However, a deeper understanding of how both *direct* and *indirect fitness* contribute to the evolutionary maintenance of cooperation helps resolve this apparent contradiction. Humans cooperate on a much larger scale than is typically observed in other species (Sapolsky, 2017). While small groups of individuals often cooperate, human cooperation extends far beyond kinship boundaries, encompassing reciprocating partners, friends, and even strangers. Efforts to unravel the puzzle of large-scale human cooperation have led to important advances across a range of disciplines (e.g., Barkow, Cosmides & Tooby, 1992; Fehr & Gächter, 2002; Henrich, 2016; Nowak, 2006; Ostrom, 1990; Tomasello et al., 2012).

Human cooperation takes various forms, each addressing distinct challenges inherent to social life. Theories like **Moral Foundations Theory (MFT; Graham, Nosek, Haidt, et al., 2011;** Haidt, 2012) and **Morality as Cooperation (MAC;** Curry, 2016; Curry, Chesters, & Van Lissa, 2019) suggest that these types of cooperation are fundamental to human interaction. Grounded in evolutionary biology and anthropology, these theories argue that, despite cultural variations in specific moral rules across societies, moral systems universally center around core cooperative principles. **MAC** and **MFT** identify several key forms of cooperation, including **kinship cooperation, reciprocal altruism, fairness and distributive justice,** and **coalitional cooperation.**

These frameworks propose that human morality evolved as a multifaceted solution—spanning biological, psychological, and cultural domains—to address recurring cooperation problems in social life. One such solution, the evolution of *moral emotions and social preferences,* will be explored in

greater depth in Chapter Two, *The evolutionary logic behind moral commitment problems*.

Table 1.1. *Overview of definitions of/views on cooperation and altruism (non-exhaustive)*

COOPERATION	
Multiple individuals working together for mutual benefit	Tomasello, 2009, XVII
A form of working together in which one individual pays a cost (in terms of fitness, whether genetic or cultural) and another gains a benefit as a result	Nowak & Coakley, 2013: 4
A subcomponent of prosocial behavior, i.e. behavior intended to benefit one or more people other than oneself. Some types of prosocial behavior are costly to the actor, at least temporarily (costly cooperation), whereas other types of prosocial behavior carry no costs or provide return benefits to the actor almost immediately (non costly cooperation)	Barclay & van Vugt, 2014: 2
Those (behaviors) that have been selected to provide benefits to recipients	Durrant & Ward, 2015: 102
…. cooperation focuses on how and why individuals make choices that help others (or avoid hurting them) at a personal cost	Henrich & Muthukrishna, 2021: 208
ALTRUISM	
With respect to humans,……. *'would from an inherited tendency be willing to defend, in concert with others, (their) fellow-men; and would be ready to aid them in any way, which did not too greatly interfere with (their) own welfare or (their) own strong desires'*	Darwin, 1871: 106
Behavior that benefits another organism, not closely related, while being apparently detrimental to the organism performing the behavior, benefit and detriment being defined in terms of contribution to inclusive fitness	Trivers, 1971, p.35
A willingness to act in consideration of the interests of the other person, without the need of ulterior motive *The direct influence of one person's interest on the actions of another, simply because in itself the interest of the former provides the latter with a reason to act*	Nagel, 1978: 79-80
An organism behaves altruistically – in the evolutionary sense of the term – if it reduces its own fitness and augments the fitness of others. In contrast, the concept of psychological altruism applies, in the first instance, to motivational states, and only derivatively to the behaviors those motives may cause.	Sober & Wilson, 1998: 199

Behavior costly to the actor and beneficial to the recipient, measured in lifetime personal fitness consequences	West et al., 2007: 416
Altruism (biological definition): behavior that increases the recipient's fitness at a cost to the performers	De Waal, 2008: 280
One individual sacrificing in some way for another	Tomasello, 2009, XVII
A motivational state with the ultimate goal of increasing another's welfare	Batson, 2011: 20
A form of (costly) cooperation in which an individual is motivated by goodwill or love for another (or others)	Nowak & Coakley, 2013: 5
Actions which decrease one's lifetime reproductive success, and it is one form of costly cooperation	Barclay & van Vugt, 2014: 2
When the recipient benefits at a cost to the actor	Durrant & Ward, 2015: 102
An entity, …., is said to be altruistic if it behaves in such a way as to increase another such entity's welfare at the expense of its own……. It is important to realize that the above definition(s) of altruism….are behavioral, not subjective. …..My definition is concerned only with whether the effect of an act is to lower or raise the survival prospects of the presumed altruist and the survival prospects of the presumed beneficiary.	Dawkins, 2016/1976: 5

3. The moral commitment problem

Building on Frank's (1988) analysis of commitment problems, we use the term *moral commitment problem* to describe situations where individuals sustain cooperative behavior not through external enforcement or coercion, but because of evolved and internalized emotions and social preferences such as empathy, altruism, or guilt. Although Frank himself does not explicitly use this phrase, his work implies that emotions act as reliable signals of genuine commitment, enabling cooperation to persist despite short-term incentives to cheat. This moral aspect highlights the crucial role of evolved psychological mechanisms that support cooperation even when cheating would yield immediate material benefits.

Commitment problems are known by various names across disciplines. Economists and political scientists often refer to it as a commitment or collective action problem (e.g., Camerer, 2003; Frank, 1988; Ostrom, 1990), while other social scientists describe it as a social dilemma (e.g., Dawes, 1980) or a social problem (Hoffman, 2014). In evolutionary biology, it is termed the problem of altruism or the issue of reciprocity (e.g., Trivers,